A Study on Multi-label Sentiment Analysis for Chinese Text

Wang Lei

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March, 2017

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Acknowledgement

How time flies! I have spent four years studying in Tokushima University. Where did the time go? At this very moment, I cannot help sighing at what I gained, comprehended and done in the past four years. I think back to not only the boring academic research, but also the care and encouragement given by my supervisor, teachers and classmates.

I would like to sincerely thank my supervisor, Professor Fuji Ren. I have been studying, living and carrying out thesis research under your meticulous care and careful guidance. I have been learning the specialized courses, publishing academic theses, selecting a topic for the doctoral dissertation, conducting research and writing the thesis under your painstaking care. Your rigorous scholarship, high academic accomplishments and broad scientific field of vision have benefited me a lot. It is my honor to have been your student, and your instruction will benefit me all my life. Meanwhile, I would thank Professor Kenji Kita and Professor Masami Shishibori, who had contributed a lot of time and efforts in reviewing my thesis. And their valuable recommendations helped to improve this thesis. I would also like to thank all my teachers who have helped me to develop the fundamental and essential academic competence.

I am indebted to many of my colleagues to support me. Special thanks go to our members of Ren Lab: Haitao Yu, Xin Kang and Chao Li. We have been communicating together, encourage one another and making progress jointly. We have enlightened each other and discussed with each other in daily life and study. Every day I spent with you in Tokushima University will be a wonderful memory in future.

Especially, I want to thank my parents and spouse for their selfless care and understanding. Each progress and achievement made in Tokushima University should be attributed to the encouragement and support from them. As thesis writing is nearing completion, let me express my profound gratitude to them.

I would like to thank all the people around me or related to me during my doctoral course. The journey of this life period will remain in my memory for ever.

Table of Contents

Acknowledgement.....I
Table of Contents.....II
List of Tables.....V
List of Figures.....VI
Abstract.....VII
Chapter 1 Introduction.....1
1.1 Motivation.....1
1.2 Background and Significance.....2
1.3 Research Status.....5
1.4 Main Research Contents.....8
1.5 Organization Structure.....9
Chapter 2 Word Sentiment Analysis based on Maximum Entropy.....11
2.1 Related Work.....11
2.2 Word Sentiment Orientation Analysis.....13
2.2.1 Basic Framework.....13
2.2.2 Maximum Entropy Model.....14
2.2.3 Feature Selection.....16
2.2.4 Smoothing Technique.....17
2.2.5 Word-Sentence Emotional Relationship.....18
2.3 Experiments.....18
2.3.1 Experimental Data.....19
2.3.2 Evaluation Method.....19
2.3.3 Result and Analysis.....19
2.4 Summary.....22
Chapter 3 Sentenc Sentiment Analysis with Multiple Features.....23
3.1 Related Work.....23
3.2 Sentence Sentiment Analysis based on Emotion Word.....24
3.3 Sentence Sentiment Orientation.....25
3.3.1 Basic Framework.....26
3.3.2 Emotion Vector Space Model.....26
3.3.3 Topic-based Emotion Vector Space Model.....27

3.3.4 Feature Lexicon....28
3.4 Experiments....34
3.4.1 Experiment Data....34
3.4.2 Experimental Result....35
3.4.3 Analysis of Experimental Results....39
3.5 Summary....40
Chapter 4 Sentence Sentiment Analysis based on Bayesian Network....42
4.1 Introduction....42
4.2 Latent Dirichlet Allocation....44
4.3 Sentence Sentiment Analysis....45
4.3.1 Basic Idea....45
4.3.2 Multi-Label Emotion Topic Model....46
4.4 Experiments....53
4.4.1 Experimental Data....53
4.4.2 Relations between Sentiment Orientation and Topic....53
4.4.3 Identification of Single Emotion Orientation 55
4.4.4 A Comparison between MLETM and NB 56
4.4.5 Macro-averaging of MLETM 57
4.4.6 Analysis of Experimental Results 57
4.5 Summary....58
Chapter 5 Document Sentiment Analysis based on Three-way Decisions....60
5.1 Introduction....60
5.2 Three-way Decisions Model....61
5.3 Document Sentiment Analysis....62
5.3.1 Senantic Similarity....63
5.3.2 Document Sentiment Orientaion....66
5.4 Experiments....68
5.4.1 Experimental Data....68
5.4.2 Standard of Experiment Evaluation....69
5.4.3 Experiment Result 69
5.4.4 Discussion 70
5.5 Summary....71
Chpater 6 Summary and Futrue Work....72
6.1 Overall Summary....72

6.2 Future Work.....74
Bibliography.....76

List of Tables

Table 2.1 Statistical Information of Data Sets.....1
Table 2.2 The multi-label Emotion Orientation Identification in Ren-CECps.....21
Table 3.1 Negation Dictionary.....30
Table 3.2 Degree Adverb Dictionary.....31
Table 3.3 Conjunction Dictionary.....33
Table 3.4 Statistical Information of Data Sets…….....35
Table 3.5 The Identification Result of Multi-label Emotion Orientations…….....38
Table 3.6 Identification of the Emotion Orientation with the Highest Intensity.....38
Table 4.1 Distribution of the 8 Basic Emotion Orientations……..…….46
Table 4.2 The Meaning of All Parameters in MLETM…..…….48
Table 4.3 A Comparison between MLETM and NB……....56
Table 5.1 Loss Functions for Decisions of Two States……...……..67
Table 5.2. Distribution of Sentiment Orientations for Texts……..……..68
Table 5.3 Comparison of Multi-label Sentiment Recognition……..69

List of Figures

Figure 1.1 The Basic Process of Text Sentiment Analysis.....5
Figure 2.1 The Basic Framework.....14
Figure 2.2 Identification Procedure of Word Emotion Orientation.....17
Figure 2.3 Emotion Orientation Identification in Hotel Review Corpus.....20
Figure 3.1 Sentence sentiment Orientation Analysis Framework.....20
Figure 3.2 LDA Model.....27
Figure 3.3 Identification Emotion Orientation of Complex Sentence……..34
Figure 3.4 The Relationship between Sentence Emotion Orientations and Topic…....36
Figure 3.5 Sentence Emotion Orientations in the Tan Data Set.....37
Figure 3.6 Identification of Basic Sentiment Orientations……....39
Figure 4.1 Graphic Structure of MLETM…….……..…….....47
Figure 4.2 The Relationship between Topic and Sentence Emotion Orientation….....54
Figure 4.3 Chage in the Running Time of MLETM.....54
Figure.4.4 The Identification Accuracy of the 8 Basic Emotion Orientations.....55
Figure 4.5 The Variation of *Macro – accuracy* and *Macro – precision* 57
Figure 5.1 Structure of Tongyi Cilin…….....64
Figure 5.2 Multi-label Document Sentiment Analysis Framework…….....68
Figure 5.3 A Comparison of 8 Basic Sentiment Orientation…….....70

Abstract

With rapid development of the Internet, a great amount of information full of personal subjective emotions appears on the Internet, such as microblogs, news comments and product reviews. Sentiment orientation of text is studied for the purpose of exploring people's attitudes and emotional states toward entities or their attributes in the texts. Research on sentiment orientation of text has become a hot research topic in the field of natural language processing and research achievements have been widely used in many social fields, including politics, economics and management.

Over the past few years, considerable progress has been made in sentiment analysis of text. However, many difficulties and challenges have been posed on text sentiment analysis because of complicated human emotions, diversified contents and forms of texts. As a new research method in the field of artificial intelligence, granular computing is mainly used for processing a huge amount of uncertain, fuzzy and imprecise information. The basic notions and principles of granular computing is to find an ideal solution by decomposing a complex problem and analyzing data at different levels of granularity. In this paper, an attempt is made to apply the idea of granular computing to explore multi-label sentiment orientation of Chinese texts. Based on differences in linguistic granularity, research on sentiment orientation of text is studied from the perspective of word, sentence and document. Several methods of multi-label text sentiment analysis based on granular computing are put forward and a brand-new interpretation on sentiment orientation of text is given from uncertainty of sentiment orientation. The main contributions of this thesis are summarized as follows:

(1) In light of complex and diversified human emotions, multi-label sentiment orientation of Chinese text is studied. In the past, most researchers only examined complimentary and derogatory emotions on the level of words, sentences or documents. In this paper, eight categories of multi-label sentiment orientation are studied from the perspective of words, sentences and documents in the Ren-CECps, in order to be more in line with abundant and complex human emotions in practice.

(2) Considering that sentiment orientation of words is influenced by the context, a method of word sentiment orientation based on maximum entropy is proposed. First of all, features, relationships, semantic characteristics and emotional relations of

words are extracted from pertinent contexts. Subsequently, sentiment orientation of words is identified by maximum entropy model, and problems about sparse features are solved by smoothing techniques. At last, uncertainties of sentiment orientation of words are further removed based on emotional connections between words and sentences. Experiments on the Ren-CECps Chinese emotion corpus have achieved good results. The emotional relationship between words and sentences also helps to improve the accuracy of recognizing sentiment orientation of words.

(3) Viewing that a sentence is made up of diversified elements, a method is put forward to analyze sentiment orientation of sentences by integrating multiple features. Although affective words are important factors that determine sentiment orientation of sentences, sentiment orientation of sentences is affected by relationships between the sentences of a text. At first, intensity of sentiment orientation is regulated for affective words of a sentence pursuant to topical features. Besides, sentiment orientation of sentences is judged in combination with other semantic features of negations, conjunctions and adverbs of degree in a sentence. The experimental results show that semantic features in a sentences can play a positive role in the recognition of sentiment orientation of sentences, which helps to improve the accuracy of recognizing sentiment orientation of sentences.

(4) By introducing topic features into sentiment orientation of sentences, a semi-supervised multi-label emotion topic model is put forward. A sentiment topic layer is incorporated into latent Dirichlet allocation that calculates probability distribution of topics-word, document-topics and word-sentence sentiment, so as to discriminate sentiment orientation of sentences. By comparing experiments, it is proved that the method can effectively discriminate the sentiment orientation of sentences, which can improve the accuracy of recognizing sentiment orientation of sentences.

(5) In view of general and complex sentiment in document, a method is proposed to analyze sentiment orientation of documents based on three-way decisions. Relatively great errors may be caused by directly identifying sentiment orientation of documents by words or sentences, so the method covering multiple steps and stages is adopted. Theory of three-way decision is introduced into sentiment orientation analysis in the paper. In this paper, decision-making process of sentiment orientation is divided into two stages with several steps. Firstly, a sentiment lexicon is composed by applying training data sets. Meanwhile, sentiment orientation of affective words and its

intensity in testing data are identified using Tongyici Cilin and HowNet. Then, sentiment orientation of documents is preliminarily identified according to intensity of sentiment orientation of affective words by which document is then divided into positive, negative and boundary regions with appropriate thresholds. At last, sentiment orientation of documents of delayed decisions are discriminated pursuant to emotional characteristics of pertinent sentences. The experimental results show that the method can improve the accuracy of recognizing the multi-label sentiment orientation and eight basic sentiment orientation of documents.

Key Words: Sentiment Orientation Analysis, Ren-CECps Chinese Emotion Corpus, Multi-label Classification, Three-way Decisions, Granular Computing

Chapter 1

Introduction

1.1 Motivation

With the increasing popularization of network technology and the rapid expansion of e-commerce, the Internet is experiencing a sharp increase in information quantity, and the development of information technology is bringing out a great change in the way of interpersonal communication. In the meantime, more and more users have begun to get essential information through mutual communication on the Internet.

In the era of Web1.0, the Internet emphasized information editing and integration, so when a user read a content provided by a website, such as Sina, Sohu and Netease, information flowed unidirectionally from the website to him. With the advent of Web2.0 technology, Internet communication got more focused on interaction with users, thus realizing bi-direction flowing of information between them, such as community, forum and blog. At present, we are in an era that is seeing the rise of Web3.0 technology. Web3.0 highlights initiative, digital maximization and multi-dimension, takes service as the major content, and offers user preference-based personalized services. Renren and Facebook are two typical representatives.

With the rapid development of Web2.0 technology and the rise of Web3.0 technology, some thinking concepts of Internet, such as “user experience and user service”, have won support among the people, and more and more users have got ready to post and gather information on the Internet. Most of such information data is semi-structured or non-structured text data, such as product review, film criticism and news comment. This kind of text information not only describes objective facts, but also conveys users’ subjective views on objective facts. The strong personal subjective emotion contained in the information reflects people’s various preferences and emotion orientations, such as pleasure, anger, sorrow and joy, from different perspectives. An analysis on emotions hidden in the text information on the Internet can help well identify how much users like a product, discover the evolution law of news events, and recognize individuals’ emotion state and variation trend. All this has

promoted the development of text sentiment analysis technology, making it a research hotspot in the field of natural language processing.

Text sentiment analysis is also known as opinion mining[1], aimed at mining, analyzing and processing the subjective information in texts that contains emotional colors, such as viewpoint, standpoint and preference, to identify emotional information in the texts, such as emotion subject, emotion object, emotional features and emotion orientation, and widely apply the analysis results to a number of fields, including social life, political life and commercial activities. At the initial stage of sentiment analysis, people primarily analyzed and applied sentiment words, such as "happy", a commendatory term, and "sad", a derogatory term. As there are more and more subjective texts containing emotional colors on the Internet, the academia has exalted the analysis and study of text sentiment from the research on simple word sentiment to the research on complex sentence sentiment and the most complicated document sentiment, upgraded simple sentiment orientation analysis to sentimental element identification, and started research on the effect of different fields and contexts on text sentiment orientation[2,3,4,5,6]. According to the difference in language granularity, text sentiment analysis can be divided into three research levels[7,8]: word sentiment analysis, sentence sentiment analysis and document sentiment analysis. Text sentiment analysis involves a number of very challenging research tasks. This chapter first sets forth the research background and significance of the thesis, then introduces the major individual research contents and innovations, and finally presents the main contents and organizational structure of the paper.

