National College of Ireland

Analysing the impact of social media on online shopping platform sales by using sentiment analysis with text mining

MSc Research Project
Programme Name

Xiyao Xu
Student ID: x18107834

School of Computing
National College of Ireland

Supervisor: Bahman Honari

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. ……xiyao xu……

Student ID: ……x18107834……..……

Programme: ……Msc in Data Analysis…… **Programme:**Msc in Data Analysis

Module: ……Research paper……..……

Lecturer: ……Bahman Honari……..……

Submission Due Date: …… …December 12th 2019……..……

Project Title: Analysing the impact of social media on online shopping platform sales by using sentiment analysis with text mining……..……

Word Count: …… **Page Count:** ……..…….

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: ……Xiyao xu……

Date: ……December 12th 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):	

Contents:

Catagory

Analysing the impact of social media on online shopping platform sales by using sentiment analysis with text mining

Xiyao Xu
x18107834

Abstract

With the continuous development of the shopping platform today, more and more people have entered the industry of online shopping platform merchants, and there are also countless consumers shopping on the Internet. Huge turnover conceals huge business opportunities. The purpose of this article is to use sentiment analysis technology to study people's attention to clothing features on social media and extract the keywords of clothing features as a bridge between seller sales and consumer to study whether this attention can be used by online merchants when they are planning the stock. While since people consider many factors when deciding whether to spend money on shopping, the study found that this attention is a very small factor to sellers.

1 Introduction

Weibo is the biggest social media platform in China, it has about 400 million users. People share their ideas every day on weibo, these ideas reacted the mood, the view and users' potential behaviours. To date, there are a lot of studies worked on the social media to find the value of these tweets. Raj et al. (2015) found social media plays an important and positive role in communicating with clients, Kim and Johnson (2016) describe the impact of environment/information stimulation on consumers' reaction. They said the brand related consumers' behavior, brand involvement and potential brand sales has important impact on brand.

Online shopping is a very popular shopping method, the platform used in this paper is Taobao, which is a part of Alibaba, the platform has 600 million active user and on November 11th (currently the most popular shopping festival in China), it has 250 billion RMB sales. Since the platform is a large data provider and a lot of famous brand companies and personal sellers have their shops on the platform, having the study based on the platform can be referenced. In the past there are also a lot of paper study the relationship between social media and online shopping platform.

1.1 The study on the relationship between social media and online shopping

In (Chung and Austria, 1973)'s research, he studies on 1. the level of satisfaction in using social media. 2. The attitude of media sales information and (3) the effectiveness of information about the value of online shopping. The satisfaction of social entertainment, including entertainment, information, and interaction, were tested as exogenous variables. People's attitudes to social media marketing messages and the value of online shopping are intrinsic variables. The results show that attitudes to social media marketing messages are closely related to social media interaction and information satisfaction instead of entertainment satisfaction. In addition, positive social media marketing messages increase the hedonic value of online shoppers.

In (Zhang et al., 2017)paper, he used a unique consumer panel data set that tracks people's shopping and social network website browsing and their online purchase activities within a year to study whether consumers' online shopping activities are related to their use of social networks, and if so, what is the law of this relationship. On the one hand, spending time on social networks can promote social discovery, which means that consumers "discover" or "accidentally discover" products by connecting with others. The authors found that greater usage of social networking sites was positively correlated with shopping activities. However, they also found a short-term negative relationship that caused online shopping activity to decrease immediately after social network usage increased for a period.

1.2 The problem faced by the sellers:

On the other hand, online shopping is now facing a important problem of warehousing management.

In (Patil and Divekar 2014)' s study, their research subjects are B2C e-commerce company or online retailers. They studied on the demand changes, Reverse logistics, inventory-free policies in inventory management and risks like sales loss, client loss, low satisfaction of clients. The study shows online retailers are having countless inventory management problems like demand changes, out of stocks, the number of inventory, the poor

quality goods and so on. The study thinks better inventory management plays an important role in improving customers' satisfaction.

Shen et al. (2016) says fashion cloth industry has become an important industry because of its great sales. In this industry, inventory needs is highly fluctuated because of market demands and since there are a lot of products, the management of fashion product inventory is very complex. In the study, they show the opinion that the inventory management, inventory sustainability and case study methods are the important study purpose in the future.

research question:

To what extent can social media affect the online sales in fashion cloth industry?

research objectives:

The purpose of the paper is according to use sentiment analysis and text mining, analysing the cloth features which are most discussed and most popular, at the same time, checking the sales on online shopping platform to see if the popularity of the cloth on social media is related with the sales on online shopping platform. Because when people have a purchase behaviour, they will consider many factors. The paper can help the online shopping platform's marketers to have a reference on their warehousing management.

hypotheses:

H0: the sales on the online shopping platform is not related with the popularity of the cloth on the social media.

H1: the sales on the online shopping platform is a little related with the popularity of the cloth on the social media

H2: the sales on the online shopping platform is very related with the popularity of the cloth on the social media

2 Related Work

Because social media and online shopping are very popular things, it is unsurprisingly that they attract the attention of many researchers, who have studied the relationship between social media and online shopping in various fields.

2.1 Research on relationship between social media and online shopping in apparel

Although there were a lot of researches study on the correlation between social media and online shopping, there were not so much in the field of clothing.

(Napompech, 2014) conducted the research on shoppers who made online purchases in the past. He used the way of making questionnaire to study on 412 respondents and found that most of them use Facebook to communicate with each other about clothing. The result was that the one of the factors that drive shoppers to shop is social networking.

(M. Kang et al., 2014)used Engel, Kollat, and Blackwell's model to conduct on 304 social network users, their purpose was to see whether the consumer's decision style was related to the usage of electronic word of mouth (eWOM) in social networking sites (SNS) to look for opinions and their attitude of the usage of SNS for shopping clothes online. As the result, among the consumer decision-making styles, creativity / fashion consciousness / decision style was the most important requirement for using eWOM to look for opinions. Originality / fashion consciousness, brand consciousness and price consciousness decision style directly affected the willingness of using online clothing shopping. The shortcoming of this article was its limited geographical location, the research range was not complete. Diversity of consumers (such as those had different decision-making styles) may had different reaction and result to different focused SNS.

