

**Machine translation**

**Lab Report**

**Title：**The analysis of emotion based on HowNet emotion dictionary

**Class：** 2004

**Number：**  2001760

**Name：** LiJiahao

**Date：** 2020年12月19日

**The analysis of emotion based on HowNet emotion dictionary**

**LiJiahao 2001760 2004**

**Abstract**：The ubiquity of social networks has brought an explosion of emotions and opinions.The monitoring of mental health based on emotion and the control of public opinion make emotion analysis of great significance in modern society.Emotion classification is an important branch of natural language processing. It is a method to analyze, process, conclude and reason subjective texts with emotion color, and use some emotion score indicators to quantify qualitative data.Based on the research background, purpose and significance of emotion analysis, this paper clarifies the research content and the techniques used, gives the detailed algorithm design and implementation steps, and discusses the experimental results and proposes the direction of improvement.

**Keyword：**Emotional analysis;Emotional dictionary; Machine learning

**Catalog**

[1.Introduction 1](#_Toc13032)

[2.Research contents 2](#_Toc13458)

[2.1 The choice of an emotional dictionary 2](#_Toc23348)

[2.2 Data set selection 3](#_Toc18335)

[2.3 Data Processing 3](#_Toc30679)

[2.4 Calculate emotional scores 3](#_Toc887)

[3.Platforms and Key technologies 3](#_Toc21931)

[3.1 Platforms 3](#_Toc25095)

[3.2 Chinese clause technology 4](#_Toc19835)

[3.3 Chinses words divided syncopation technology 4](#_Toc19763)

[4.Design and Implementation 5](#_Toc27838)

[4.1 Overall Design 5](#_Toc17207)

[4.2 Training data acquisition 6](#_Toc11506)

[4.3 Clause design 6](#_Toc17149)

[4.4 Participles design 7](#_Toc30490)

[4.4 Statistical emotion word and Degree word processing 8](#_Toc26679)

[4.5 Score calculation 8](#_Toc23536)

[5.Experimental results and Analysis 8](#_Toc7492)

[5.1 Environment configuration 8](#_Toc28350)

[5.2 Dictionary of degree words 9](#_Toc14621)

[5.3 Chinese clause result 10](#_Toc24583)

[5.4 Chinese participle result 11](#_Toc20968)

[5.5 Final result 12](#_Toc7774)

[6.Interpretation of result 14](#_Toc4318)

[Reference： 14](#_Toc14117)

# 1.Introduction

With the popularization of computer applications and the development of Internet technology, especially after the emergence of Web2.0, Internet applications such as blog, email, forum, chat room and online comment are everywhere. More and more users post words, pictures or emoticons with opinions and emotions through the Internet.In these information with people's viewpoints, a large amount of useful and rich information is extracted through cleaning, feature extraction and other operations. Among these massive information, the text information data volume is the largest and easy to obtain. Therefore, text-based emotion analysis has become a hot and cutting-edge research field of artificial intelligence.

Since it was proposed by Bo Pang in 2002, textual emotion analysis has attracted wide attention and made some achievements, especially in the analysis of online comments' emotion tendency, with the accuracy reaching more than 90%.Emotional analysis also gradually experienced from coarse to fine, from simple to complex process.Meanwhile, textual emotion analysis has also expanded from simply understanding users' preferences to providing decision support, online public opinion risk analysis and information prediction for individuals, enterprises and institutions.

Emotion analysis is one of the most active research fields in common natural language processing (NLP). It is a method to analyze, process, induce and reason subjective texts with emotion color, and use some score indicators to quantify qualitative data.

There are two mainstream emotion analysis methods: one is dictionary-based method and the other is machine learning algorithm method.The dictionary-based method mainly makes a series of emotion dictionaries and rules, disintegrates the text, extracts the keywords, calculates the emotion value, and finally uses the emotion value as the basis of the emotional tendency of the text.However, the text emotion analysis based on machine learning requires a large number of manually annotated corpus as training sets, which can realize the classification of emotions by extracting text features and constructing a classifier.