1.2 Background and Significance

The 21st century has seen mankind entering an "age of information explosion". Along with the unceasing expansion of the Internet information superhighway, Internet information technology has penetrated into every corner of human social life, and is influencing all aspects of human production and life at an unprecedented speed. According to the 37th Statistical Report on Internet Development in China issued by China Internet Network Information Center (CNNIC), by December, 2015, there had been 688 million netizens in China, the Internet penetration rate had reached 50.3%, and half of the Chinese people had accessed the Internet. Mobile netizen number reached 620 million, with the percentage rising to 90.1%. More and more individuals began to surf the Internet on the mobile phone; online education, medical treatment

and car rental took shape on the Internet. In a word, the Internet started to influence the overall society in a new manner. From this perspective, the Internet containing mass information has become an important information source for humans. On the other hand, in the face of so violent a "flood of data", people usually feel at a loss what to do, and this is the phenomenon known as "information overload". So, for the current information processing technology, one of the primary goals is to solve the problem of "data overload" to the greatest extent, remove useless junk data quickly to avoid data disorders, and mine useful information fast and accurately.

Sentiment analysis research began in the late 1990s, and Professor Picard proposed the concept of "affective computing" in 1997[9]. From then on, sentiment intelligence research drew attention from lots of scholars and aroused their interest. They dedicated themselves to acquiring human emotion information from communication media, such as text, voice, facial expression, gesture and physiological signal. Since the 21st century began, the huge research value and economic potential brought about by sentiment analysis, in particular text sentiment analysis technology, have been known to the business circles and scientific community. Also, after over twenty years of development, sentiment analysis has become a research hotspot in the field of natural language processing, artificial intelligence, machine learning and data mining.

Text sentiment analysis technology can be applied very widely. Specifically, it can be used to acquire the public's attitudes, views and feelings for things from the Internet, so as to facilitate human activities and create huge social values and wealth. However, owing to the unending changes of Internet language and the emergence of a great many Internet catchwords, text sentiment analysis has become increasingly difficult. The absence of a diversified text emotion corpus has brought enormous difficulties to the research of text emotion analysis.

Granular computing[10] is a new concept of information processing and a new computing paradigm. It covers the research of all theories, methods, technologies and tools related to granularity, and its main idea is to granulate every problem at different levels for data analysis, and make use of the information at different granular levels for intelligent problem solving. When solving a complex problem, human usually does not adopt a systematic, accurate method to directly develop an optimal solution for this problem, but often take a progressive strategy to analyze the problem at different levels to develop an approximate solution that meets the requirement. Granular computing is similar to the process in which human solves a complex

problem. Granular computing theory not only is reflected in its idea concerning hierarchical problem solving, but also includes specific models[11], such as word computing[12], rough set[13] and quotient space[11].

With regard to natural language comprehension, lots of researchers have pointed out that the most complicated thing in human world is nothing other than linguistic "vagueness". Vagueness is a common phenomenon in languages, and a ubiquitous objective reality in every natural language. On the other hand, natural language has a hierarchical structure. For instance, a word is composed of letters, a sentence is composed of words, and a document is composed of sentences. When analyzing text sentiment, human also adopts a similar method for a bottom-up, low-to-high layer-by-layer research to solve all problems of text emotion research.

In conclusion, text sentiment analysis has a characteristic similar to that of granular computing theory. We try to apply the hierarchical operation process of granular computing to the research of the multi-label sentiment orientation in Chinese texts, to simulate human's identification strategy for text sentiments, aimed at intensively exploring and researching the multi-label sentiment orientation of words, sentences and documents in texts. The research significance of this subject is mainly reflected in the following two aspects:

(1) Task research provides a new research approach for the analysis of the multi-label sentiment orientation in Chinese texts

The successful application of granular computing theory to artificial intelligence, pattern recognition, data mining and many other fields has convincingly demonstrated the richness and flexibility of granular computing thought. Considering the vagueness of natural languages, we believe we will be able to well combine granular computing thought with text sentiment analysis to effectively analyze and solve the problems concerning the analysis of multi-label sentiment orientation in Chinese texts. A multi-granular, multi-level research model can be set up to research the problems of sentiment orientation at each level and further infer and analyze the sentiment orientation at each level by full use of the emotional connection between different levels to open up a new way for the research of multi-label emotion orientation in Chinese texts.

(2) Task research can enrich the application fields and research directions of granular computing

Granular computing theory has been commonly researched and applied in artificial intelligence, pattern recognition and other fields. But it is still less feasible to research and analyze the problem of Chinese text sentiment orientation from a multi-granular, multi-level perspective with granular computing thought. We apply granular computing thought to the analysis of multi-label sentiment orientation in Chinese texts, and further enriches the research of granular computing theory via application feedback. This is a further attempt at the practical application of granular computing theory and a supplement to it.

Therefore, the task of the thesis has important research values both in theory and practical application.

1.3 Research Status

As there are more and more review texts on the Internet, manual method has been unable to cope with the collection and process of the online mass information, and a more efficient, intelligent method has been urgently needed to help users search and acquire related review information. Text sentiment analysis has become an important application field in natural language processing, on which more and more scholars have begun to carry out a research. Here we are going to review the research status of text sentiment analysis based on the research hotspot.

1. Basic Process of Text Sentiment Analysis

The basic process of text sentiment analysis is shown in Figure.1.1.

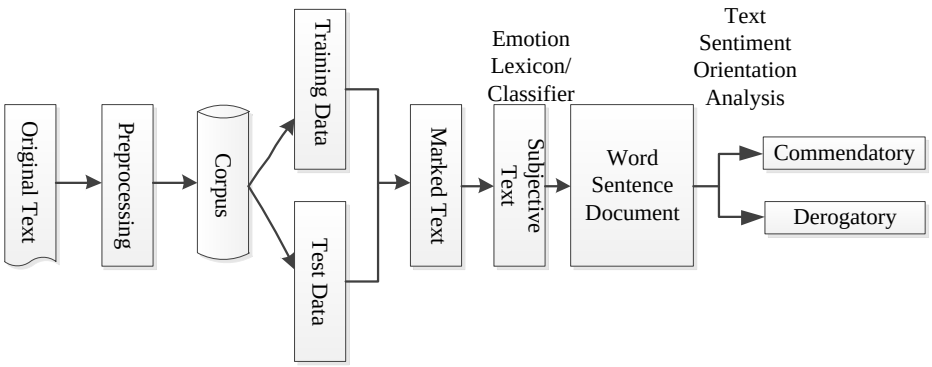

Figure1.1 The Basic Process of Text Sentiment Analysis

Firstly, we get raw materials from the Internet using crawler tool, and collate the materials. Second, we conduct text preprocessing for the materials collected, such as noise elimination, label filtering, word segmentation, and word marking, to set up a corpus to provide the follow-up work with relatively standard texts. Then, we judge

whether the original texts contain subjective or objective information using the corpus set up above or a classification method, and then remove the objective texts exclusive of subjective emotional colors. Finally, we identify the sentiment orientation of the subjective texts at word, sentence and document levels using a machine learning or semantic rule-based method based on the corpus.

2. Sentiment Information Extraction

Sentiment information extraction aims to extract valuable emotional information from emotional texts, which can be considered as a basis for text sentiment analysis. Since the rise of sentiment analysis research, emotional information extraction has been a research hotspot in the academic fields, and the enthusiasm hasn't ever been undiminished today. Valuable emotional information primarily includes evaluative words (affective words in most cases, such as good and bad), evaluation objects (target and its attribute, such as mobile phone and memory capacity), and view givers (those that give views, such as user and government).

3. Subjective Text Recognition

On the Internet, there are a large number of evaluative texts, including a great many subject texts with emotional colors, as well as many objective texts excluding view or emotion. Objective texts merely involve an objective description of external objects, while subjective texts involve people's personal views on externalities or feelings or them. The latter is the major object of research on text sentiment analysis. So, there are two major methods for the classification of text sentiment information: first, binary classification of subjective and objective texts; second, sentiment orientation classification of subjective texts, including the common binary classification based on single-labeled commendatory/derogatory or the more defined multi-label sentiment orientation classification[14].

Most of the subjective sentence identification methods identify subjective sentences based on emotion words by the use of various text features and classifiers. Finn[15] et al. identified subjective sentences by part-of-speech tagging. Yu[16] et al. constructed Bayes classifier by various POS information, and used multiple classifiers to improve the quality of the corpus. Pang[17] established a graph structure based on the distribution features of the sentences with the same attributes, and transformed the problem of subjective sentence identification into that of minimum graph cut, enabling the classification of subjective and objective sentences. In terms of Chinese,

Ye[18] et al. extracted POS combination patterns by CHI statistical method, combining subjective emotion words together, identifying subjective sentences.

4. Text Sentiment Orientation Analysis

Text sentiment orientation analysis aims primarily to identify text sentiment orientation or classify texts according to commendatory and derogatory meanings emotionally. According to the difference in text analysis granularity, text sentiment orientation analysis can be divided into three levels: word, sentence and document. Accordingly, there are three major research methods: simple statistics method, machine learning method and correlation analysis method. Simple statistics method is to carry on simple statistics of all emotional words or emotional phrases in texts, and then compare the statistical results with threshold to identify text sentiment orientation. This method is usually used for sentence and document analysis. Machine learning method is to train a lot of labeled corpuses by machine learning to generate sentiment orientation classifiers to classify test texts according to sentiment orientation. This method is usually used for sentence and document analysis. Correlation analysis method is to realize word-level sentiment orientation identification by semantic labeling or the co-occurrence relationship between seed emotional words and characteristic words.

5. Emotion System

As text sentiment analysis research deepens, some text sentiment analysis systems have come into being. Product information analysis system Opinion Observer[19] is a visual analysis system that is used to review online product features. It extracts features from the subjective texts concerning online product review, and then carries on statistics of product features in accordance with commendatory and derogatory reviews to finally identify the overall quality of product features. In the Chinese fields, Yao Tianfang[20] developed a Chinese opinion mining system for automotive products, realizing analyses and statistics of the commendatory and derogatory reviews on automotive product features, leading a visual result.

Overall, people have recognized the importance and practical application value of text sentiment analysis technology, and achieved some research results. In the research process of text sentiment analysis, some scholars have, consciously or unconsciously, introduced granular computing thought into it to analyze and research text sentiment issues from a multi-granular, multi-level perspective. Some other scholars have begun to consider the emotional integration at different granularities

and levels. This paper intends to research the multi-label emotion orientation issues in Chinese texts.

1.4 Main Research Contents

Text sentiment orientation analysis is an important task in text sentiment analysis, and a research hotspot in the natural language processing. Along with the rapid development of the Internet, as well as the increasing complication and diversification of text contents and expression forms, text sentiment orientation analysis has run up against more and more challenges. From the perspective of linguistic granularity, text sentiment orientation analysis mainly involves word sentiment orientation analysis, sentence sentiment orientation analysis and document sentiment orientation analysis, yet the research on them is still far from being systematic. With granular computing theory as a master tool, we do research into multi-granular, multi-label Chinese text sentiment orientation from the perspective of the uncertainty and diversity of emotion orientation, studies the cause of the uncertainty of multi-label text emotion, and sets up a corresponding research model, algorithm or framework, aimed at realizing the identification of multi-label Chinese text sentiment orientation. This paper will not only enrich and improve the research of text sentiment orientation, but also expand the application of granular computing theory to practical problems. The major research contents of this paper are shown as follows:

1. Word Sentiment Orientation Analysis based on Maximum Entropy

The uncertainty of word sentiment orientation may be affected by the context. According to the information concerning context, we put forward a word sentiment orientation analysis method based on maximum entropy to extract word features, word relationship features, word semantic features and word sentiment features from the context, then identifies word sentiment orientation based on the maximum entropy model, and solves the problem of data sparseness by smoothing technique. Finally, according to the emotional connection between words and the relevant sentence, this paper further removes the uncertainty of word sentiment orientation.

2. Sentence Sentiment Orientation Analysis with Multiple Features

Sentence sentiment orientation is not only decided by the emotional features of the emotion words in the sentence, but also affected by the negative words, degree adverbs and conjunctions in the sentence. Meanwhile, sentences also interact with each other and depend on each other emotionally. In a context, if different sentences

describe the same topic feature, these sentences based on the same topic feature may be similar to each other in terms of sentence sentiment orientation. So, sentence sentiment orientation should be decided by the emotional word in the sentence concerned that is better able to topic feature, and then the topic feature should be used to adjust the sentiment orientation intensity of the emotion word, so as to adjust the sentiment orientation and intensity of the sentence. According to this idea, we put forward a method of sentence sentiment orientation analysis with multiple features, analyzes multiple grammar features and text topic features in sentences, and finally uses the feature information to identify sentence sentiment orientation.

3. Sentence Sentiment Orientation Analysis based on Bayesian Network

Latent Dirichlet allocation (LDA) model is an important text classification method, which is commonly used for text classification. By introducing topic feature into the research of sentence sentiment orientation, we put forward a new semi-supervised multi-label emotion theme model (MLETM), and add a sentence emotion topic layer to the three-layer structure of LDM to label each sentence with emotion orientation and label each word with topic so as to realize the identification of word emotion orientation.