(Darben and Li, 2012) tried to find out what steps an online social network can have influencing a consumer's purchasing decision when it affected a food retailer; and why these steps were affected by an online social network. The author conducted face-to-face and in-depth interviews on the phone with 11 interviewees, combined with theoretical framework analysis, and found that online social networks had different degrees of impact on each step of the food purchaser's consumer purchase decision process. When it came to purchasing decisions about food retailers, online social networks have the biggest impact on information search. The main reason was that Facebook's features bring convenience to consumers, so consumers spend more time, and Facebook's features allow consumers to interact with supermarkets and other consumers and see other consumers' Facebook pages on the supermarket and leave a comment. Consumers can express satisfaction or dissatisfaction with their experience to the company, product or service after consumption, or share their knowledge and perspectives on their online social networks and with others. Finally, as this study only covers consumer perceptions of online social networks, further research can look at online social networks from a company perspective.

(Nadeem et al., 2015) believed that consumers are increasingly searching, evaluating and purchasing products through social media and websites, but little is known about how these activities affect their trust, attitudes to online retail and their online shopping behaviour. So, they conducted an online survey of Gen Y Italian consumers who used Facebook to search various sites to buy clothing online. Validate the structure using confirmatory factor analysis and test hypotheses by using a structural equation model (SEM). The survey results confirmed that the quality of website services and consumers 'tendency to use Facebook for

online shopping directly affected consumers' trust in electronic retailers (whether they would buy or not).

2.2 Research on studying the relationship between social media and online shopping in other areas (mainly data mining)

(Erkan and Evans, 2018) tested and compared the impact of friend recommendations on social media and anonymous reviews on shopping sites under online purchase intent. They analysed the impact of the two platforms based on information adoption model (IAM) components and found that anonymous comments had a far greater impact on consumers 'online purchase intentions than friends' suggestions on social media. At the same time, contrary to expectations, information volume, information readiness, detailed information and dedicated information are factors that make shopping websites better than social media in terms of the impact of electronic word of mouth (eWOM).

(Gaikar and Marakarkandy, 2015) used sentiment analysis to make predictions on movie sales, he studied the impact of these reviews on viewers by analysing positive, negative, strongly positive, and strongly negative online reviews of movies. In the article, he collected the required data in the form of a questionnaire and performed sentiment analysis by using various theories. The author believed that in addition to the need to expand the research scale, the work of this research could also be carried out in other fields.

(Karthika et al., 2016) used data which collected from social media and transformed the data into smart data to help product owners analyse how people think about products The author used Apache Flume, Apache hive, and Apache HDFS to obtain and analyse the data, then analyse the tweets based on the filtered words, and then compare them to already available data sets. Based on that comparison, tweets are rated and classified as positive, negative, and neutral. The author used this method to help relevant people understand what consumers think, and consumers can also analyse these reviews.

(Dijkman et al., n.d.) also used sentiment analysis on Twitter to predict sales, and he analysed products that were less socially oriented than movies and books. He believes that for movies and books, the number of tweets is directly related to sales, but not for other products. When he classified the tweets, he divided the tweets into positive, neutral, and negative. He found that many people on Twitter had positive opinions, and he thought it was not surprising, because spam was often considered positive. He found that although existing research has shown a link between activity on Twitter and movie or book sales. But this rule does not work for all products. This research shows that the initial correlation between Twitter activity and products that generate less Twitter activity is much smaller. One can use the number of positive tweets to predict sales. However, the evidence only applies to the four countries studied, and only if the importance of the steadiness of the tweet is accepted, or it is in long-term forecasts (next five weeks or more). The conclusions of this article are too narrow, if others want to understand the situation in other regions, they need to study separately. In addition, the object of this study: the characteristics of clothing are small change cycles, and the requirements change frequently, so the conclusion of this article is not applicable. At the same time, the article also suggests that not all products have the same rules, which need to be considered with the characteristics of the product itself.

(Pai and Liu, 2018) using Twitter data and stock market value to determine car sales, he applied Bayesian algorithm for sentiment analysis, and finally found that using mixed data including social media sentiment analysis and stock market data can improve prediction accuracy. In addition, due to Twitter keywords will significantly affect the accuracy of search results and has an impact on the accuracy of predictions, so more systematic techniques for selecting the right keywords from Twitter may become the direction of future research. Another possible direction for future research is to use other social media data, such as Facebook and YouTube, to predict car sales. Finally, Twitter's geographic information collection may be an important issue for future research to improve tweet analysis.

2.3 Sentiment Analysis in Short Text Analysis of Social Media Types:

At present, the research work of text sentiment analysis is mainly divided into semantic-based sentiment dictionary method and machine learning-based method.

(Wang et al., 2013) proposed a Chinese bag of words model based on the dependency grammar of Weibo sentences. Then, they calculated the sentiment polarity score for each opinion and weighted summed the sentiment evaluation for each sentence. Confidence values for polarity scores of the sentences are also defined. With it, we can extract sentences with high confidence as annotated data, which can guide further analysis. They applied the model to summation evaluation and semi-supervised methods. Their experiments on Chinese sentiment analysis on the NLP & CC 2012 dataset prove the effectiveness of the method.

(Shuoqiu and Chaojun, 2019) used TextRank and word2vec models (combining a general sentiment dictionary and an online course review corpus) to identify and extract sentiment words. Then, a tag propagation algorithm is applied to distinguish the sentiment polarity of sentiment words, thereby constructing a sentiment dictionary for online course reviews. The experimental results show that this method is an accurate and effective method to achieve sentiment classification in online courses.

(Sun et al., 2019) used automate analysis to the sentiment of Tibetan microblog text, the sentiment in text was identified, and an analysis method based on Tibetan sentiment dictionary and rules was proposed, and sentiment characteristics were expressed in it. The experimental results show that the accuracy of emotion recognition based on the Tibetan sentiment dictionary is 78.6%, which provides a basis for establishing a high-precision Tibetan text sentiment classification system.