In recent years, with the application of deep learning, deep learning has made great breakthroughs in the traditional feature selection and extraction framework, In natural language processing, biomedical analysis, remote sensing image interpretation and many other fields of more and more important influence, and in the field of computer vision and speech recognition has made revolutionary breakthroughs.Especially, it is widely used in the field of NLP, based on abstract features of deep learning, it can avoid the work of manual extraction of features. Moreover, it simulates the connection between words through Word2Vec technology, and has the function of local feature abstract painting and memory, which has great advantages in emotion classification.

Since machine learning requires a large amount of annotated data and deep learning lacks large-scale training data and other disadvantages and bottlenecks, textual emotion analysis based on emotion dictionary has the advantages of simple thinking, easily implementation,areas are easy to identify and high precision, so we choose the emotion dictionary method for emotion analysis.

# **2.[Research](C:/Users/Administrator/AppData/Local/youdao/dict/Application/8.9.3.0/resultui/html/index.html" \l "/javascript:;) [contents](C:/Users/Administrator/AppData/Local/youdao/dict/Application/8.9.3.0/resultui/html/index.html" \l "/javascript:;)**

This paper carries out emotion analysis based on the emotion dictionary method, which can be roughly divided into four steps: emotion dictionary selection, data set selection, data processing and emotion score calculation. The main research contents are as follows:

## 2.1 The choice of an emotional dictionary

At present, the open source sentiment dictionaries on the Internet include HowNet Emotion Dictionary, Taiwan University (NTSUSD) and Dalian University of Technology emotion dictionary.The HowNet emotion dictionaras 12 files, divided into English and Chinese.The Chinese emotion dictionary includes the emotional text of evaluation, emotion, opinion and degree (positive and negative).In this paper, evaluation and affective words are integrated as emotion dictionary. The degree words contained in the table of degree words are classified into: most -very -more -ish - insufficciently -over .There are more than 700 stop-words in the "Stop-Words Lexicon of Harbin Institute of Technology". Due to the richness of HowNet emotion Dictionary and Stop-Words lexicon of Harbin Institute of Technology, we choose the above two as dictionary.

## 2.2 Data set selection

Because it runs on CPU and is limited by memory and runtime , the open source micro-blog with small data set is selected. Micro-blog contains news, finance and entertainment, with a total of 65 data sets.

## 2.3 Data Processing

Data processing consists of two tasks, clause and word segmentation.Firstly, the data is divided into the minimum analysis unit according to punctuation marks, and then the jieba tool in Python is selected for word segmentation according to the stop-word thesaurus of HIT.

## 2.4 Calculate emotional scores

Analyze the words in the sentence and match them with the emotional vocabulary, find out the positive words and negative words, processing the degree words before the words, weighting them according to the degree word vocabulary, calculate the emotional score of the whole sentence and output the emotional polarity.

# 3.Platforms and Key technologies

## 3.1 Platforms

Python is an object-oriented interpreted computer programming language, which can be applied in artificial intelligence, scientific computing, Web development, system operation and maintenance, big data and cloud computing, finance, game development and other fields.It is cross-platform, simple, extensible and embedded, and can be built and used in Linux, MAXOS and Windows. The code needs a few changes when it runs on different platforms. In addition, Python has a large number of standard libraries and third-party libraries with relatively complete functions.Through the reference to the library, the development of different areas of business can be realized.Python is good for our tasks.

In addition, because of the large number of Python libraries, managing them and maintaining them in a timely manner is an important but complex task.Anaconda is open source, easily to install, supports Python and R languages, which can helps us manage our environment and Python libraries.

## 3.2 Chinese clause technology

Chinese has more punctuation marks than English. Chinese clauses usually rely on punctuation marks, so we divide words according to the punctuation marks. The common Chinese clauses are "?, "!", ".", "...", ":", "\n".The design of cut\_sen function can divide sentences more accurately.

## 3.3 Chinses words divided syncopation technology

Chinese word segmentation is different from English, because every word in English sentence is separated by a space, so it can be divided directly by a space, while Chinese is much more difficult.However, with the development of technology and the deepening of research, many Chinese word segmentation technologies have emerged.Common word segmentation technique including Jieba, Ltp (Language technology platform), ICTCLAS (Institute of Computing Technology, Chinese Analysis System) and various others.