4. Document Sentiment Orientation Analysis based on Three-way Decisions

Document sentiment has high generalization, as well as the highest complexity and diversity, so it has high uncertainty as well. Because of this, a big error will be caused if document sentiment orientation is identified by direct use of words or sentences. We introduce three-way decisions theory into document sentiment orientation analysis and divide emotion orientation decision-making process into two stages to carry out work. First, we construct an emotional lexicon by a training set, and identifie the emotional orientation, as well as intensity, of emotional words based on HowNet dictionary and thesaurus. Then, in accordance with a three-way decisions model, we identify document sentiment orientation based on the emotional intensity of emotional words, and divide document set into positive region, negative region and boundary region based on document sentiment orientation intensity. Finally, in line with the emotional features of sentences, we make a final judgment on the document sentiment orientations decided tardily.

1.5 Organizational Structure

The thesis mainly studies and analyzes multi-label Chinese text sentiment orientation

issues from three different granularities, such as word, sentence and document. It first analyzes the current research methods of word sentiment orientation, puts forward a maximum entropy model, and applies it to the identification of word sentiment orientation. Second, it integrates the emotional features, topic features and other features of word together for identification of sentence sentiment orientation. On this basis, it further improves the LDM to build an MLETM for direct identification of sentence sentiment orientation. Finally, on the basis of word and sentence sentiment analysis, it identifies document sentiment orientation through three-way decisions model. The organizational structure of this paper consists of the above research contents as follows:

Chapter 1 talks about the background of the task and research significance, and introduces the main research contents and organizational structure of this paper.

Chapter 2 introduces word sentiment orientation issues and research methods. A maximum entropy model based on multiple grammar features is proposed for Chinese word multi-label sentiment orientation identification, and emotional relationship between words and sentences improve the classification capacity of the model.

Chapter 3 analyzes sentence-level sentiment orientation issues and research method. We propose a multi-feature fusion-based sentence sentiment orientation analysis method, and combine contextual theme features with syntactic features for identification of the multi-label emotion orientation of Chinese sentences.

Chapter 4 explores the LDA and its application. Proposes a new semi-supervised MLETM based on textual theme features for identification of the multi-label emotion orientation of Chinese sentences.

Chapter 5 researches and analyzes the decision-theoretic rough set theory, and puts forward a multi-stage, multi-step, multi-strategy Chinese document sentiment orientation identification method based on three-way decision theory, as well as the emotional information hidden in words and sentences.

Chapter 6 sums up the results achieved in the research process and the weaknesses, and points out a direction for the future research.

Chapter 2

Word Sentiment Analysis based on Maximum Entropy

Emotion word refers to the words with sentiment orientation, and plays a decisive role in text sentiment analysis. Word sentiment orientation analysis intrigued the researches of different countries when sentiment analysis just came into being. As a basis for text sentiment analysis, it decides both sentence sentiment analysis and document sentiment analysis. A word emotion orientation analysis method is presented based on maximum entropy, and integrated the emotional information of the targeted sentence into it to improve the identification accuracy of word sentiment orientation.

2.1 Related Work

From the linguistic perspective, linguistic granularity consists of document, paragraph, sentence, phrase, word and morpheme from big to small. In the existing sentiment analysis research field, most scholars choose word as the basic linguistic granularity and research basis and use word emotion information to further analyze sentence and document sentiment.

As for word sentiment analysis research, lots of scholars first think of constructing a seed emotion lexicon, and then use this seed emotion lexicon to analyze the emotion features of new words. For this reason, many emotion lexicons (e.g., MPQA Subjective Dictionary[21], SentiWordNet[22] and LIWC[23]) have appeared successively. With different characteristics, these lexicons have offered great help to text sentiment analysis research. However, these lexicons have two weaknesses: field independence and context independence. In other words, the sentiment orientation of a word must be determined in a specific field or context, while there may be a different sentiment orientation of the same word in different fields or contexts.

For word sentiment orientation analysis, there are two methods: lexicon-based method and corpus-based method.

The lexicon-based method is mainly to set up a more comprehensive lexicon by expanding an existing lexicon, to identify the emotion of new words. English word

sentiment identification is mainly carried out by WordNet. That is, the synonym set, antonym set and semantic hierarchy in WordNet are used to expand the number of emotion words[24,25,26]. Esuti and Sebastiani[27] put forward a semi-supervised learning method, and expanded the positive and negative word sets using the synonym set and antonym set in WordNet. After that, they used these two emotion word sets as training sets generating multiple binary classifiers. Andreevskaia and Bergler[28] put forward multi-round Bootstrapping algorithm and used word's fuzzy fraction to evaluate the credibility of word classification. The Chinese scholars usually adopt HowNet for Chinese word classification. The core idea of this method is to calculate the semantic similarity between new words and seed words. Liu Qun and Li Sujian[29] proposed a HowNet-based word semantic similarity calculation method, finally translating word similarity calculation into sememe similarity calculation. This method considers synonymous relationship, but ignores antonymous relationship. Some scholars[30,31] took antonymous relationship into consideration, but they just made an external adjustment rather than gave consideration to the essence of similarity calculation. Lin[32] defined the similarity between two objects as the ratio of common information to different information, thus having both synonymous and antonymous relationships taken into consideration.

Semantic similarity can help conduct emotion analysis[31,33] on the independent words in the context to find out the several words similar to this word semantically that can be found in the seed lexicon. Then, these seed words can be used to infer the sentiment orientation of new words. However, this method depends too much on the number and quality[34] of the seed words in the seed lexicon, so field or context linkage cannot be realized.

The corpus-based method is mainly to mine the emotion words in corpuses, such as adjective clustering[36] and pointwise mutual information[35] by using a small number of seed emotion words in line with the statistic characteristics of large-scale corpuses, and judge their orientation. With the advent of probabilistic graph model and social network technology, the graph model has been gradually applied to word sentiment orientation analysis. Rao[37] et al. and Velikovich[119] et al. iteratively updated the edge weight and node emotion label in the graph model using semi-supervised label propagation algorithm and graph propagation respectively. Qiu[120] et al. proposed a two-way propagation algorithm, and extracted emotion words interactively according to the dependence between emotion words and

attributes. Zhang and Liu[40,41] paid attention to the emotion information contained in objective words.

Since word sentiment orientation may change with the context, semantic similarity cannot be directly used to accurately identify the sentiment orientation of context-related words.

Example 1:

Sentence 1: “多少共同的等待和期盼，多少共同的热情和祝愿。” (“We have waited and expected together with countless passions and wishes for many times.”)

Sentence 2: “不经意间，我们发现他们之中的很多人还有一个共同的符号。” (“Inadvertently, we found that many of them had a common symbol.”)

In Sentence 1, “共同” means “joy”, while in Sentence 2, “共同” means surprise.

With given seed emotion words, semantic similarity calculation can be used for sentiment orientation identification of context-irrelevant words. Let “good”, “beautiful” and “kind” be standard emotion words with a commendatory meaning, “bad”, “ugly” and “evil” be standard emotion words with a derogatory meaning, and the semantic orientation of a new word be equal to the sum of its similarity with the standard emotion words with a commendatory meaning minus the sum of its similarity with the standard emotional words with a derogatory meaning. This method relies on the size of standard emotion word sets and the selection of standard emotion words, but it ignores the possible emotion difference in words in the context. A maximum entropy model is put forward, which can be applied to word sentiment orientation identification by means of the context-relationship.

2.2 Word Sentiment Orientation Analysis

2.2.1 Basic Framework

We put forward a word emotion orientation analysis method based on maximum entropy model, whose frame structure is shown in Figure.2.1.

With the emotion words in sentences as the objects of study, we can identify the sentiment orientation of these emotion words in accordance with the maximum entropy model and word-sentence emotion relationship. Features are extracted from the training corpus with maximum entropy classifiers (Part A), and then the features extracted from the training corpus are input into the classifiers (Part B) for

preliminary identification of word sentiment orientation. Afterwards, word-sentence emotional relation features are used to correct word sentiment orientation.

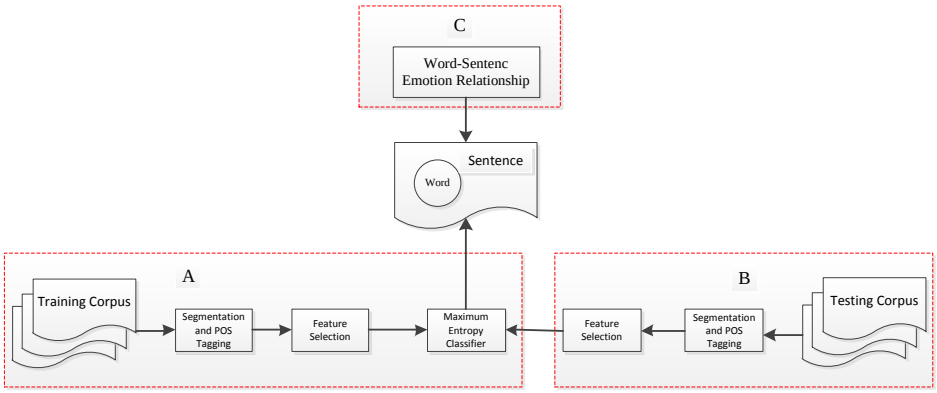

Figure.2.1 The Basic Framework

2.2.2 Maximum Entropy Model

Maximum entropy model[42] is a probability statistical model based on information entropy theory. When an optimum system distribution is selected in certain constraint conditions, if the constraint conditions cannot be used to determine an only system distribution, the best distribution will be the maximum distribution of system information entropy as long as all the conditions are satisfied. A major advantage of maximum entropy-based model is that it can describe different features under the same framework, and feature independence hypothesis is not required.

Probability evaluation method is adopted for maximum entropy model. Suppose x is an event, and y is the context for the occurrence of Event x, the joint probability of x and y is denoted by $p(x, y)$. With regard to word sentiment orientation identification, that an emotion word belongs to a sentiment category can be regarded as an event, and the various features related to this emotion word in the text can be considered as the circumstance in which the event occurs. If an emotional word needs to be put under its category, a training corpus can be used for statistical calculation.

Definition 3.1: preset a training set, and define $E = \{e_1, e_2, \cdots, e_m\}$ as the sentiment orientation set of emotion words, and $C = \{c_1, c_2, \cdots, c_n\}$ as the feature set concerning emotion words acquired from documents, then Formula 2.1 can be used for probability estimation.

$$\tilde{p}(e_i, c_j) = \frac{count(e_i, c_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} count(e_i, c_j)} \tag{2.1}$$

For this probability estimation method, there is a problem know as "sparse matrix" problem, and for a large training set, if there are lots of two-tuples (e_i, c_j) not co-occurring, the probability of this two-tuple will be estimated to be zero, and this is obviously incorrect. In maximum entropy model, the probability distribution of unknown events can be made uniform as far as possible to solve the sparse matrix problem, but this method also has weaknesses. This chapter will adopt a smoothing technique to solve this problem in accordance with the specific situation of training set.

According to the definition proposed by Shannon, the calculation formula of entropy is shown below:

$$H(p) = -\sum_{x} p(x) \log_2 p(x) \tag{2.2}$$

The formula for solving the probability distribution that meets the maximum entropy principle is shown below:

$$p^* = \arg\max H(p) \tag{2.3}$$

Despite the inability to evaluate the joint probability value of all two-tuples, it is feasible to evaluate the joint probability value of some two-tuples or acquire some constraint conditions. This way, the problem will be turned into a problem concerning maximum entropy calculation.

For constraint condition and feature description, the concept of feature function is usually adopted. Under normal circumstances, feature function is a two-valued function, as shown in Formula 2.4.

$$f(e, c) = \begin{cases} 1, & e \text{ and } c \text{ satisfy some condition} \\ 0, & \text{Otherwise} \end{cases} \tag{2.4}$$

It is expected that the empirical probability of feature function f should be equal to the sum of all the empirical probabilities that meet the feature requirements, as shown below:

$$E_{\tilde{p}} f = \sum_{e,c} \tilde{p}(e, c) f(e, c) \tag{2.5}$$

The expected value of feature function f in the random event learned is shown in Formula 2.6:

$$E_p f = \sum_{e,c} \tilde{p}(c) p(e \mid c) f(e, c) \tag{2.6}$$

In the training set, the empirical probability expectation and expected probability of features should be identical as a constraint condition, as shown below:

$$E_p f = E_{\tilde{p}} f \tag{2.7}$$

Several feature functions can be defined in accordance with word features and linguistic context. These feature functions can be selected from different perspectives and granularities of the research problem, and connected to each other or unrelated to each other, so that dispersed and fragmented knowledge could be integrated together to perform every task jointly.

Definition 3.2: suppose there are n feature functions $f_1, f_2, \cdots, f_n$, and $E_P f_i = E_{\tilde{p}} f_i$, then the maximum entropy model can be described as the model with the maximum entropy among all the models that meet the constraint conditions, as shown in Formula 2.8 and 2.9:

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i, i = 1, 2, \cdots, n\} \tag{2.8}$$

$$p* = \arg\max_{p \in P} H(p) \tag{2.9}$$

Lagrange multiplier algorithm is a classical approach used to develop an optimal solution in the constraint conditions, and we also adopt this model for solving, thus the following formula is obtained:

$$p*(e \mid c) = \frac{1}{Cxt(c)} \exp(\sum_{i=1}^{n} \lambda_i f_i(e, c)) \tag{2.10}$$

where $Cxt(c)$ is normalization factor, expressed as below:

$$Cxt(c) = \sum_{e} \exp(\sum_{i=1}^{n} \lambda_i f_i(e, c)) \tag{2.11}$$

λ_i is the weight of feature function f_i, and in the training process, every λ_i value is evaluated according to the learning process in the training set.

Maximum entropy model is widely used in natural language processing, particularly for text classification and information retrieval. We first adopt maximum entropy model as a basic classifier for preliminary identification of word sentiment orientation, and then further integrate the word-sentence emotional relationship for correction of word sentiment orientation.