(Xue et al., 2014) thought that analyzing the hidden emotions in this information can benefit online marketing, brand promotion, customer relationship management, and public opinion monitoring. This paper proposed a new model based on the Word2vec tool to build a sentiment dictionary based on our semantically oriented point-to-likeness distance (SO-SD) model. Then they used sentiment dictionary to get the sentiment of Weibo information.

(Zhang et al., 2018) proposed a kind of sentiment dictionary-based sentiment analysis method for Chinese Weibo text to better support the work of network regulators. First, the emotion dictionary can be extended by the extraction and construction of degree adverb dictionary, network word dictionary, negative word dictionary and other related dictionaries. Secondly, the emotional value of Weibo text can be obtained by calculating the weight. Finally, Weibo

texts on a certain topic can be divided into positive, negative and neutral. Experimental results show the effectiveness of the method.

(Shuoqiu and Chaojun, 2019) thought the construction of sentiment dictionary was one of the most important and basic tasks in the field of text mining. At present, there was no universal and complete sentiment dictionary in the field of text mining. This paper proposed a method of constructing an emotional dictionary based on Word2vec. This method integrated the existing sentiment dictionary first, then used the SO-PMI algorithm to determine the sentiment polarity of unrecorded network words. Word2vec was used to correct the results. The modified network words were added to the emotional dictionary to complete the construction of the emotional dictionary. Finally, in order to verify the effectiveness of the proposed method, we used the constructed sentiment dictionary to distinguish the sentiment polarity of the text. The experimental results showed that the microblog sentiment dictionary constructed by this method has high accuracy and reliability.

(Alemneh et al., 2019) provided an algorithm for constructing Amharic sentiment dictionary. The proposed method relies on the Amharic-English dictionary to transfer emotional tags from one language (such as English) to a language with limited resources (such as Amharic). Used a bilingual / monolingual dictionary as a bridge, two Amharic sentiment dictionaries would be automatically generated, the first one was based on the SO-CAL polarity dictionary and the second one was based on SentiWordNet 3.0. For each Amharic word, the algorithm found the meaning of the corresponded English word. For these English words, the emotional information is searched from the above emotional dictionary.

2.4 TF-IDF:

(Ramos, n.d.)checked the results of applying the term frequency inverse document frequency (TF-IDF) to determine which words in the document corpus might be more suitable for use in queries. TF-IDF calculates the value of each word in a document by inversely proportion of the frequency of words in a document and the percentage of documents in which that word appears. Words with a high TF-IDF value has a relationship with their appearance in the document, indicating that if the word appears in a query, the user may be interested in the document. We provide evidence that this simple algorithm can effectively classify related words that can enhance query retrieval capabilities.

(Qaiser and Ali, 2018)checked the relevance of keywords to the corpus document. The main limitation of TF-IDF is that even the changes slightly, the algorithm still cannot recognize words, such as the tense. Another limitation of TF-IDF is that it cannot check the semantics of the text in the document. TF-IDF algorithm is easy to realize and has a strong function but its limitation still cannot be neglected.

(Zhang and Ge, 2019) believed that many improved methods of the TF-IDF algorithm have achieved good results, but they are not effective for desensitized data or encrypted data. In this paper, we proposed a new concept, which was the strength of class discrimination, and used it to improve TF-IDF. The new algorithm is called TF-IDF-ρ, and the author used it to represent desensitized data for text classification. It was worth to be mentioned that the experimental results of the testing set showed its effectiveness. Finally, experiments performed on a desensitization tester showed that compared with traditional TF-IDF, TF-IDF-ρ can increase the F1 measurement by up to 4.07%.

2.5 Conclusion：

Through the literature collected above, it can be found that researchers are constantly improving these two methods. In sentiment analysis, scholars have found that add more words related to the field of study into the dictionary can improve the accuracy of the results better. At the same time, since many new words are constantly being created, when using basic words, it is also necessary to add these new words. TF-IDF, as a keyword extraction algorithm that has been used, researchers are no longer satisfy accuracy it had in the past, and they are thinking about using it in combination with other algorithms to improve his accuracy. On the other hand, when reading the article (Dijkman et al., n.d.), he found that not all products are the same to movie or book which sales are related to the effect social media. This broke my inherent impression of the impact of social media on clothing sales. On social media, there are so many recommendations about clothing and the sharing of clothing, but it is unknown whether the more you talk about the cloth, the better it sales well online, if the warehouse management is not proper, or if the sellers just follow the trend and recommendations in social media. The purchase is a great loss to individual online stores.

3 Research Methodology

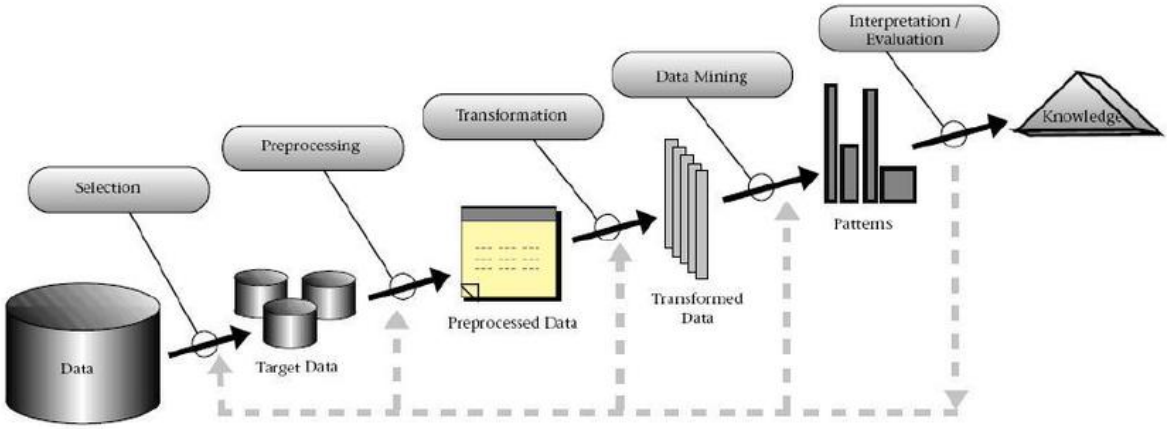
KDD, SEMMA, and CRISPDM are the three most widely used methodologies, so it is more appropriate to associate their role with the research purpose and process in this article to find the most appropriate one.