Jieba word segmentation is selected in this paper. It supports traditional word segmentation and is also provided by custom dictionary. There are three modes for word segmentation, which are fast and accurate.Jieba word segmentation is characterized by its strong expansibility, high accuracy and powerful functions, so it is selected in this paper for word segmentation.

# 4.Design and Implementation

## 4.1 Overall Design

The following steps are needed to realize the emotion analysis based on the HowNet emotion dictionary:

Training data acquisition: The acquisition of open source, high quality and small amount of microblog data sets.

Chinese clause: Divide sentences according to Chinese punctuation marks.

Chinese word segmentation: according to the stop-word list of Harbin Institute of Technology, accurately word segmentation is performed for the dividing sentences.

Statistical emotion words and degree word processing: According to HowNet emotion dictionary, words are divided into positive emotion and negative emotion.Divide the degree words into six and assign different weights to them.

Score calculation: The sum of each clause is calculated, and the polarity is divided according to the score of positive emotion and negative emotion.

The specific flow chart is shown in Figure 1

****

Figure1. Overall flow chart

## 4.2 Training data acquisition

According to the experimental needs, it is hoped to find a small and high-quality open data set of emotion classification.This paper selects Microblog corpus data set, covering three aspects and 65 pieces of data in total.The text emotion analysis in this paper is at the sentence level. It processes the data set to form a format corresponding to the emotion of the sentence, which can obtain a total of 65 pieces of data.

## 4.3 Clause design

In this paper, the cut\_sent function is used for word segmentation, according to the Chinese punctuation marks to judge each line in the file, judge the sentence length and each word, if the word is a punctuation mark, the sentence is broken, otherwise continue to judge.

## 4.4 Participles design

This paper chooses the jieba segmentation tool for word segmentation.In order to improve the word accuracy and bring convenience for model training.In the process of word segmentation, two measures are used to divide the dictionary and stop using words.

1. Divide the dictionary

The dictionary used in this paper is HowNet emotion dctionary. Using the Chinese dictionary, positive evaluation words, positive emotion words, negative evaluation words and negative emotion words are combined into positive word dictionary and negative word dictionary respectively to expand the scope of the dictionary. The two dictionaries contain 4566 and 4370 words respectively.

Secondly, divide the degree word dictionary.Invent the degree dictionary of degree words in HowNet emotion dictionary, divided into six degree dictionaries of most -very -more -ish - insufficiently -over.Used for weighting calculation.

After joining in the dictionary, the effect of word segmentation has been significantly improved.

2. Stop using words

The purpose of stop words is to remove words irrelevant to emotion analysis, reduce the amount of text to be processed, and thus improve the speed of model learning.In order to reduce the training burden and improve the training efficiency, this paper adopts the way to stop words, but does not remove Chinese words, only remove symbols and numbers, so as to retain the meaning of sentences as much as possible.

Remove all symbols from the stop list.Before processing a large number of corpus sets, all corpora need to be converted to a half-angle pattern in case the regular expression does not match.The text will be much cleaner after the stop words are removed.

## 4.4 Statistical emotion word and Degree word processing

Invent venue of different degrees according to the positive and negative dictionaries. After determining the polarity, we find the degree words before it and invent them with different weights, most (with weight 6), very (with weight 5), more (with weight 4), ish (with weight 3), insufficiently(with weight 2), and over (with weight 1).

## 4.5 Score calculation

The sum of the score of positive emotion and negative emotion of each clause is calculated, and the score of negative emotion is subtracted by the score of positive emotion. If the difference is greater than 0, it is positive; if it is less than 0, it is negative; if it is equal to 0, it is neutral.