2.2.3 Feature Selection

We selected the following contextual features from the data training set and applied them to the maximum entropy model. These features are described as follows:

(1) WEF: directly identify word sentiment orientation in line with word sentiment

orientation features according to the statistical information in the training corpus.

(2) NWF: identify the sentiment orientation $w_{i-n},\cdots,w_i,\cdots,w_{i+n}$ of candidate w_i in accordance with the contextual feature that it is surrounded by n words in the sentence.

(3) PWF: identify the sentiment orientation of candidate w_i according to its feature and the features of the n words around it in the sentence. Word segmentation and POS tagging have been completed in all sentences in Chinese emotional corpus Ren-CECps. The accuracy is no less than 97%, with 35 word categories involved.

(4) PNEF: identify the sentiment orientation of candidate w_i according to the emotion features of the n emotional words prior to w_i .

The procedure of identifying word emotion orientation based on maximum entropy model is shown in Figure.2.2.

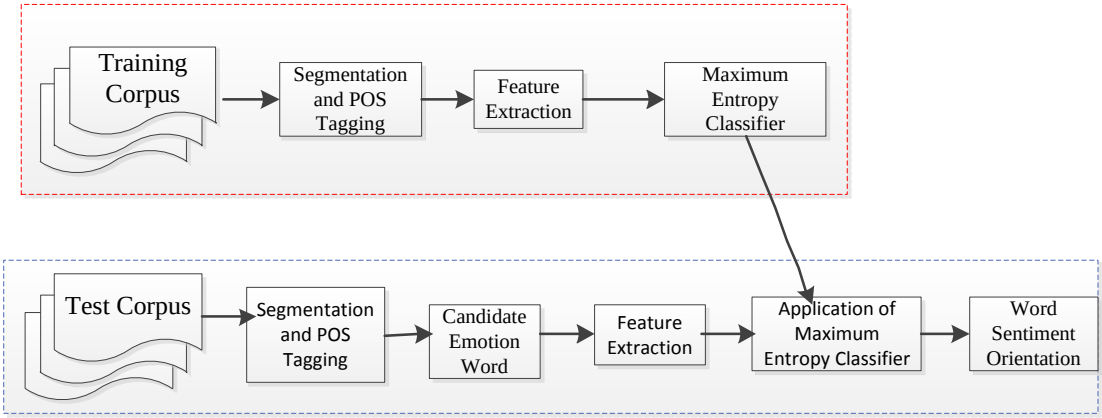

Figure.2.2 Identification Procedure of Word Emotion Orientation

2.2.4 Smoothing Technique

For the multi-label sentiment orientation identification of words, most words have one or two sentiment orientations, while only few words have three emotion orientations, thus many words have very sparse sentiment features. Data sparseness is a common problem in machine learning. For such a problem, we deal with it by absolute discounting smoothing technique[43].

Absolute discounting smoothing technique is to discount all events that can be identified in a model by subtracting a fixed value and distribute this fixed value equally to all events that fail to appear. For word sentiment orientation identification, we directly assign a value d to all features that haven't appeared yet, and transform feature function formula 2.4 into Formula 2.12.

$$f_{w,e}(e,\mathrm{c}) = \begin{cases} \dfrac{count(w,c)}{N}, & e = e' \\ d, & \text{otherwise} \end{cases} \tag{2.12}$$

where $d = 0.05$, N represents the number the times that word w appears in the training set, and $count(w,c)$ represents the number the times that word w in the training set appears in the context c.

2.2.5 Word-Sentence Emotional Relationship

Maximum entropy model can help identify word sentiment orientation and intensity. In order to further improve the accuracy of word sentiment orientation identification, we can adjust word sentiment orientation according to the word-sentence emotion relationship. An iterative computation made in accordance with the following relations can help identify a stable word sentiment orientation and intensity.

$$e(w)^{(i+1)} = \alpha \times e(w)^i + (1-\alpha) \times e(Sens(w))^i \tag{2.13}$$

$$e(Sens(w)) = \sqrt{\sum_{w \in sen_w} \frac{e(sen_w)^2}{n_w}} \tag{2.14}$$

$$e(sen_w) = \sum_{w \in sen_w} \frac{e(w)}{n_{sen}} \tag{2.15}$$

α is a adjustable parameter and set equal to 0.64 in the experiment. $e(w)^i$ represents the emotion orientation intensity of word w during the ith iteration, $Sens(w)$ represents the set of all sentences containing word w in a text, $e(Sens(w))$ represents the sentiment orientation intensity of the sentence set containing all sentences, n_w represents the number of sentences in the sentence set, sen_w represents a sentence containing word w, $e(sen_w)$ represents the sentiment orientation intensity of a sentence containing word w, and n_{sen} represents the number of the words contained in a sentence.

In the entire iterative computation process, full use is made of the word-sentence emotion relationship to correct the sentiment orientation intensity of words and then identify the sentiment orientation of words. Parameter α is adjusted for the sentiment intensity of words that was identified during the last iteration. Besides, the emotion intensity of all words needs to be standardized and normalized in each iterative computation.

2.3 Experiments

2.3.1 Experimental Data

The experimental data comes mainly from two data sets: one is a Chinese emotion corpus provided by Tan Songbo at Chinese Academy of Sciences, from which 2000 positive and negative non-repeated hotel reviews are selected respectively. The sentiment orientations of words and sentences fall into 2 kinds: positive and negative; the other is Chinese emotional corpus Ren-CECps, in which the emotion orientations of words and sentences fall into 8 kinds, which can be used for multi-label emotion orientation identification of words. The statistical information of the above data sets is shown in Table.2.1.

Table.2.1 Statistical Information of Data Sets

Data Set	Field	Sample Size	Sentiment Orientation
Hotel	Hotel Reviews	4000	2 Kinds
Ren-CECps	Blog Text	1476	8 Kinds

Two experiments are conducted in this chapter. In Experiment 1, hotel reviews are used to identity the sentiment polarity of emotional words. This involves the binary classification of a sentiment orientation; Experiment 2 is conducted in order to identity the several sentiment orientations of each emotional word in Ren-CECps. This involves the multi-label classification of sentiment orientation.

2.3.2 Evaluation Method

A sample-based evaluation measure is adopted, represented by the mean error between the actual label set and predictive label set in the test data. The evaluation method[44] prescribed by the Fifth Chinese Opinion Analysis Evaluation is adopted, and for each sentiment orientation, precision, recall and F-measure are taken as evaluation criteria.

$$\mathrm{Pr}ecision = \frac{\#System.Correct}{\#Golden} \tag{2.16}$$

$$\mathrm{Re}call = \frac{\#System.Correct}{\#System.\mathrm{Pr}oposed} \tag{2.17}$$

$$F1 = \frac{2 \times \mathrm{Pr}esion \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{2.18}$$

2.3.3 Result and Analysis

Experiment 1 is conducted in order to identify the positive and negative sentiment orientations of the emotion words in the hotel review corpus. The adjectives in corpus are used to set up an experimental data set consisting of training corpus and testing corpus to identify the emotion orientation of the candidates in the testing corpus.

ICTCLAS is adopted at the preprocessing phase for word segmentation and POS tagging. The emotional lexicon is a basic emotional lexicon composed of 2090 emotional words and 6846 evaluative words from HowNet, with the individual-character emotional words left out.

The method by which maximum entropy model is used for word sentiment orientation identification is denoted by MaxEnt, and the method by which the word-sentence emotional relationship is integrated into the maximum entropy model for word sentiment orientation identification is denoted by Combine. Figure.2.3 shows the result of word sentiment orientation identification.

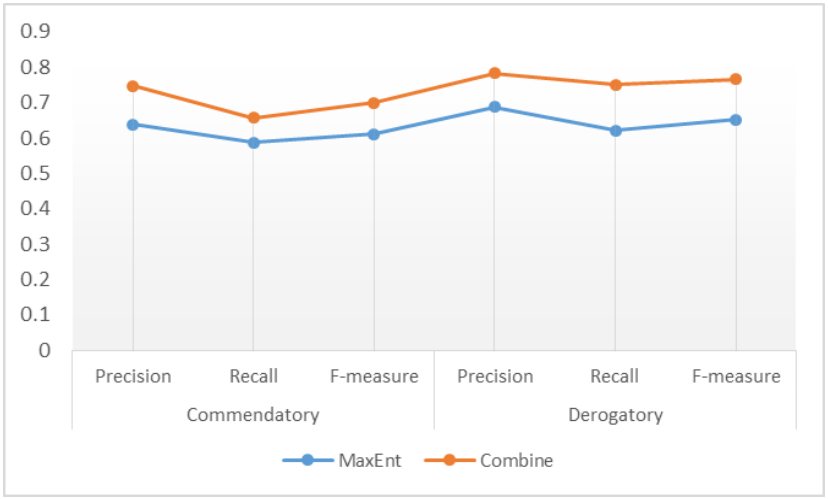

Figure.2.3 Emotion Orientation Identification in Hotel Review Corpus

For the experimental results, we see that both methods have achieved good effect in identifying the commendatory and derogatory emotion orientations of the candidates, and that after the integration of word-sentence emotion relations, the identification effect of emotion orientations has been further improved.

Experiment 2 is conducted in order to identify the multi-label sentiment orientations of the emotional words in Ren-CECps. 1476 blog texts containing 34630 sentences and 101842 emotional words were chosen from Ren-CECps. The emotion lexicon required for the experiment is composed of the emotion words extracted from the training corpus. For this data set, a 5-fold cross-validation method is adopted for experiment.

The multi-label sentiment orientation identification result of emotion words is shown in Table.2.2.

Table.2.2 The multi-label Emotion Orientation Identification in Ren-CECps

Feature Type		MaxEnt			Combine		
		Precision	Recall	F1	Precision	Recall	F1
WEF	$f_1 = w_i$	51.4	47.6	49.4	53.7	52.9	53.3
NWF	$f_2 = w_{i-1}, w_i, w_{i+1}$	50.7	45.3	47.8	52.8	52.2	52.5
WEF +NWF	$f_1 = w_i$ $f_2 = w_{i-1}, w_i, w_{i+1}$	53.3	51.8	52.5	58.6	58.2	58.4
WEF +NWF +PWF	$f_1 = w_i$ $f_2 = w_{i-1}, w_i, w_{i+1}$ $f_3 = pos_{i-1,} pos_i, pos_{i+1}$	56.1	55.2	55.6	60.3	59.8	60.0
WEF +NWF +PWF +PNEF	$f_1 = w_i$ $f_2 = w_{i-1}, w_i, w_{i+1}$ $f_3 = pos_{i-1,} pos_i, pos_{i+1}$ $f4 = pre\_e_{i-1}$	56.3	55.1	55.7	62.4	61.9	62.1

As can be seen from the experiment result in Table.2.2, the sentiment orientation identification result is basically satisfactory, but the sentiment orientation identification rate is not high. This reveals not only the complexity of human emotion, but also the lower operability of multi-label emotion orientation identification than single-label emotion orientation identification.

The experiment result in Table.2.2 shows that the precision, recall and F1 value will be very low if only a single feature, such as WEF, is used to identify the sentiment orientation of candidates. As more and more features are chosen, the precision, recall and F1 value will go up continuously.

The causes of the errors arising in the multi-label sentiment orientation identification of words are analyzed in Experiment 2, as listed below:

(1) There are some short sentences in the data set, and in the short sentences there are very few extractable contextual features, so the sentiment orientation of words couldn't be identified effectively. It is a future research direction about how to extract more contextual features and use them for word sentiment orientation identification.

(2) For most emotion words, only one of their sentiment orientations is strong and easy to identify, while the rest emotion orientations are weak and hard to identify. So the overall sentiment orientation identification rate is very low.

2.4 Summary

In the chapter, we mainly analyze the identification of word sentiment orientation by putting forward a word sentiment orientation identification method based on maximum entropy model, and make full use of the contextual features and word-sentence emotionl relations.

In the previous analyses of word sentiment orientations, people primarily researched how to judge emotion words and non-emotion words in texts, and just divided emotion polarity into two categories: commendatory and derogatory. Moreover, only the lexicon-based method and corpus-based method were mostly adopted. In this chapter, we solve the problem of word multi-label sentiment orientation identification based on massive corpus data.

It identified the sentiment orientation of the candidate by extracting word sentiment features, adjacent word features, POS feature and the features of the prior N words from the corpus data set. A comparison between the results of the experiments based on hotel review corpus and Ren-CECps shows that the effect of multi-label sentiment orientation identification is lower than that of commendatory and derogatory motion orientation identification. This reveals not only the complexity of human emotion, but also the lower operability of multi-label sentiment orientation identification.

After preliminary identification of word sentiment orientations, we analyze the word-sentence emotion connection and set up a correction formula for word sentiment orientations, and then adjust word sentiment orientations by the correction formula, finally improving the effect of word sentiment orientation identification to some extent. The experimental results indicate that there is some relationship between word emotion orientations and the sentence emotion orientations, and this emotion relationship can be used for assistant identification of word emotion orientations.

It is a future research direction as to how to apply three-way decisions theory, in particular decision-theoretic rough set theory to the analysis of word sentiment orientations. Besides, as there are more and more network catchwords and symbols, the sentiment identification of network language will become one of the research hotspot in future.

Chapter 3

Sentence Sentiment Analysis with Multiple Features

At present, there have been lots of researches on word sentiment orientations, most of which involve the use of emotional lexicon, language knowledge and machine learning, with the scope limited to the sentences analyzed. However, the whole context has great effects on sentence sentiment orientation recognition, which has drawn enough attention from lots of scholars. We adjust the emotion orientation intensity of the emotional words in sentences by topic features, transforms topic information into emotional prior information by calculating word topic probability, and identifie word sentiment orientation by reference to the grammar features of the adjacent words.