KDD：(Fayyad et al, 1996) said that the KDD process is a process of extraction which use the DM method. According to the specifications of the metrics and thresholds, which database to use as knowledge, and any required database preprocessing, sub-sampling, and conversion, there are five stages in KDD

SEMMA was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Evaluate, and refers to the process of conducting a data mining project. (Azevedo and Santos, n.d.)

CRISP-DM stands for the data standard flow in Cross-industry. It is consisted by a six steps ‘circulation: Business understanding，Data understanding，Data preparation，Modeling， Evaluation and Deployment.

After (KDD-CRISP-SEMMA.pdf)’s comparison, he thought SEMMA process can be seen as a training pattern of the KDD, and the difference between KDD and CRISP-DM is the business requirement.
Since in the paper, there is no directly business purpose and SEMMA is used for the users of SAS software, I think KDD is the most proper to me.



An overview of the steps that compose the knowledge discovery in databases (Fayyad et al. 1996)

The Knowledge Discovery in Databases (KDD) process (Fayyad et al., 1996).

3.1 Data collection

The data in this study was extracted by using data acquisition software, and 20,000+ pieces of data were extracted from taobao (online shopping platform) and weibo (social platform) respectively using keywords. Due to the unique limitations of the textile and apparel market in the apparel industry, such as fluctuations in demand, peak sales seasons, a large variety of products, and short life cycles (Thomassey, 2010), the data extracted are data for the past three months. The keywords are related to clothing and dressing: wearing, women's clothing etc.. Since the data of the Weibo platform is always frequently refreshed, and the number of data has been kept at the maximum of 50 pages, so the extracted data can be guaranteed to be the latest, however, because the product data of the taobao platform will not be the same as weibo, its old data will not automatically disappear. If the seller did not removed the data, the data will still be extracted, which has been listed for a long time, such as the spring of 2019 clothes, since these clothes have been on the shelf for a long time, the sales volume has been accumulated a lot, and it is easy to cause deviations in the data. Therefore, the keyword extraction on taobao side will bring autumn and winter words.

The original dataset sample of Taobao (There are 14 columns):

Shop name	location	Name	Price	Payment	Link	Picture address
Goods link	Goods ID	Shop type	Current page' website	Current time	Page	Advertisment?

The original dataset of weibo(There are 14 columns):

User_name	Twee t	Favorite	Retweet	comments	Thumbs	Time	From	Website link	Twee t link

3.2 pre-processing

In the data preprocessing, unnecessary columns are deleted, at the same time, the null values in different columns are processed differently. In Weibo data, the null values of tweets are retained, and if values of likes and retweets are null, replace it with 0. In the taobao data, the data of sales volume is filled with 0 where the value is empty. In the extracted original data, the original data of the number of payers in taobao shows that xxx people pay. Delete it and only keep numbers to bring convenience to the later study. At the same time, due to the large crowd of social media, the related tweets from famous stars are deleted, because the users who liked them liked the idols instead of the clothing itself, and these tweets often have many thumbs and retweets. The number of likes and retweets is easy to cause large deviations in subsequent data analysis. At the same time, advertisements and repeated tweets are also deleted.

3.3 Feature extraction

In feature extraction, firstly use the word segmentation dictionary to segment the sentence, cut each sentence into daily vocabulary, and then use stop word list to remove the stop word. The reset word refers to the phrase that is usually repeated but not do not have meaning. Give the scores to each degree words, for example: very, a little, almost, etc.

3.4 Transformation

In order to perform sentiment analysis on Weibo tweets later, copy the tweets to a txt file. In this study, the object format of sentiment analysis is text. Make sure that the data format to be analyzed is text format and confirm that its Unicode Transformation Format is correct and executable.

3.5 Data mining

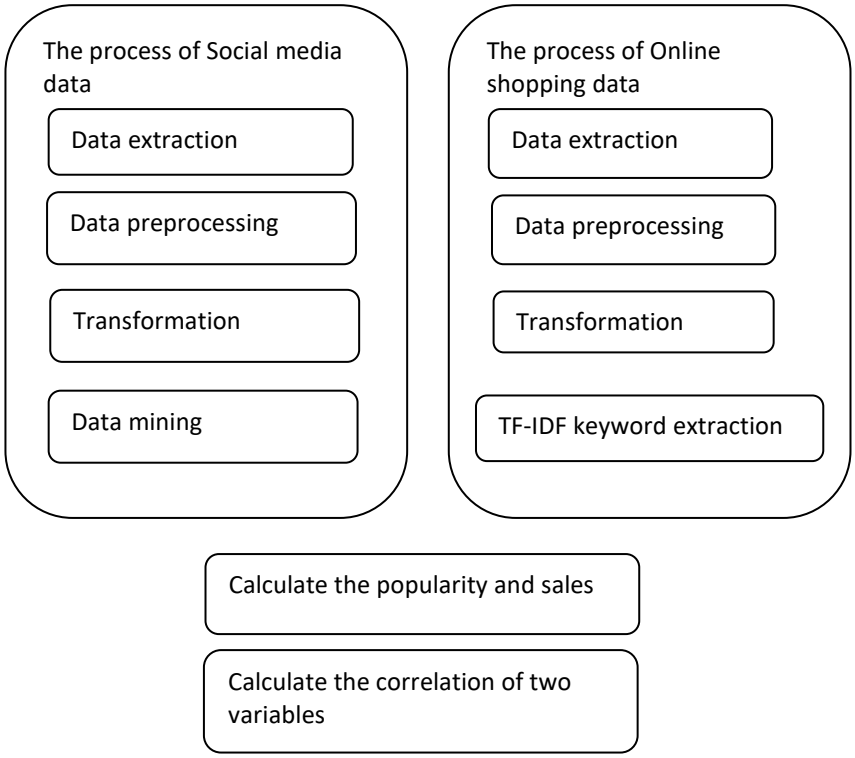
In this study, the part of data mining uses text mining for sentiment analysis. The method of this article is to use an emotional dictionary to score tweets, and according to the constructed emotional dictionary, extract the emotional words of the text to be analyzed, and then calculate their emotional tendencies. The affective dictionary generally includes: affective words and degree words. Users use emotion words to express their attitudes, such as: like, hate, etc..; while users use degree words to express their strength, such as: very, average, etc. Different emotional strengths have different scores. First, the sentence is segmented, and the words in the segmentation result are matched according to the emotional dictionary. The sum of the scores of each word can be divided into positive and negative numbers. In the results of this article, sentiment score greater than 0 are recorded as 1 (positive emotion), results equal to 0 are recorded as 0 (neutral emotion), and results less than 0 are recorded as -1 (negative emotion).