# 5.Experimental results and Analysis

## 5.1 Environment configuration

This paper relies on Python3.6 and Anaconda for code writing, and python libraries of time, Jieba, numpy and re are needed, and the specific environment configuration is shown in Figure 2:

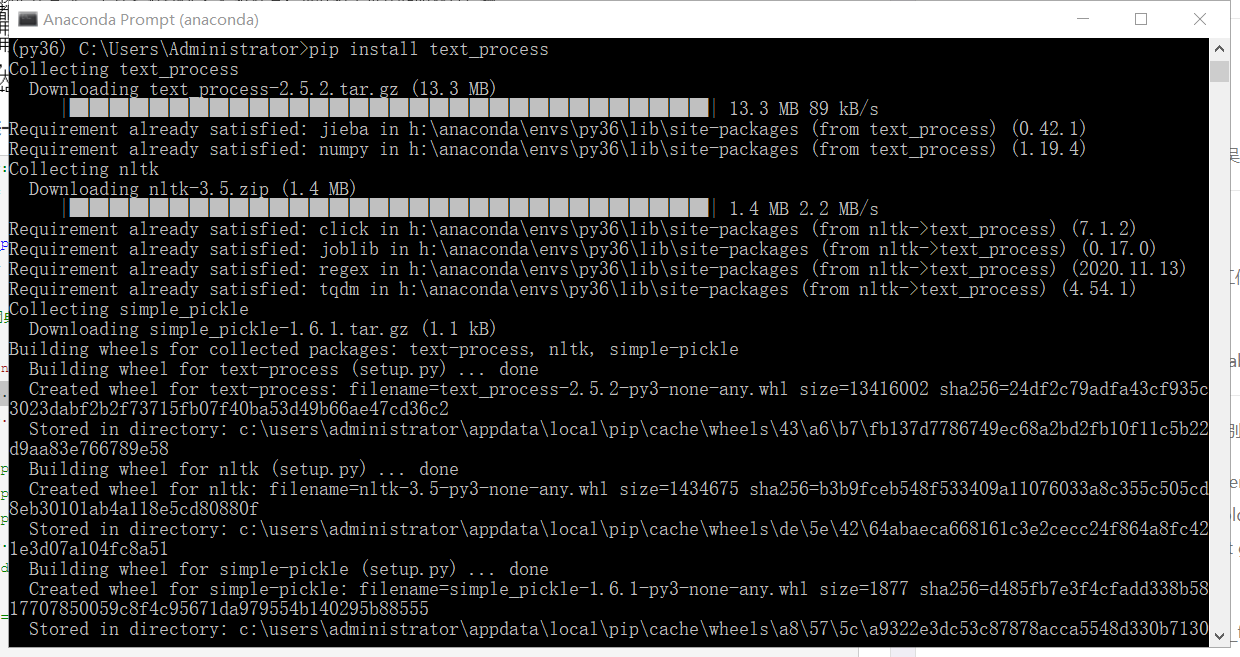


Figure2. Environment Configuration

## 5.2 Dictionary of degree words

This article is based on HowNet emotion dictionary analysis, the degree words are divided into six dictionaries, which are most very more ish insufficiently over .The specific dictionary is shown as follows:

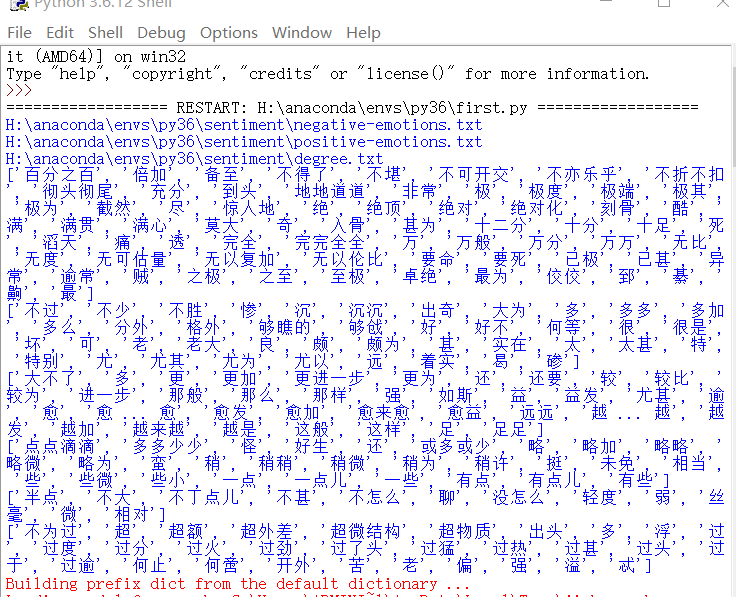


Figure3. Degree word dictionary

## 5.3 Chinese clause result

In this paper, the custom cut\_sent function is adopted to divide sentences according to Chinese punctuation marks and line breaks, which can obtain better word segmentation results, as shown in figure. 4:

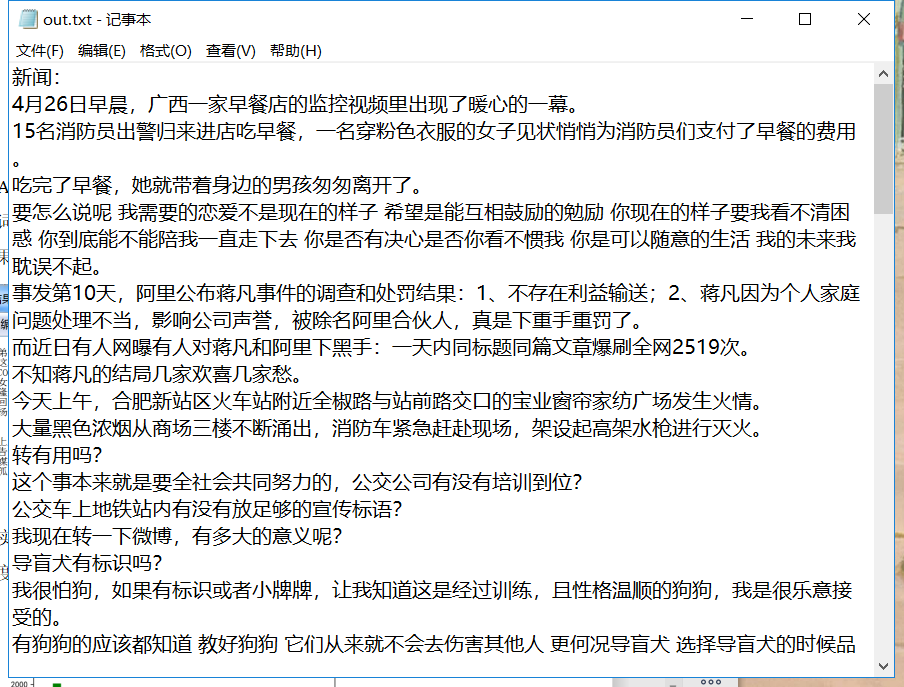


Figure4. Clause result

## 5.4 Chinese participle result

Jieba word segmentation method is adopted in this paper. In the design of data word segmentation module, specific operation methods for the two steps of dictionary division and word stopping have been clearly defined. After these steps, the text finally gets relatively good word segmentation results, as shown in Figure5:

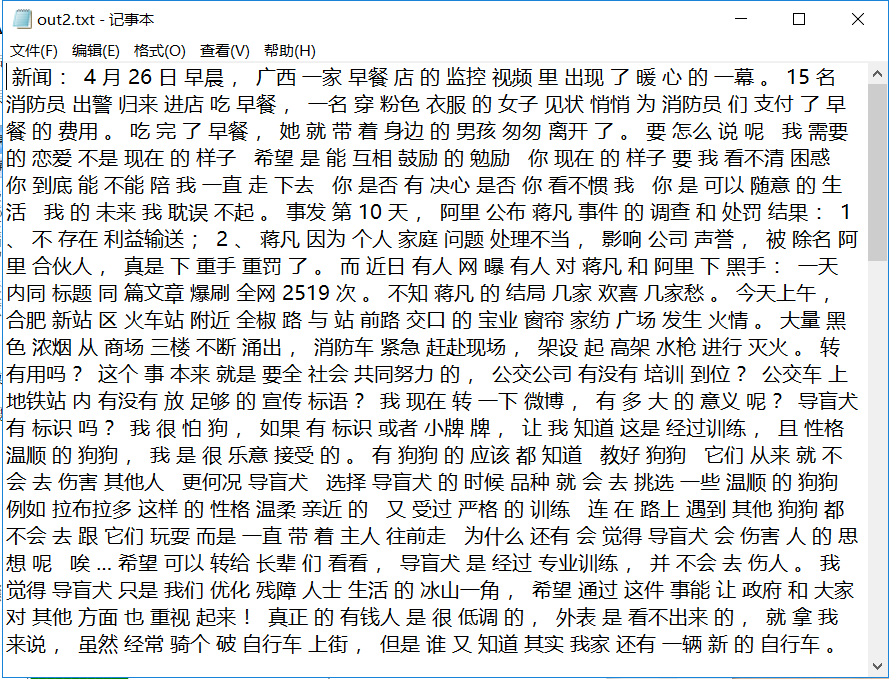


Figure5. Segmentation results

## 5.5 Final result

The weighted sum of degree words and the difference between positive and negative emotions were obtained as shown in Figure6:

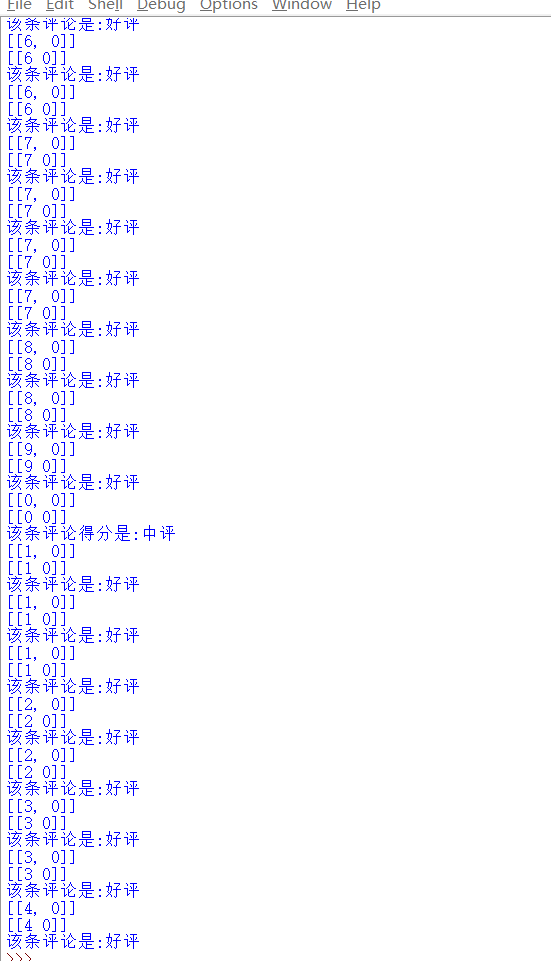


Figure6. Result diagram

# 6.Interpretation of result

Through constant modification of the model, the results obtained in this paper are shown in figure 7.The factors influencing the results are analyzed:

(1) Inadequate consideration: This paper simply considers the effect and influence of stop words and degree, but does not consider the influence of negative words or punctuation marks, so the accuracy of sentence score is low. In the follow-up, continuous training, model modification and parameter modification are needed to achieve higher accuracy.

(2) The complexity of the micro blog this: microblog this is colloquial and diversified characteristics, microblog exists in all kinds of sentence patterns, using irony, emoticons, and so on, which makes microblog data don't like hotel reviews or is the language of movie review results so unified, eventually lead to microblog sentiment analysis more difficult, difficult to extract.At the same time, due to the rapid updating of microblog language, the dictionary update speed lags behind, so many novel words cannot be found in the dictionary.

# Reference：

1. Ma Yu-kun,Peng Hai-yun，Khan T，et al. Sentic LSTM： A hybrid network for targeted aspect-based sentiment analysis [J].Cognitive Computation，2018，10(4)：639-650.
2. Liang Jun，Chai Yu-mei，Yuan Hui-bin，et al． Polarity shifting and LSTM based recursive networks for sentiment analysis [J]．Journal of Chinese Information Processing，2015，29(5)：152-159．(in Chinese)
3. Tian Zhu, Research on sentiment analysis based on deep feature representation [D]．Weihai：Shandong University，2017．(in Chinese)
4. Zhou Yu,Xu Rui-feng，Gui Lin.A sequence level latent topic modeling method for sentiment analysis via CNN based diversified restrict Boltzmann machine [C]Proc of International Conference on Machine Learning and Cybernetics，2016：356-361.
5. Shervin Minaee,Nal Kalchbrenner.Deep Learning Based Text Classification: A Comprehensive Review,2020