3.1 Related Work

Along with the rapid development of e-commerce and social network, a lot of text information, such as blog, micro-blog, news commentary and shopping review, has sprung out on the Internet. All these texts are made up of sentences with emotional colors, and these sentences reflect people's preference for objective things or individuals' emotion. So, the research of sentence sentiment orientation analysis has received attention of many scholars at home and abroad, and offered help to the analysis of paragraph or short text sentiment orientation and even the analysis of document sentiment orientation.

A word or phrase is a research object of word sentiment analysis, while a sentence in a context is a research object of sentence sentiment analysis. Sentence sentiment analysis involves not only the identification of sentence sentiment orientation, but also the analysis and extraction of various subjective information in sentences. This subjective information mainly includes sentiment orientation-related reviewers, review objects and sentiment intensity.

Hu and Liu[24] identified word sentiment orientation using the synonymous and antonymous relationships in WordNet, and treated the sentiment orientation with advantages in a sentence as the sentiment orientation of the sentence. Dave[45] et al.

collected more than 1000 review articles marked with sentiment orientation, rated the features extracted with the statistical information composed of n-grams word, and then judged the sentiment orientation of new words in accordance with these features and their value. Yang[46] et al. integrated context into the conditional random fields model, putting forward a context-based sentiment analysis method. Narayanan[47] et al. researched the sentiment orientation of conditional clauses. Khan[48] et al. extracted subjective sentences by naive Bayesian method, established SVM classifiers by data training, and identified the sentiment orientation of sentences based on the context. Domestically, Liu Xiaohua[49] et al. at Harbin Institute of Technology decomposed a sentence into a series of subsequences, and then inferred the emotion orientation of the whole sentence by analyzing the sentiment of the subsequences, thus putting forward a sequence model-based sentence sentiment analysis method. Zhao Yanyan[50] et al. integrated inter-textual and intra-textual factors together to improve the precision of sentence sentiment analysis. Song Rui[51] at Dalian University of Technology researched comparative sentences, and adopted CRF model for emotion classification.

So far, there is still lack of enough researches on multi-label sentiment analysis. Bhowmich[52,53] et al. implemented multi-label sentiment classification for news sentences. In detail, they just classified multi-label emotions into 4 categories: antipathy, fear, happiness and sadness.

We take sentence emotion orientation analysis as a research focus, integrate topic feature into multi-label sentence sentiment orientation recognition, and fused some grammar features, such as negation, degree adverb and conjunction, together, putting forward a sentence sentiment orientation recognition method with multiple features to for identification of the multi-label sentiment orientation of sentences.

3.2 Sentence Sentiment Analysis based on Emotion Word

For the identification of sentence sentiment orientations, the simplest and most common method is the rule-based emotional word analysis method. This method is also often applied to the research of document sentiment orientation analysis. The basic idea of the rule-based emotional word analysis method is to summate emotional words or emotional phrases in sentences by weighting. If there are more words with a certain sentiment orientation and these words have higher emotion intensity, the

accumulated value of this emotion will be higher after summation. So, it will be more possible that the sentence concerned possesses this emotion.

The concrete analysis process of the rule-based emotional word analysis method is shown as follows:

(1) Preprocess sentences to remove the neutral words without any emotion orientation. Summate the emotional words or phrases in sentences. The calculation formula is shown below:

$$S-emotion(i)=\sum_{j=1}^{n} ew(w_j) \tag{3.1}$$

where $S: w_1, w_2, \cdots, w_n$, w_j is the *jth* emotional word in sentence S , $S-emtion(i)$ represents the *ith* category of emotion, and the $ew(w_j)$ represents the emotion intensity of w_j 's *ith* category of emotion.

(2) The calculation formula of sentence emotion categories is shown as follows:

$$Sentiment\_Class(S)=\arg\max(S-emotion(i)) \tag{3.2}$$

The emotion category with the maximum sum value is chosen as the major sentiment orientation of sentences.

The rule-based emotional word analysis method is used for identification of sentence or document sentiment orientation. This method, though simple, can achieve a good identification effect. But it has the following two weaknesses:

(1) Unitary feature: only emotional words or phrases are considered, while other words, such as negations, conjunctions, and degree adverbs, are ignored. But all these words may affect sentence sentiment orientation, and even change sentence sentiment orientation.

(2) Unanalyzable sentence structure: this method has disadvantages in the sentences of complex syntactic structure, such as negative sentences and complex sentences.

We make reference to this method in the process of sentence sentiment orientation identification, improve it with the influence of different types of words and syntactic structures on sentence sentiment orientation identification, and put forward a multi-feature sentence sentiment orientation analysis method by integrating text topic features into the process of sentence sentiment orientation identification.

3.3 Sentence Sentiment Orientation

3.3.1 Basic Framework

We put forward multi-feature sentence sentiment orientation analysis method, as shown in Figure.3.1. The dotted line partly represents the training process, which is aimed at constructing multi-label sentiment orientation classifiers. The focus of this method is to extract multiple features from sentences and represent the corresponding sentences with these features. As can be seen in the frame diagram, eature extraction need various dictionaries, including emotional lexicon, negative word dictionary and conjunction dictionary. Besides, syntactic structure also affects sentence emotion to some extent.

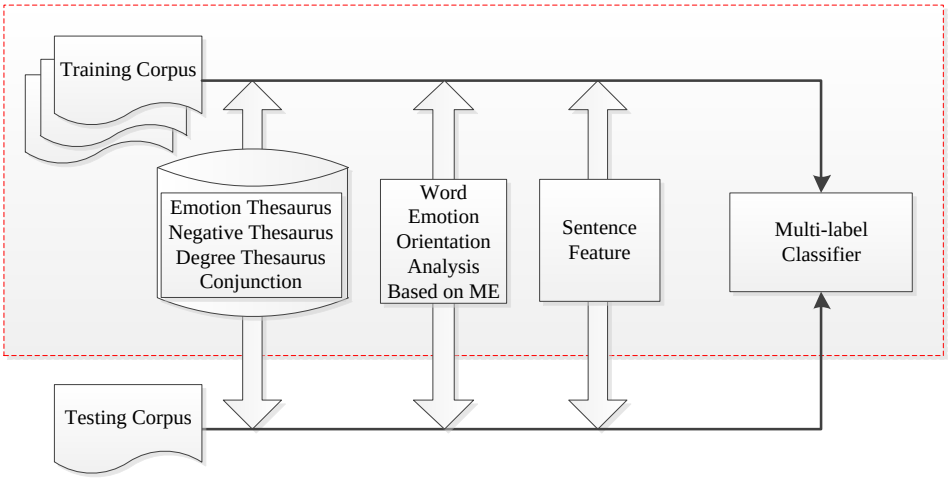

Figure.3.1 Sentence sentiment Orientation Analysis Framework

3.3.2 Emotion Vector Space Model

In order to precisely identify the multi-label sentiment orientations of sentences, we extract as many features as possible from sentences or texts and then apply these features to sentence sentiment orientation analysis. Word emotion information is the most important feature. After word segmentation, POS tagging, and removal of neutral words and stop words, only emotion words are retained.

Each word in Chinese emotional corpus Ren-CECps is marked with emotion orientation and intensity. In this chapter, all the emotion words in the training data in this corpus are extracted and made up an emotion lexicon, and then the emotion lexicon is used for sentence sentiment orientation analysis.

According to “BOW” model, sentence is regarded as an emotion word set, then sentence S can be expressed as:

$$S = (e(w_1), e(w_2), \cdots, e(w_n)) \tag{3.3}$$

$$e(w_i) = (e_i^1 \cdots, e_i^j, \cdots, e_i^8),\ 1 \le i \le n, 1 \le j \le 8 \tag{3.4}$$

$e(w_i)$ is an 8-dimensional emotion orientation intensity vector and represents the ith emotional word w_i's emotion orientation and intensity, and e_i^j represents the intensity of emotional word w_i's jth emotion orientation.

In order to identify the emotion orientation of sentence S, we represent the "BOW" model of sentence S with 8-dimensional emotion orientation. So, Formula 3.3 is transformed into the following one:

$$S - emotion(S) = (\sum_{i=1}^{n} e_i^1, \cdots, \sum_{i=1}^{n} e_i^j, \cdots, \sum_{i=1}^{n} e_i^8) \tag{3.5}$$

3.3.3 Topic-based Emotion Vector Space Model

In an article, sentence sentiment orientation should be decided by the core emotional word that is best able to reflect the text topic. If there is the same topic feature among sentences in a context, it will be extremely possible that the sentences based on the same topic feature have the same sentiment orientation, so text topic feature is introduced into the sentence emotion orientation discrimination process, and topic feature is used to adjust the emotion intensity of emotion words, so as to adjust the sentiment orientation and intensity of sentences.

Latent Dirichlet allocation (LDA), which was proposed by Blei[54] et al. in 2003, is a "text-topic-word" 3-layer Bayesian generative model. As a classic topic model, LDA represents topic mixture distribution in each text, and each topic is a concept distribution in words. Initially, only a hyper-parameter was introduced into the model letting the text-topic probability distribution follow the Dirichlet allocation. Later, Griffiths[55] et al. also introduced a hyper-parameter into the text-topic probability distribution to make it follow the Dirichlet allocation, thus establishing a complete generative model. This model is shown in Figure 3.2.

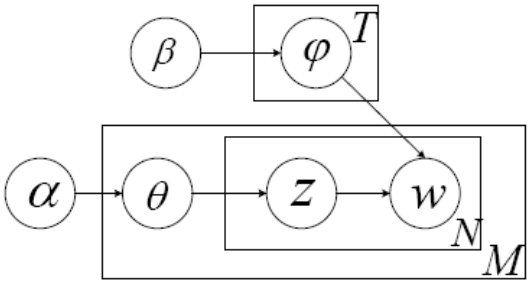

Figure.3.2 LDA Model

The number parameters in LDA model is only related to the number of topic and words, while irrelevant to the size of corpus, so it can be used to process large-scale corpuses.

Latent topic features is integrated into the sentence sentiment orientation discrimination process, and meanwhile LDA model is introduced into document D, obtaining T latent topics $T=\{t_1,t_2,\ldots,t_T\}$ and topic-word probability distribution φ. Then, the "text-topic-word" probability distribution is adopted to identify the emotional words that accords with the text topic features. The topic with maximum probability weight t_m is found among the T latent topics and applied to sentence emotion orientation discrimination formula 3.5, developing a sentence emotion orientation discrimination containing topic features, as shown below:

$$S-emotion(S)=(\sum_{i=1}^{n}(1+\varphi_{mi})e_i^{\,1},\ldots,\sum_{i=1}^{n}(1+\varphi_{mi})e_i^{\,7},\sum_{i=1}^{n}(1+\varphi_{mi})e_i^{\,8}) \qquad (3.6)$$

where φ_{mi} represents the topic-word probability distribution of the ith emotional word based on topic t_m.

3.3.4 Feature Lexicon

Apart from the emotional words, there are also other meaningful words in a sentence, and these words may affect and even change sentence sentiment orientation. In order better identify sentence sentiment orientation, we further extracted some supplementary features from sentences and applied them to sentence sentiment orientation discrimination. These supplementary features include negation feature, degree feature and disjunction feature.

3.3.4.1 Negation Feature

Negation feature is an important syntactic feature, and negative words can change the sentiment orientation of the emotional words within the range of jurisdiction and thus change the emotion of sentences. The range of jurisdiction of negative words, as well as the modified, has become a research hotspot in the linguistic domain[56,57,58,59]. The range of jurisdiction of a negative word usually begins with the negative word until the end of the sentence, while the modified is generally located behind the negative word.

Example 1:

Sentence 1: 中国队击败了韩国队。(Chinese team defeated Korean team.)

Sentence 2: 中国队没有击败韩国队。(Chinese team didn't defeat Korean team.)

Sentence 3: 不是中国队击败了韩国队。(It was not Chinese team that defeated Korean team.)

Sentence 4: 中国队没有不击败韩国队。(Chinese team did not lose out to Korean team.)

Sentence 5: 中国队击败的不是韩国队。(The one that was defeated by Chinese team was not Korean team.)

In the above sentences, Sentence 1 does not contain any negative word, and the whole sentence expresses a sentiment orientation known as happiness. Sentence 2-5 contain some negative words, which have different ranges of jurisdiction and modify different objects, and have changed the sentiment orientation of the sentences. It can be seen in this example that the object modified by negative word changes with context, so it is hard to define a complete rule to identify whether negative word has influence on the change of sentence sentiment orientation. Besides, negation of negation expresses a positive meaning, the sentiment orientation of sentences changes somewhat.

In this chapter, we adopte the proximity principle for sentence sentiment orientation identification. That is, negation just modifies the first emotional word behind it. For all emotional words in a sentence, it should be judged whether there is negation prior to it. If there does exist a negative word, it will affect the sentiment orientation of the sentence. We directly adopte a relatively simple processing rule to adjust the sentiment orientation intensity of the emotional words modified by the negative word, thus changing the emotion orientation of the sentence.

Definition 3.1: if there are N emotional words such as $w_1, w_2, \cdots, w_n$ in a sentence, and there is a negative word appearing ahead of emotional word w_i, the emotion orientation and intensity of w_i are:

$$e^*(w_i) = \begin{cases} \lambda_n \times e(w_i), \textit{odd number negative words ahead} \\ e(w_i), \text{even number negative words ahead} \end{cases} \tag{3.7}$$

$e^*(w_i)$ represents the sentiment orientation and intensity of the *ith* emotional word w_i after adjustment, and $\lambda_n = 0.5$ is an empirical value.