3.6 Interpretation/Evaluation

Firstly, using TF-IDF to extract the keywords related to the characteristics of the clothing from the product name of taobao, and use the keywords to do the calculation. In social platforms, use the result of the emotional score as the basis to calculate the popularity of each feature. Popularity is equal to (Weibo Likes + Retweets) * Emotional score (which is 1,0 or -1). At the same time, each Weibo's (Weibo Likes + Retweets) plus one point, one point means the thumb from the person who published the tweet.

Popularity = （Weibo Likes + Retweets+1）* Emotional score (which is 1,0 or -1)
After the calculation,
Use the keywords extracted by TF-IDF to categorize the products name in the data set of the online shopping platform and calculate the total sales of each category. Finally, we get two columns of popularity and sales, and use spss to remove the extreme values in the data to calculate their correlation. Correlation calculation I used spss to do linear correlation, univariate linear regression and curve regression.

4 Design Specification

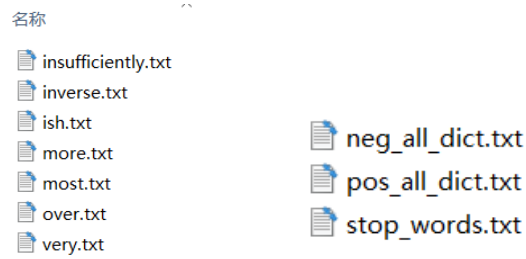


5 Implementation

5.1 Sentiment analysis：

5.1.1 Data pre-processing

Domain-oriented sentiment dictionary is widely used for fine-grained sentiment analysis of reviews. (Du et al., 2010) The sentiment analysis uses python. In the process of using emotion dictionaries to analyze emotions, external emotion dictionaries are necessary. A total of 10 external dictionaries were used. In the previous literature survey, it was proved that adding new common words into the word segmentation and using words which are unique to the field of research into the dictionary can improve the accuracy of the analysis, so in addition to these 10 lists, I also collected some dictionary in e-commerce field and some popular words and add them into degree list and emotion list.



The seven files on the left are words that indicate degree, the three files on the right are positive, negative words, and stop words are stop words. In the dictionary, each word is associated with a series of distributions of human emotions. (Rao et al., 2014)

When doing text processing, in order to divide the sentence into words that we are familiar with, we should import the dictionary inside jieba package and add the segment words we collected first, then divide the sentences according to the dictionary. Jieba generated a directed acyclic graph, in which based on the Trie tree structure (a dictionary tree that uses the common variables of strings to save query time), all Chinese characters in the analyzed sentence can be turned into words. Then, a dynamic programming algorithm is used to find the maximum probability path, and the maximum segmentation form is obtained according to the word frequency to achieve Chinese text segmentation. You can recognize the words and expressions in Weibo posts and comments by defying word segmentation. （wang2019）

For example: "我好开心（I'm so happy）" will be divided into "I (I) 好 (so) 快乐" (happy) ". The English translation currently seen in this sentence will be a bit weird, because ‘am ‘ is missing. Since ‘am’ has no practical meaning in this sentence, and that is the stop word going to be deleted later.

The most common and no meaning words in English are called stop words. Stop words are language-specific function words and contain no information. It may be of the following types, such as pronouns, prepositions, conjunctions.(Ramasubramanian and Ramya, 2013)
In this study, using the stop-words dictionaries produced by Harbin Institute of Technology, baidu stop words and Sichuan University machine learning lab, which are currently mainstream and widely used stop word lists.

Read the words with six weights, return the list according to the requirements, and get a list with six weights. These six lists represent six different degrees, most, very, more, ish, insufficiently and inverse.

5.1.2 Sentiment analysis

The process of the code:

1. In python, the code read the emotion dictionaries and give score to each degree. The degree of most is 2, very is 1.75, more is 1.5, ish is 1.2, insufficient is 0.5 and inverse is -1
2. Calculating the score of each sentence:
3. Cut the sentence by sentence, so that the sentence can be calculated one by one.
4. When there is a positive emotion word, +1, if there is a negative emotion word, -1.
5. If there is an exclamation mark, score+2 and break the circle.
6. After breaking the circle, calculating the final score by cyclically accumulate the segments.
7. Output the result of the test.

Then calculate the priority:

Popularity = （Weibo Likes + Retweets+1）* Emotional score (which is 1,0 or -1)

5.2 TF-IDF **keyword extraction**

TF-IDF (short for term frequency–inverse document frequency) is a statistical method used to evaluate the importance of a word to a file set or a file in a corpus. The importance of a word increases proportionally with the number of times it appears in the file, but at the same time decreases inversely with the frequency of its appearance in the corpus.
The main idea of TFIDF is: If a word or phrase appears frequently in one article and has a high frequency of TF, and rarely appears in other articles, it is considered that the word or phrase has a good class discrimination ability and is suitable for classification . TFIDF is actually: TF * IDF, TF Term Frequency, IDF Inverse document frequency. (Menaka and Radha, 2013).

In the paper, TF-IDF is used for keyword extraction. The process of extraction is as following:
1. Read stop words list
2. Load the name of the goods, use Jieba package to segment and delete the stop words.
3. Acquisition the frequency of the words in the product's name
4. Calculate the times the word appeared in all the products' name.
5. TF-IDF value calculation

5.3 SPSS analysis

The most commonly used techniques for studying the relationship between two quantitative variables are correlation analysis and linear regression. (Bewick et al., 2003)
SPSS software was used in the last step of the correlation analysis. SPSS was used for linear regression analysis and correlation analysis. Correlation analysis can be used to find if two variables are correlated but it cannot read the causality of these two variables, and it cannot tell how these variables are related with each other.