When there are an odd number of negative words prior to emotional word w_i, the emotion orientation intensity of emotional word w_i will change; when there isn't negative word or are an even number of negative words prior to emotional word w_i,

the emotion orientation intensity of emotional word w_i will not change;

The selection of negative words has direct effects on the identification of sentence sentiment orientation. We constructed a negation dictionary based on HowNet. The negative words in HowNet contain "negation", according to which the negative words is searched, and some common negative words are added manually. Besides, the negative words often appearing on the Internet is also included in this dictionary. 65 negative words are included in the negation dictionary, with some listed in Table.3.1.

Table.3.1 Negation Dictionary

不 No	不是 Not	没 None	没有 Nothing
不能 Cannot	不会 Won'	无法 Unable	毫不 Not at all
不要 Don't	不可 Not allowed	从不 Never	从未 Not ever
不曾 Not yet	绝不 Absolutely not	未曾 Ne'er	木有 Without

3.3.4.2 Degree Feature

Among adverbs there is a special kind of adverb known as degree adverb, which generally modifies verbs and adjectives[56]. Degree adverb cannot change the sentiment orientation of the emotion words modified, but it may affect the emotion intensity of emotion words, such as increasing or decreasing the emotion intensity of emotion words.

A degree dictionary containing 140 degree adverbs is constructed in accordance with the degree adverbs in HowNet. Each degree adverb is graded and set equal to 2, 3, 4 and 5. Some is listed in Table.3.2 for some degree adverbs and their weight.

Definition 3.2: given a degree adverb dw, the weight of this degree adverb is:

$$\lambda_d(dw) = 0.25 \times (3 + Rank(dw)) \tag{3.8}$$

where $Rank(dw)$ represents the grade of degree adverb dw, and that $\lambda_d(dw) \in (1, 2]$.

When a degree adverb modifies an emotional word, it can be put ahead of this emotional word or behind it.

Table.3.2 Degree Adverb Dictionary

Word	Grade	Word	Grade
有点	2	格外	4
Somewhat	2	Exceptionally	4
稍微	2	实在	4
Slightly	2	Indeed	4
较为	3	万分	5
Relatively	3	Extremely	5
更加	3	绝顶	5
Especially	3	Remarkably	5

The emotion orientation intensity of the emotion words modified by degree adverbs is defined as follows:

Definition 3.3: if there are N emotional words such as $w_1, w_2, \cdots, w_n$ in a sentence, and there are degree adverbs around emotional word w_i, the emotion orientation and intensity of emotional word w_i can be expressed as:

$$e^*(w_i) = \lambda_d \times e(w_i) \tag{3.9}$$

where $e^*(w_i)$ represents the emotion orientation and intensity of the *ith* emotional word w_i after adjustment, and λ_d is the weight of the degree adverbs.

3.3.4.3 Disjunction Feature

Owing to the complexity of grammar and the polysemy of emotional words, it is very difficult to identify sentence sentiment orientation only by using the word features contained in sentences. An understanding of sentence structure relationship will help discriminate sentence sentiment orientation. According to structure features[56,57], Chinese sentences can generally be classified into two categories: simple sentence and complex sentence. It is easy to identify the emotion orientation of a simple sentence, for which the method introduced above can be adopted. However, it is hard to identify the emotion orientation of a complex sentence, since it is usually composed of two or more clauses separated by comma.

Example 2:

Sentence 6: “你们的获奖当之无愧。”(You fully deserve this award.)

Sentence 7: “你们获取的虽仅是数千美金，但其价值、其意义远胜于百万英

镑。(Although you have just got several thousand dollars, its value and significance are much higher than that of one million pounds.)

Sentence 6 is a simple sentence, and Sentence 7 is a complex sentence composed of two clauses separated by comma.

The principal-subordinate relationship between clauses can be classified into three categories: coordinate relationship, causal relationship and transition relationship. For a complex sentence containing a coordinate relationship, there is an equal relationship between the clauses, which express the same emotion orientation. In a complex sentence containing causal relationship, there is a causal relationship between the clauses, and the clauses express the same emotion orientation, but emotion intensity differs. In a complex sentence containing transition relationship, the clauses express contradictory or diametrically opposite emotion orientations, and the sentiment orientation of the whole sentence is decided by the emotion orientation of the final clause.

Example 3:

Sentence 8: “他们一边唱歌，一边游戏。”(They sang when they played games.)

Sentence 9: “因为比赛获奖，他非常高兴。”(Because he won a prize in the contest, he was very pleased.)

Sentence 10: “虽然天气越来越差，但同学们仍然兴高采烈的比赛着。”(Although weather was getting worse, the students were still playing the match happily.)

Sentence 8 is a complex sentence containing a coordinate relationship, in which both clauses coincide and express the same emotion orientation. Sentence 9 is a complex sentence containing causal relationship, in which the second clause is the result. Sentence 10 is a complex sentence containing transition relationship, in which the first clause describes a state, while the second one reflects the behavior in this state.

The above analysis shows that conjunction is the key to the identification of the sentiment orientation of complex sentences. Some conjunctions appear in pairs, while some appear separately, and that different types of conjunctions have different effects on the sentiment orientation of complex sentences. A conjunction can be used to divide a complex sentence into two clauses to identify their emotion orientation separately, and then this conjunction can be used to judge the orientation of the whole sentence. In Sentence 8, conjunction expresses a coordinate relationship, and both

clauses reflect happiness, so happiness represents the sentiment orientation of the entire sentence. Sentence 9 is a causal complex sentence, and both clauses express happiness, so the whole sentence has sentiment orientation of happiness. Sentence 10 is a transitional complex sentence, and the first clause expresses an anxious emotion, while the second one expresses a happy emotion. The sentiment orientation of the whole sentence depends on the second clause, so its emotion orientation is happiness.

To sum up, in a complex sentence containing a coordinate relationship or causal relationship, the sentiment orientation of the whole sentence is basically consistent with that of the clauses, so these two kinds of complex sentences can be treated as a simple sentence. However, for a transitional complex sentence, the sentiment orientation of the second clause is contradictory and even diametrically opposite to that of the first one, while the sentiment orientation of the whole sentence is always decided by that of the second clause. Transitional complex sentence is the important task in the paper.

For transitional complex sentence, we construct a conjunction dictionary consisting of 45 conjunctions based on the corpus. The conjunctions in this conjunction dictionary mainly reflect transition relationship, concessive relationship and hypothetical relationship. Some is listed in Table.3.3 for part of the conjunctions.

Table.3.3 Conjunction Dictionary

Relationship	Quantity	Example
Transition Relation	12	But, however, yet, and while
Concessive Relationship	7	Although, though, despite and even if
Hypothetical Relationship	26	If, in case, once and suppose

In a transitional complex sentence, the first clause is called front sentence (FS), and the second clause is called back sentence (BS). In Figure.3.3, the process in which sentence sentiment orientation is identified based on sentence structure characters, as described in detail below:

(1) Input sentence *S* whose emotion orientation is to be discriminated, and judge whether there is a comma in the sentence; enter Step 2 if there is a comma; extract its features if there isn't comma to calculate the emotion orientation intensity of the sentence to identify its emotion orientation.

(2) Identify whether there is a conjunction in Sentence *S* that is included in the conjunction dictionary. Enter Step 3 if there is a conjunction. If there isn't conjunction, this sentence may be a coordinate complex sentence or a causal complex sentence.

Treat this complex sentence as a simple sentence and extract its features to calculate the sentiment orientation intensity of the sentence to identify its sentiment orientation.

(3) For a transitional complex sentence, divide it into two parts: FS and BS, and then extract their features. Afterwards, calculate the emotion orientation intensity of this sentence by Formula 3.10 to identify its emotion orientation.

$$S-emotion^{*}(S)=\alpha\times S-emotion(FS)+\beta\times S-emotion(BS) \tag{3.10}$$

$$(\alpha,\beta)\in\{(0.0,1.0),(0.5,1.0),(0.5,1.0)\}$$

where $(\alpha,\beta)=(0.0,1.0)$ means that conjunction represents a transition relationship, $(\alpha,\beta)=(0.5,1.0)$ reflects concessive relationship, and $(\alpha,\beta)=(0.5,1.0)$ reflects hypothetical relationship.

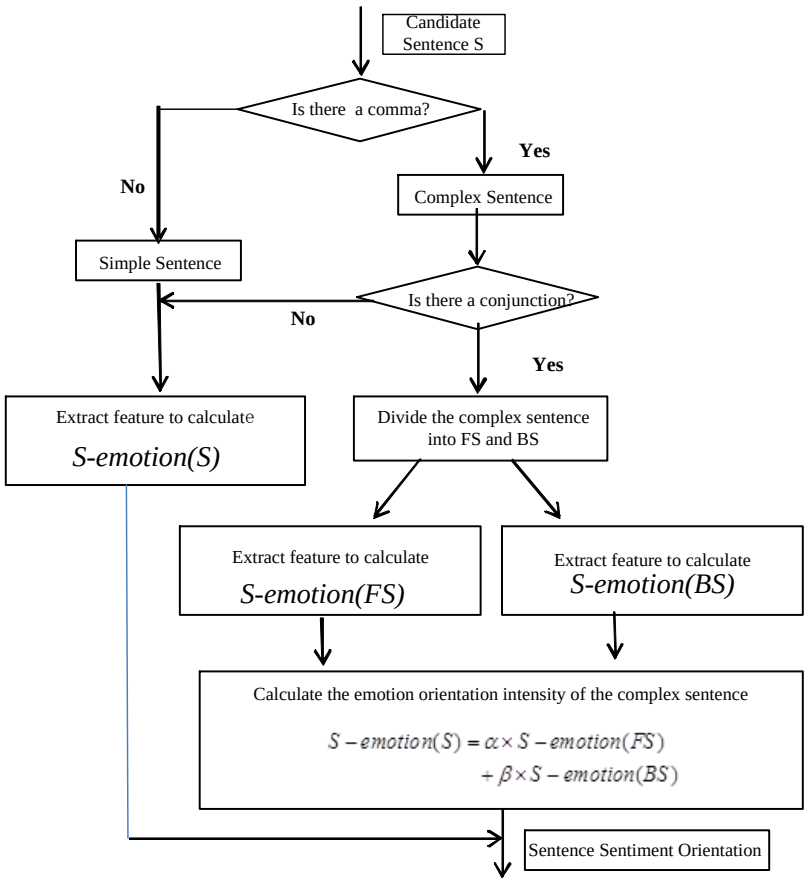

Figure.3.3 Identification Process of the Emotion Orientation of Complex Sentence

3.4 Experiments

3.4.1 Experiment Data

The experiment data in this experiment comes mainly from three data sets, two of which are Chinese emotion corpuses provided by Tan Songbo at Chinese Academy of Sciences, from which 2000 positive and negative non-repeated hotel reviews and book reviews are selected respectively. The sentiment orientations of words and sentences fall into 2 kinds: positive and negative; the other is Chinese emotional corpus Ren-CECps, in which the sentiment orientations of sentences fall into 8 kinds, which can be used for multi-label emotion orientation identification of sentences. See Table.3.4 for the statistical information of the above data sets.

Table.3.4 Statistical Information of Data Sets

Data Set	**Domain**	**Sample Size**	**Sentence Number**	**Emotion Orientation**
Hotel	Hotel Review	4000	4000	2 kinds
Book	Book Review	4000	4000	2 kinds
Ren-CECps	Blog Text	1000	21225	8 kinds

Two experiments are conducted in this chapter. In Experiment 1, hotel and book reviews are used to identity the polarity of sentence emotion. This involves the single-label classification of sentiment orientation; the other is conducted in order to identity the several emotion orientations of the sentences in Ren-CECps. This involves the multi-label classification of sentiment orientation.

In both experiments, the traditional bag-of-word-based method for sentence sentiment orientation is denoted by BOW, the topic feature-based method for sentence sentiment orientation is denoted by TM, and the multi-feature method for sentence sentiment orientation is denoted by Combine.

3.4.2 Experimental Result

3.4.2.1 Tan Data Set

For the first experiment, 20 sentences are selected randomly from the hotel review corpus to form one text, with 200 short texts generated finally. Similarly, 200 short texts are generated from the book review corpus with sentences randomly selected. The 400 short texts are taken as a data set for Experiment 1, from which 300 short texts are selected randomly as training data, while 100 as testing data. At the preprocessing stage, ICTCLAS is used for word segmentation and POS tagging for the data set. The emotion lexicon is composed of 2090 emotion words in HowNet and

6846 evaluation words in HowNet, which are applied to Experiment 1 with individual-character emotional words left out.

In this chapter, text topic features is applied to the sentence sentiment orientation identification process. Figure.3.4 shows the relationship between the identification accuracy of sentence emotion orientations and the topic features.

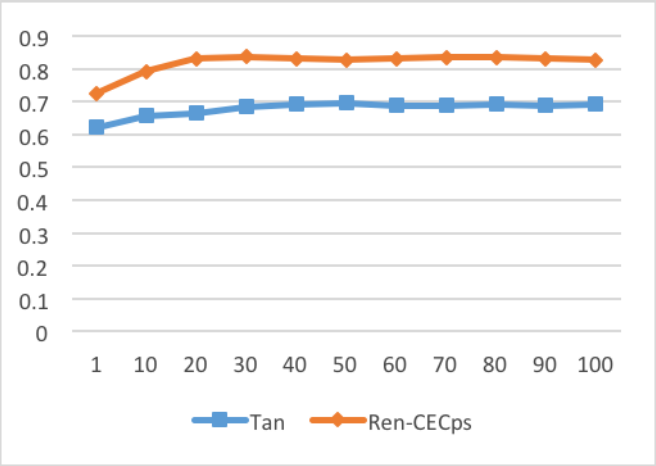

Figure.3.4 The Relationship between Sentence Emotion Orientations and Topic Feature

As can be seen in Figure.3.4, in both data sets, when the topic number increased from 1 to 10, the identification accuracy of sentiment orientations rise the fastest, and then achieved slow growth. When the topic number exceeded 30, the identification accuracy of sentence sentiment orientations would no longer rise, and even might fall sometimes. So in the experiment, it would be best the topic number is set equal to 1~30. In addition, as we can see, the identification accuracy of multi-label sentiment orientations based on Ren-CECps data is higher than that of single-label emotion orientation based on Tan data. An analysis on the features of the data sets indicates that the major reason lies in the following: in the Ren-CECps data set, the sentence relationship was closer, and the topic feature is more prominent and has a greater effect on emotional words, while in the Tan data set, the texts is generated from the sentences randomly selected from the raw corpus, and the sentences might had nothing to do with each other in topic feature, so the topic feature do not have an obvious effect.

In the Tan data set, we compare the effects of these three methods on the identification of the commendatory and derogatory meanings of sentence emotion, as shown in Figure.3.5.

As can be seen in Figure.3.5, when a text topic feature is added to the traditional bag-of-word, the identification effect of sentence sentiment orientations rise

somewhat, but not obviously. After additional features is integrated into sentence sentiment orientation identification, the effect rise significantly, suggesting that additional features are also important factors that have influence on sentence sentiment orientation identification.

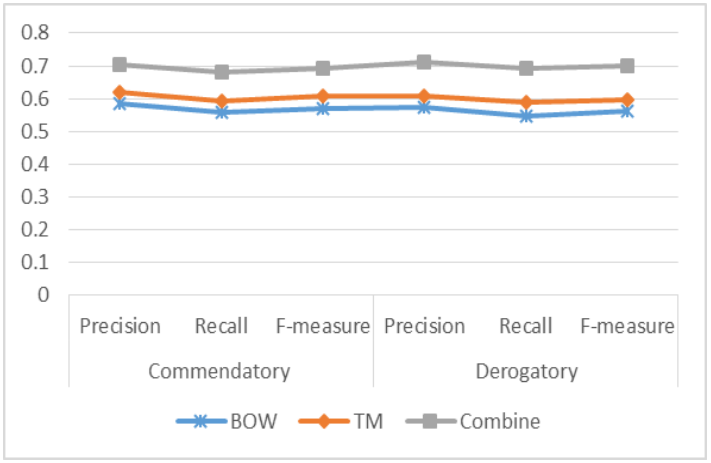

Figure.3.5 Sentence Emotion Orientations in the Tan Data Set

3.4.2.2 Ren-CECps Data Set

In the experiment, the multi-label sentiment orientations of the sentences in Ren-CECps is judged. BR (Binary Relevance) is a classic method of multi-label learning, which is particularly applicable when the label number is not large. The emotion labels in Ren_CECps are divided into 8 kinds, so BR applies to this data set. Besides, naive Bayes algorithm could be chosen as a basic classification algorithm in BR.

In Ren-CECps, the emotion orientation of each sentence is labeled as a subset of the following 8 emotion categories: surprise, sorrow, love, joy, hate, expect, anxiety and anger. 1000 texts are chosen from Ren-CECps as a data set. After a small number of neutral emotional sentences are left out, 800 texts are selected randomly as training data, and 200 texts as testing data. Emotional words are extracted from the training data to construct an emotion lexicon as required by the experiment. LDA model is adopted for topic feature discovery, and the parameters are set as follows: $\alpha = 0.5, \beta = 0.01, L = 8, T = 20$. All the above parameters are empirical values.

For Ren-CECps data set, we do the following three experiments on multi-label sentiment orientation identification.

(1) Correct identification of any multi-label sentence sentiment orientations;

(2) Correct identification of the multi-label sentence sentiment orientation with the highest intensity;

(3) Sentence sentiment orientation identification with multi-feature fusion;

In Experiment (1), macro-average and micro-average are used to compare the effects of multi-label sentence sentiment orientations by BOW, TM and Combine, as shown in Table.3.5.

Table.3.5 The Identification Result of Multi-label Emotion Orientations

	BOW	TM	Combine
Macro – accuracy	0.682	0.803	0.836
Macro – precision	0.567	0.626	0.658
Micro – accuracy	0.701	0.810	0.839
Micro – precision	0.582	0.631	0.657

Table.3.5 shows that topic feature has great effect on Chinese sentence sentiment orientation identification, and after the integration of topic feature, the identification effect of sentiment orientations increases obviously. Besides, after additional features are integrated into sentence sentiment orientation identification, the identification effect of sentiment orientations is further improved, suggesting that these additional features are also influencing factors to sentence sentiment orientation identification.

In Experiment (2), macro-averaging F1, micro-averaging F1, macro-averaging accuracy and micro-averaging accuracy are used to measure the effect of these three methods on the identification of the sentiment orientation with the maximum intensity, as shown in Table.3.6.

Table.3.6 Identification of the Emotion Orientation with the Highest Intensity

	BOW	TM	Combine
Micro-averaging F1	0.355	0.394	0.422
Macro-averaging F1	0.298	0.332	0.374
Micro – accuracy	0.574	0.663	0.697
Macro – accuracy	0.527	0.626	0.658

Table.3.6 shows that topic feature plays an important role in the sentiment identification process related to sentiment orientation intensity, and multi-feature fusion method also achieved a good experiment effect. A comparison with Experiment (1) indicates that both macro-averaging accuracy and micro-averaging accuracy decreases somewhat, suggesting that it is relatively easy to identify one or more multi-label emotion orientations, while it is relatively hard to identify the

sentiment orientation with the highest intensity

For the 8 kinds of basic sentiment orientations, we uses macro-averaging accuracy to analyze the identification effect of sentence sentiment orientations, as shown in Figure.3.6.

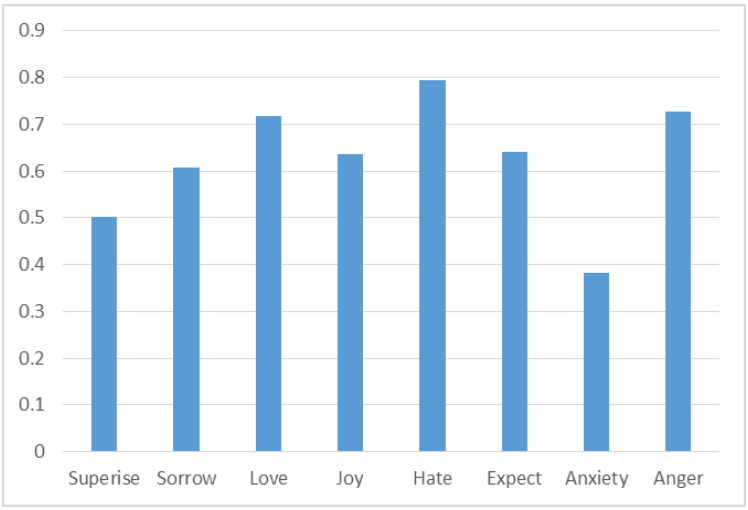

Figure.3.6 Identification of Basic Sentiment Orientations

Figure.3.6 shows that the identification accuracy of surprise and anxiety is relatively low. This is because there is little data in Ren_CECps that is related to surprise and anxiety. This defect should be remedied in future to further improve Ren_CECps.

3.4.3 Analysis of Experimental Results

For sentence sentiment orientation identification, bag-of-word and emotional words are adopted, but the identification effect is not good in both the Tan corpus and Ren-CECps. On this basis, topic feature is integrated into sentence sentiment orientation identification, leading a big increasing in the identification effect. The application of additional features, such as negative word, degree adverb and conjunction, further improve the identification effect. The experimental results indicate that emotional word, topic feature, negative word, degree adverb and conjunction could help identify sentence sentiment orientations, so the sentence sentiment orientation analysis method with multiple features proposed in this chapter is effective.

For the errors arising during sentence sentiment orientation identification, there may be the following two reasons:

(1) Misjudgment of the sentiment orientation and intensity of emotional words. For the emotion words not existing in the emotional lexicon, this chapter judged their the emotion orientation and intensity by the method introduced in Chapter 2, so some identification errors arise, finally causing errors in sentence sentiment orientation identification.

(2) Human emotion has subjectivity and complexity. The generation of an emotion usually gives rise to another emotion, and there is some dependence between different emotion orientations. The statistics of the emotion orientations in Ren-CECps indicate that “anger” and “hatred”, “anxiety” and “sadness”, as well as “happiness”, “fondness” and “expectation”, often appear simultaneously. This phenomenon has perplexed sentence emotion orientation identification.

3.5 Summary

We analyze the identification of the multi-label sentiment orientations of sentences, put forward a multi-feature fusion-based sentence sentiment orientation identification method, and make full use of emotional words, topic features and other additional features to identify the multi-label sentiment orientations of sentences.

In the research of sentence sentiment analysis, the focus is put on the identification of the subjective and objective sentences in texts. Specifically, for a subjective sentence, its emotion orientation and the emotional elements in it are mainly identified. The sentiment orientations of sentences fall into two major categories: commendatory and derogatory. In recent years, some scholars have realized the complexity of emotions, and the fact that commendatory and derogatory meanings are not enough for a complete description of all sentiment orientations, so they have begun to research multi-label sentiment orientation identification. The research methods primarily include dictionary-based method and corpus-based method. For the former, a complete emotion lexicon needs to be constructed and then used for sentence sentiment orientation identification, while for the latter, machine learning should be used for reference, and meanwhile the statistical information of corpus needs to be used for sentence sentiment orientation identification. We combine the advantages of dictionary information and corpus statistical information together, realizing the identification of multi-label sentiment orientations.

An emotion lexicon is constructed with the emotional words extracted from the training data, and then the emotion lexicon is used to set up an initial sentence

sentiment orientation identification model based on bag-of-word. When sentence emotions is analyzed by bag-of-word, the contextual dependency between sentences would often be ignored, so text topic feature information is introduced, thus solving the problem of the lack of the contextual dependency between sentences. The experiment results of Tan and Ren-CECps data sets show that topic feature is important to sentence sentiment orientation identification, since it can effectively improve the identification effect of sentence sentiment orientations.

Apart from emotional words and topic features, there are also other word features in sentences. Since all negative words, degree adverbs and conjunctions have effects on sentence sentiment orientation identification, these word features are also integrated into sentence sentiment orientation identification. As a result, the identification effect of multi-label sentiment orientations is further improved.

The identification of the multi-label emotion orientations of sentences is studied in the chapter, but there are also many special sentences, such as negative sentence, conditional sentence and comparative sentence in the texts. Special sentences usually contain some unique feature information, and these features are good for the identification of the sentiment orientations of special sentences. The identification of the multi-label sentiment orientations of special sentences will be a research hotspot in the future, worthy of further exploration and research.

Chapter 4

Sentence Sentiment Analysis based on Bayesian Network

At present, most of the sentence sentiment orientation researches are carried out by linguistic knowledge or machine learning. Machine learning includes supervised learning, unsupervised learning and semi-supervised learning. But for supervised learning, it is not easy to acquire the corpus resources labeled in detail. So, a semi-supervised multi-labeled emotion topic feature model (MLETM) is proposed with emotion topic features integrated into the LDA. MLETM is used to sample each sentence with emotion orientation, and sample each word with topic feature. As a result, sentence sentiment orientation can be identified.

4.1 Introduction

Text sentiment analysis aims to find human's delicate emotions in texts, so as to promote human-computer interaction, identify accurate product reviews and diagnose some mental illnesses[59,60]. In the traditional text sentiment analysis research, most scholars divide text sentiment orientations into two categories: commendatory and derogatory. This is known as single-label sentiment orientation analysis. Actually, text sentiment orientation has high complexity, so it cannot be described effectively with just commendatory and derogatory. Single emotion labels may not only narrow the coverage of emotions, but also affect the exactness of emotion forecasting. In Ren_CECps and some other corpus resources, emotion orientation is further detailed and divided into several kinds, thus realizing multi-label emotion orientation forecasting.

The two focuses of text emotion analysis research are emotion feature extraction and emotion orientation identification. At present, the above research is carried out mainly by means of linguistic rules and corpus statistics. However, with the rapid development of Internet applications, constantly emerging net neologisms, ever-changing linguistic expression ways and complex language processing have complicated the linguistic rule-based text emotion analysis method, so that this method hardly applies to different language environments and different types of texts.

For the corpus statistics-based method, text modeling and machine learning are two core contents. Machine learning mainly includes supervised learning, semi-supervised learning and unsupervised learning. In the supervised learning and semi-supervised learning-based method, lots of labeled training samples need to be adopted in the training process. However, the manual labeling process of samples is relatively time-consuming, strenuous and costly. Unsupervised learning method can be used for classifier construction even there is no labeled training sample.

Text modeling has been researched for many years. Vector space modal (VSM) is the most common method for text modeling. VSM is used to represent a document with dictionary space. That is, a document is a one-to-many document-word mapping. With the development of natural language processing and text processing techniques, people have further deepened the understanding and research of texts, finding that VSM ignores synonymy and polysemy, as well as the semantic relations between words. To mine the latent semantic meanings hidden in texts, people begin to seek for a text representation model that is better able to reflect text semantic meanings. Latent semantic analysis (LSA) is a mode of thinking that can find a certain latent semantic structure between words and use this latent semantic structure to represent words and texts to make it unnecessary to represent documents in the dictionary space. Its main idea is to put a semantic dimension between texts and words and then use a linear algebra method to lower the high-dimension vector of documents to the low-dimension latent semantic space. With the development of statistics, probabilistic latent semantic analysis (PLSA) has overcome some defects of LSA, so probability model can be used to represent the relationship between "documents, potential semantic meanings and words" to map documents and words into the same semantic space. So, inter-document similarity relation can be quantized through calculation of the included angle of the semantic space. However, PLSA model sometimes brings about over-fitting phenomena. In order that over-fitting phenomena should be avoided, a widely applicable maximum likelihood estimation method is adopted in PLSA model, but in this case, the training parameters in the model may show a linear increase trend with the increase of text number in the data set. PLSA model can generate a model for the documents in the data set where it is, but cannot generated a model for new documents. The defects of PLSA urged people to seek for a better theme model. In 2003, Blei[54] et al. proposed latent Dirichlet allocation (LDA) model on the basis of PLSA. LDA model is a complete generative model. With a

good mathematical foundation and expansibility, it has been widely applied to a number of fields, such as text analysis. We put forward a semi-supervised MLETM, integrating the advantages of semi-supervised machine learning and LDA model together. It then samples the emotion orientation of each sentence in the texts and samples the topic feature of each word, thus solving the problem about sentence emotion orientation identification and word topic discovery.