Linear Logistic Regression:
Binomial (or binary) logistic regression is applied when the dependent variable only has two different possible values. (reference:学校 ppt)It has independent variable and dependent variable, when there is a math's formula Y=f(X) and it is a linear formula it can be called linear regression. If the regression formula is reliable, how big the deviation is needed to be examined by Significance test and error calculation.

$$E(Y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

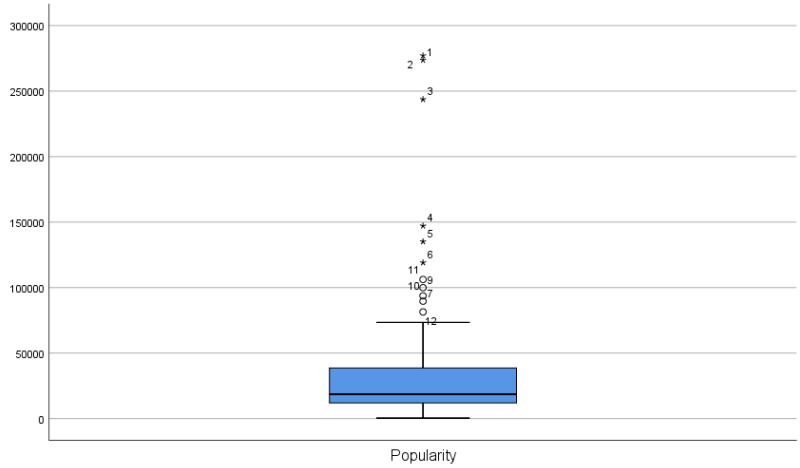
6 Evaluation

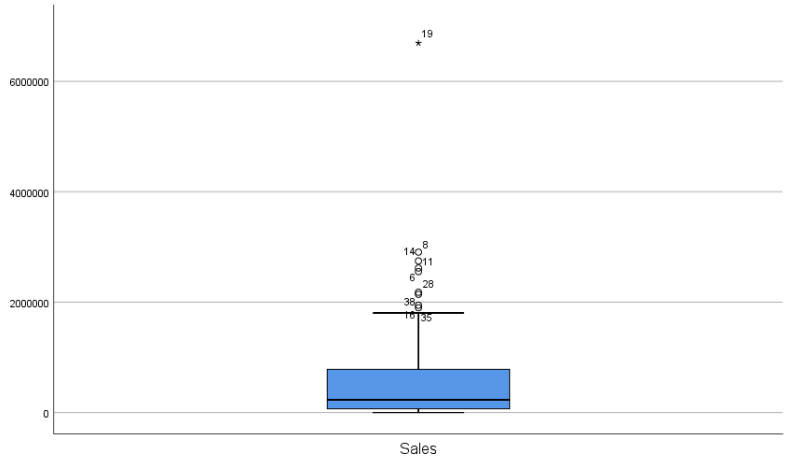
6.1 Requirement check:

6.1.1 General check：

1. Outlier check:
   An outlier is a set of measured values whose deviation from the expected value exceeds the two-way standard deviation, and whose deviation from the deviation exceeds three standard deviations, which is called a highly abnormal outlier. (Bremer, 1995)

Using boxplot to test ouliers:





As we can see, there are many outliers, outliers affect the process in the analysis significantly, (for example: average and standard deviation), it makes the result wrongly estimated. So, The results of data analysis depend greatly on how missing and outliers are handled.

The process of outliers usually we can choose delete it, use Robust estimation method and Winsoriz it. (Kwak and Kim, 2017)

Sometimes, discarding the outliers can improve the quality of analysis(Last and Kandel, n.d.),while considering outliers are also a part of the research data, it shows the factors, it is not reasonable to neglect it directly, on the other hand, in this paper, there are about 10 outliers, deleting them will probably affect the result greatly, so I tend to use winsorizaion and robust estimation to deal with outliers.

6.1.2 Check for Correlation analysis:

1. Normal distribution:

Normal distribution for sales:

Descriptive Statistics

	N	Minimum	Maximum	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Sales	79	112	6692986	3.406	.271	16.198	.535
Valid N (listwise)	79						

In the descriptive part of the result output, a basic statistical description of the variable sales is made, and the skewness of the distribution is 3.419 (standard error 0.271). Z-score =

3.419/0.271，Kurtosis value = 16.198（Standard error 0.535），Z-score = 16.198/.535 = 1.036。Skewness and kurtosis values are both≈0，Z-score of both are greater than ± 1.96, and it can be considered that the data do not obey the normal distribution.

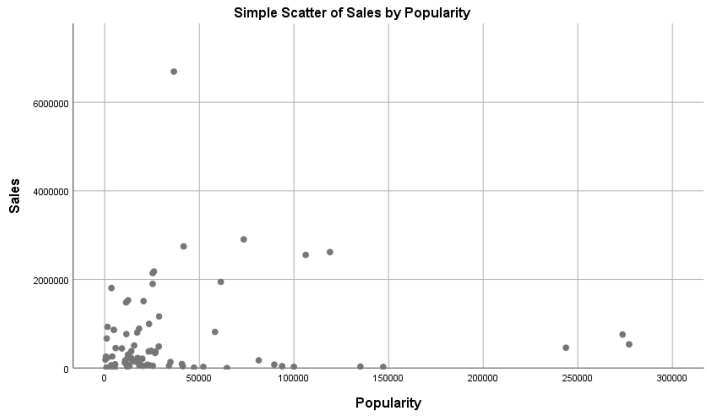
Normal distribution for popularity:

Descriptive Statistics

	N	Minimum	Maximum	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Popularity	79	387	277092	2.953	.271	9.418	.535
Valid N (listwise)	79						

In the Descriptives section of the result output, a basic statistical description of the variable popularity is given, and the skewness of its distribution 2.953 is given.（Standard error 0.271），Z-score = 2.953/0.271，Kurtosis value = 9.418（Standard error = 0.535），Z-score = 9.418/.535 = 1.036。Skewness and kurtosis values are both≈0，Z-score of both are greater than ± 1.96, and it can be considered that the data do not obey the normal distribution.

2. Whether the data is linear: Use a scatter plot to check:



It can be seen from the above tests that the data are neither normally distributed nor linearly related.

Although many people can now use box_cox to change the data to normal distribution. It is a technology used to reduce anomalies such as non-additivity, non-normality and heteroscedasticity. (Sakia, 1992)

， But the domain and scope of the transformation in this method is usually bounded (Kemp, 1996), and I try to not to change the data as much as possible, so I choose to use non-parametric test. Non-parametric test is a method do not need the assumption of population distributed.(Hoeffding, 1948) . Non-parametric test is applied in a lot of ways, it is easy but the parametric test has higher efficiency. It is better to use parametric test is the data satify the assumption. According to the data which is not normally distributed and the purpose in this

study is to check the relationship between two variables, Spearman's non-parametric correlation test is more suitable for this study. (Artusi et al., 2002)

6.1.3 Check for Linear Logistic Regression:

Although Linear Logistic Regression do not need normal distribution, it still has the premise that whether there is a linear relationship between the continuous independent variable and the logit transformed value of the dependent variable. (Wright, 1995)

As can be seen from the figure, the two variables do not have a linear relationship. So the logistic regression method is also not applicable with this data.

6.2 Spear Correlation analysis / outliers winsorization

R is used to do the winsorization. (Hoo et al., 2002) is a way of handling outliers. It can replace the part of value which exceed the specific percentage range with a number at specific percentage position in the dataset.

```
7  data
8
9  length(data)
0  summary(data)
1  benth <-783632 + 1.5*IQR(data)
2  benth
3
4  data[data > benth]
5  data[data > benth] <-benth
6  data
7  summary(data)
8  boxplot(data)
9
```

It replace the number which is bigger than 75% with a number which is much closer to the normal value.

Correlations

			popularity	sales
Spearman's rho	popularity	Correlation Coefficient	1.000	.048
		Sig. (2-tailed)	.	.675

		N	79	79
	sales	Correlation Coefficient	.048	1.000
		Sig. (2-tailed)	.675	.
		N	79	79

From the picture we can see the correlation coefficient is small which means the relationship between these two variables is weak.
And the possibility of no relationship between these two variables is high.

6.3 Spear Correlation analysis / robust estimation

Robust estimation is realized by R, robust regression is used in the test. It can identify the extreme words and make their impact on the estimates to be as small as possible. The number of weights assigned to each observation in robust regression is controlled by a special curve called the influence function.(Rousseeuw and Leroy, 2005)

	Sales	resid	weight
6	2617734	2322591.432	0.00000000
8	2905154	2634178.081	0.00000000
11	2553575	2265245.717	0.00000000
14	2745657	2491512.874	0.00000000
16	1944018	1679463.658	0.00000000
19	6692986	6441592.843	0.00000000
28	2142190	1896730.242	0.00000000
35	1898800	1653341.832	0.00000000
38	2178952	1933103.184	0.00000000
71	1805145	1571227.119	0.00000000
45	1529637	1291036.642	0.02853598

As the result, the numbers which are very high are calculated as no weight in the algorithm.

Correlations

		popularity	weighted_sales
popularity	Pearson Correlation	1	.100
	Sig. (2-tailed)		.410
	N	70	70
weighted_sales	Pearson Correlation	.100	1
	Sig. (2-tailed)	.410	
	N	70	70

Although the result is better than the previous one, it is still not a good result. The relationship between two variables is low and the possibility of no relationship between two variables is still higher than it should be.

6.4 Discussion

The result is not the same with what I thought before the study. In this experiment, the collection of data seems to be the key. If the amount of data is big, this data may become a normal distribution, just like some data from 100 people is not normally distributed but if the data come from 100 million people, it might be normal distributed. What's more, in this experiment, if the sentiment analysis method can be used based on the comparison of sentiment dictionary methods and machine learning methods, the result might be better, and at the same time, we cannot deny the possibility that the platform data is fake. Because the sellers might buy their own goods to make the sales looks better. Compared with other data, some outliers are extremely high. In this survey, according to the result literature review research, I added domain-related dictionaries and popular words into the dictionary for sentiment analysis, it helps me to analyze more accurately. The results of this study are matched with the results from (Dijkman et al., n.d.). Not all the sales in different areas are related to its social media hot topic. To the cloth industry, there are some features that people naturally need, so they will not post it on social media, such as keeping warm and thickening. However, since the survey period is autumn and winter, the sales of clothing with these characteristics will increase, but this will not be shown on social media

7 Conclusion and Future Work

The purpose of this thesis is to study the impact of social media on sales of online sales platforms. From the results, the relationship between the popularity of a clothing feature on social media and its sales on online sales platforms is not significant. Instead of paying attention to the current fashion styles, focusing on the necessary demand from the customers (like the warmth just mentioned) and the price will have lower risks. In addition, from the results, people like a kind of clothes does not mean they will go to buy it. From the perspective of clothing and dressing, this may be related to everyone's opinion on consume, whether the clothing is suitable. Some consumers like the cloth, so they will give a thumb for it on social platforms, but they will not buy it if it is not suitable to them. Similarly, some necessary clothes will be bought if they are not published on the public platform. The result shows the widely spread topic of a cloth does not mean people will really buy it. In future research, investigators can conduct questionnaires to ensure the quality of the data, but more people are needed to conduct to improve the applicability of the data. In addition, they can also focus on necessary needed clothes to help the sellers to defeat their competitors. Also, in future surveys, if the survey's target is a fashion cloth, when looking for the data, it is necessary to separate the fashion cloth with the necessary needed clothes to improve the reliability of the data.

References

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.
Menaka, S. and Radha, N. (2013). Text classification using keyword extraction technique, International Journal of Advanced Research in Computer Science and Software Engineering 3(12).
Alemneh, G.N., Rauber, A., Atnafu, S., 2019. Dictionary Based Amharic Sentiment Lexicon Generation, in: Mekuria, F., Nigussie, E., Tegegne, T. (Eds.), Information and Communication Technology for Development for Africa, Communications in Computer and Information Science. Springer International Publishing, Cham, pp. 311–326. https://doi.org/10.1007/978-3-030-26630-1_27
Artusi, R., Verderio, P., Marubini, E., 2002. Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval. Int. J. Biol. Markers 17, 148–151. https://doi.org/10.1177/172460080201700213
Azevedo, A., Santos, M.F., n.d. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW 6.
Bewick, V., Cheek, L., Ball, J., 2003. [No title found]. Crit. Care 7, 451. https://doi.org/10.1186/cc2401
Bremer, R., 1995. Outliers In Statistical Data. Technometrics 37, 117–118. https://doi.org/10.1080/00401706.1995.10485900
Chung, C., Austria, K., 1973. Social Media Gratification and Attitude toward Social Media Marketing Messages: A Study of the Effect of Social Media Marketing Messages on Online Shopping Value 7.
Dijkman, R., Ipeirotis, P., Aertsen, F., van Helden, R., n.d. USING TWITTER TO PREDICT SALES: A CASE STUDY 14.
Du, W., Tan, S., Cheng, X., Yun, X., 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10. Presented at the the third ACM international conference, ACM Press, New York, New York, USA, p. 111. https://doi.org/10.1145/1718487.1718502
Erkan, I., Evans, C., 2018. Social media or shopping websites? The influence of eWOM on consumers' online purchase intentions. J. Mark. Commun. 24, 617–632. https://doi.org/10.1080/13527266.2016.1184706
Gaikar, D., Marakarkandy, B., 2015. Product Sales Prediction Based on Sentiment Analysis Using Twitter Data 6, 11.
Hoeffding, W., 1948. A Non-Parametric Test of Independence. Ann. Math. Stat. 19, 546–557.
Hoo, K.A., Tvarlapati, K.J., Piovoso, M.J., Hajare, R., 2002. A method of robust multivariate outlier replacement. Comput. Chem. Eng. 26, 17–39. https://doi.org/10.1016/S0098-1354(01)00734-7
Karthika, I., Gokulraj, P., Saravanan, S., 2016. Prediction of sales using Big data analytics. N 12, 4.
Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: management of missing values and outliers. Korean J. Anesthesiol. 70, 407–411. https://doi.org/10.4097/kjae.2017.70.4.407
Last, M., Kandel, A., n.d. Automated Detection of Outliers in Real-World Data 10.
M. Kang, J.-Y., K.P. Johnson, K., Wu, J., 2014. Consumer style inventory and intent to social shop online for apparel using social networking sites. J. Fash. Mark. Manag. Int. J. 18, 301–320. https://doi.org/10.1108/JFMM-09-2012-0057
Nadeem, W., Andreini, D., Salo, J., Laukkanen, T., 2015. Engaging consumers online through websites and social media: A gender study of Italian Generation Y clothing

consumers. Int. J. Inf. Manag. 35, 432–442. https://doi.org/10.1016/j.ijinfomgt.2015.04.008
Pai, P.-F., Liu, C.-H., 2018. Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values. IEEE Access 6, 57655–57662. https://doi.org/10.1109/ACCESS.2018.2873730
Qaiser, S., Ali, R., 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. Int. J. Comput. Appl. 181, 25–29. https://doi.org/10.5120/ijca2018917395
Ramasubramanian, C., Ramya, R., 2013. Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm 2, 3.
Ramos, J., n.d. Using TF-IDF to Determine Word Relevance in Document Queries 4.
Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M., 2014. Building emotional dictionary for sentiment analysis of online news. World Wide Web 17, 723–742. https://doi.org/10.1007/s11280-013-0221-9
Rousseeuw, P.J., Leroy, A.M., 2005. Robust Regression and Outlier Detection. John Wiley & Sons.
Sakia, R.M., 1992. The Box-Cox Transformation Technique: A Review. J. R. Stat. Soc. Ser. Stat. 41, 169–178. https://doi.org/10.2307/2348250
Shuoqiu, Y., Chaojun, X., 2019. Research on Constructing Sentiment Dictionary of Online Course Reviews Based on Multi-source Combination, in: Proceedings of the 2019 2Nd International Conference on Data Science and Information Technology, DSIT 2019. ACM, New York, NY, USA, pp. 71–76. https://doi.org/10.1145/3352411.3352452
Sun, B., Tian, F., Jia, M., 2019. Emotion recognition method of Tibetan micro-blog text based on sentiment dictionary. J. Phys. Conf. Ser. 1314, 012182. https://doi.org/10.1088/1742-6596/1314/1/012182
Thomassey, S., 2010. Sales forecasts in clothing industry: The key success factor of the supply chain management. Int. J. Prod. Econ. 128, 470–483.
Wang, J., Song, D., Liao, L., Zou, W., Yan, X., Su, Y., 2013. The Chinese Bag-of-Opinions Method for Hot-Topic-Oriented Sentiment Analysis on Weibo, in: Li, J., Qi, G., Zhao, D., Nejdl, W., Zheng, H.-T. (Eds.), Semantic Web and Web Science, Springer Proceedings in Complexity. Springer, New York, NY, pp. 357–367. https://doi.org/10.1007/978-1-4614-6880-6_31
Wright, R.E., 1995. Logistic regression, in: Reading and Understanding Multivariate Statistics. American Psychological Association, Washington, DC, US, pp. 217–244.
Xue, B., Fu, C., Shaobin, Z., 2014. A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec, in: 2014 IEEE International Congress on Big Data. Presented at the 2014 IEEE International Congress on Big Data (BigData Congress), IEEE, Anchorage, AK, USA, pp. 358–363. https://doi.org/10.1109/BigData.Congress.2014.59
Zhang, S., Wei, Z., Wang, Y., Liao, T., 2018. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. Future Gener. Comput. Syst. 81, 395–403. https://doi.org/10.1016/j.future.2017.09.048
Zhang, T., Ge, S.S., 2019. An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data, in: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, ICIAI 2019. ACM, New York, NY, USA, pp. 39–44. https://doi.org/10.1145/3319921.3319924
Zhang, Y., Trusov, M., Stephen, A.T., Jamal, Z., 2017. Online Shopping and Social Media: Friends or Foes? J. Mark. 81, 24–41.