4.2 Latent Dirichlet Allocation

LDA model is a probability generative model, characterized by the independence between the scale of parameter space and the number of documents, so it applies to the processing of large-scale corpuses. At present, some scholars have applied LDA model to text emotion analysis. Titov[61] et al. proposed a multi-granularity LDA model (MG-LDA) based on online reviews. Afterwards, they used topic feature in the generation process of text emotion summarization, putting forward a multi-topic emotion model known as MAS [62]. Although they proved the satisfactory effect of MG-LDA on fine-grained topic extraction by experiment, for MG-LDA, a data set containing labeled should be adopted for training, so it is a supervised learning method and has some defects such as difficulty of sample data and poor domain portability. Maximum entropy method is combined with topic method in the ME-LDA proposed in reference[63], and supervised learning method was adopted as well, so a mass of labeled sample data is needed.

In order to maintain the advantages of unsupervised learning method and acquire fine-grained topic features, many scholars improved topic model from different perspectives. Brody[64] et al. directly treated every single sentence reviewed online as a document, building a relationship between “sentences, topic and words”. This method don’t consider the inter-sentence connection, but actually, entirely different words might have the same topic and be distributed in different sentences. In addition, this method is just used for emotion identification of topic word, not for sentence sentiment identification. Jo[65] et al. argued that all words in a sentence should have the same topic feature and emotion feature. Therefore, during topic label sampling, each sentence rather than each word is sampled, and a relationship is built between “documents, topic and sentences”. This method weakened the topic-related connection between words.

Topic-emotion mixing model blended topic feature and emotion feature together based on LDA model to analyze text emotion problems. It has two forms of expression in the linguistic model. The first form of expression is to fuse topic feature and emotion feature together to make up a single linguistic model. In this model, each word may be related to both topic and emotion, such as the ASUM model proposed in reference[65] and the JST model proposed in reference[66]. The second form of expression is to separate topic feature from emotion feature. Each word in a text has either emotion feature or topic feature, such as the TSM model proposed in reference[67]. TSM model divides all words of text into two parts: topic word and emotion word, and holds that emotion word has nothing to do with topic feature discovery. As a matter of fact, an emotion word is an important term that can reflect topic feature and has topic feature, so it should be considered part of topic word.

In the MLETM proposed in this chapter, each word is related to topic and emotion, and this is the biggest difference from TSM. Reference[64] just identified the sentiment orientation of topic word, and didn't analyze that of document or sentence. In this chapter, MLETM model samples each sentence with emotion, and samples each word with topic and emotion. This not only accords with the emotion expression way, but also won't weaken the topic-related connection between words.

4.3 Sentence Sentiment Analysis

4.3.1 Basic Idea

In the traditional topic issue research, all topics are regarded as the probability distribution of words, and a topic represents a word set. It is unnecessary to learn about the concrete semantic meaning of a topic before its generation, since its semantic meaning will be understood from a set of words related to it.

According to the close observation and analysis of Ren-CECps, there is some relation between the topic feature of the words in a sentence and the emotion feature of this sentence. For instance, "spring" and "children" are often associated with "happiness" and "fondness", while "war" and "confrontation" are often associated with "hatred". Based on such an observation, we propose an LDA sentence sentiment orientation identification model, linking the topic feature of words with the emotion feature of the sentence concerned.

In order to describe the relation between emotion feature and topic feature, an MLETM is proposed to identify the complex, diverse of sentiment orientations of sentences. This model is a hierarchical Bayesian network structure, by which emotion variable is introduced into the LDA model. Its basic idea: in the emotion topic model, word distribution depends not only on document topic feature, but also on the emotion feature of the sentence concerned. The LDA model is improved, with sentence emotion feature introduced into topic feature. This model is a semi-supervised learning model, in which sentence emotion generation depends on the training set, while topic generation is in accordance with the LDA model, known as unsupervised learning.

For most corpuses, unbalanced distribution is a difficulty problem to overcome. In Ren-CECps there are many sentences containing “fondness” and “happiness”, while few sentences containing “surprise”, and this may affect the judgment of infrequent sentence emotion orientations.

4.3.2 Multi-Label Emotion Topic Model

4.3.2.1 Structure of MLETM

The multi-label sentiment orientations of sentences is researched based on Ren-CECps. There are 1487 Chinese blog articles in this emotional corpus, and each article is labeled with 8 emotion orientations at the level of document, paragraph, sentence and word. Table.4.1 shows the distribution of the 8 emotion orientations included in Ren-CECps.

Table.4.1 Distribution of the 8 Basic Emotion Orientations

Sentiment Orientation	Document Number	Paragraph Number	Sentence Number
Surprise	124	503	1118
Expect	656	2145	4588
Joy	565	2740	6211
Sorrow	693	3643	8166
Hate	335	1589	3555
Anxiety	762	4128	10115
Love	911	4991	11866

Anger	189	900	2221
Total	4205	20639	47840

An MLETM is used to identify the multi-label sentiment orientations of sentences. Figure.4.1 shows the generative graphic structure of MLETM.

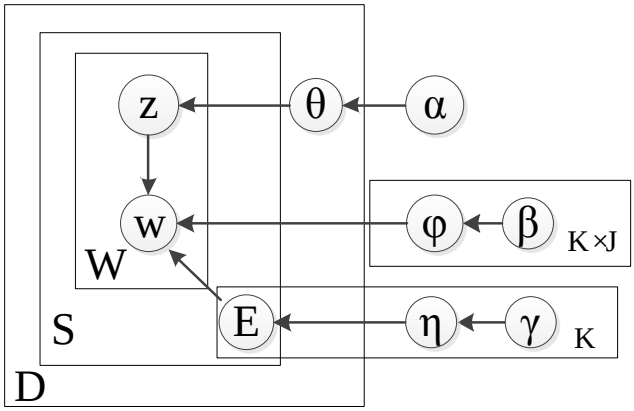

Figure.4.1 Graphic Structure of MLETM

According to the basic idea of MLETM[68], topic z is found in words, while emotion variable E is introduced into sentences, thus developing a graphic structure for MLETM.

In Figure.4.1, node represents random variable such as word node w, and directed edge $z \to w$ describes the condition dependence between word variable and topic variable. In the graphic model structure there are three kinds of variables: class variable, proportional variable and observation variable. E, z and w are class variables, and their value is discrete data, representing node object. In order to identify the multi-label sentiment orientation of sentences, we define $K(K=8)$ binary random variables E_{dsk} to indicate whether the sth sentence in text d has an emotion orientation belonging to category k. So, we generate K different planes on emotion node E to represent K categories of emotion orientations. As the ith word in the sth sentence in text d, w_{dsi} is connected with parameter node φ, topic node z and emotion node E, so it follows $(\varphi \to w)$, a random distribution based on parameter φ, and meanwhile is affected by theme random variable z and emotion random variable E. θ, η and φ are proportional variable, respectively represented as the prior probability of E, z and w. θ_d is a J -dimensional vector, and each θ_{dj} represents the prior probability of the dth topic in text d. The dependence between variable z and θ_d is reflected by $\theta \to z$. η is a K-dimensional vector, which can be used to calculate the prior probability of different emotion orientations. φ is a $K \times J \times N$ -dimensional vector, and φ_{kjt} can be

used to calculate the probability distribution of w, the tth word based on topic and emotion. Since sentences may have emotion orientation k, or not have emotion orientation k, variable φ_{kjt} is divided into two cases for differentiation and description. φ^1_{kjt} represents the prior probability of the tth word w that has topic j and emotion k, and φ^0_{kjt} represents the prior probability of the tth word w that has topic j but no emotion k. α, β and γ are observation variables acquired from the training set, and all the above proportional variables rely on them. γ is a K-dimensional vector, and in K planes, γ^1_k represents the number of the sentences containing emotion orientation k in the training data set, while γ^0_k represents the number of the sentences not containing emotion orientation k in the training data set. β is a word-related observation variable, a $K \times J \times N$-dimensional variable. Variable φ^1_{kjt} corresponds to variable β^1_{kjt}, while variable φ^0_{kjt} corresponds to variable β^0_{kjt}. A detailed description of all parameters in the model is shown in Table.4.2.

Table.4.2 The Meaning of All Parameters in MLETM

D	Document	w	Concrete word
S	Sentence	z	Theme
W	Words in a sentence	K	Number of emotion orientations
N	Word number	J	Theme number
α	J-dimensional vector, and α_j represents the number of theme j in the training data set	E	K-dimensional vector, representing a multi-label emotion orientation
β^0, β^1	$K \times J \times N$ vector, β^1_{kjt} represents the number of the words with theme j and emotion orientation k in the training data set, and β^0_{kjt} represents the number of the words without theme j and emotion orientation k in the training data set	φ	$K \times J \times N$ variable, φ^1_{kjt} represents the prior probability of word w with theme j and emotion orientation k, and φ^0_{kjt} represents the prior probability of word w without theme j and emotion orientation
γ^0, γ^1	γ^1 represents the number	θ	$D \times J$ -dimensional

	of the sentences containing emotion orientation k in the training data set, while γ^0 represents the number of the sentences not containing emotion orientation k in the training data set		variable, and θ_{dj} represents the probability of word w with theme j in document d
The η	K -dimensional variable, and η_k represents the prior probability of the kth emotion orientation		

4.3.2.2 Probability Hypothesis

In this section, we will put forward a probability hypothesis for all condition dependences in the MLETM. According to model definition, the directed edge in the graphic structure reflects the condition dependence between random variables. For example, directed edge $z \to w$ reflects the condition dependence between word w and topic z.

In the graphic structure, word w has three input directed edges, which come from emotion node E, topic node z and parameter node φ respectively. According to the Bayesian theory, word w depends on these three nodes. However, since each sentence S in text d has K emotion orientations E_{dsk}, each K emotion orientations E_{dsk} and topic z_{di} codetermine the probability distribution of word w. In addition, suppose every E_{dsk} is independent of another. Suppose word w follows the categorical distribution of random variable φ, and meanwhile condition depends on emotion E and topic z, then the hypothetical formula is expressed as below:

$$w_{dsi} \mid E^1{}_{dsk}, z \sim Categorical(\varphi^1{}_{E_{dsk} z_{di}}) \quad E_{dsk} = 1 \tag{4.1}$$

$$w_{dsi} \mid E^0{}_{dsk}, z \sim Categorical(\varphi^0{}_{E_{dsk} z_{di}}) \quad E_{dsk} = 0 \tag{4.2}$$

Topic node z has an input directed edge, which comes from variable θ. Suppose every topic z_{di} is independent of another and obeys categorical distribution, then the conditional hypothetical formula of topic variable z_{di} is shown as follows:

$$z_{di} \sim Categorical(\theta_d) \tag{4.3}$$

Suppose K emotion orientation classifiers are independent of each other. In this case, the interaction between the E_{dsk} can be removed, and that each E_{dsk} follows a Bernoulli distribution of parameter η. So, the hypothetical formula is followed:

$$E_{dsk} \sim Bernoulli(\eta_k) \tag{4.4}$$

Bernoulli distribution is also known as 0-1 distribution. It is a discrete probability distribution. If Bernoulli experiment succeeds, variable will be equal to 1 and the success probability to η_k; if Bernoulli experiment fails, variable will be equal to 0 and the success probability to $1-\eta_k$.

The proportional variable of word φ corresponds to whether there is a sentence emotion orientation, so $\varphi^1{}_{kjt}$ is defined to represent represents the tth word w that has the jth topic and the kth emotion orientation, while $\varphi^0{}_{kjt}$ is defined to represent the tth word w that has the jth topic but no the kth emotion orientation. Suppose they obey the Dirichlet distribution of parameter β, and the hypothetical formula is shown below:

$$\varphi^1{}_{kjt} \sim Dirichlet(\beta^1{}_{kjt}) \tag{4.5}$$

$$\varphi^0{}_{kjt} \sim Dirichlet(\beta^0{}_{kjt}) \tag{4.6}$$

K-dimensional random variable η describes the prior probability of dualistic emotion orientation classifier. Suppose η follows the conjugate prior Beta distribution of Bernoulli distribution of parameter γ, then the hypothetical formula is shown below:

$$\eta_k \sim Beta(\gamma^1{}_k, \gamma^0{}_k) \tag{4.7}$$

Beta distribution is a density function for the conjugate prior distribution of Bernoulli distribution and binomial distribution, and the parameters in Beta distribution can be understood as false counts. γ^1 represents the number of the sentences containing emotion orientation k in the training data set, while γ^0 represents the number of the sentences not containing emotion orientation k in the training data set.

4.3.2.3 Model Inference

In the MLETM, latent variable is used to represent the features demanding forecasting, such as emotion variable E and topic feature z. The value of latent variable can be derived according to a given probability distribution hypothesis and observation variable value. This process is called derivation process. At present, some approximate algorithms, such as mean-field variation, Gibbs sampling and loopy belief propagation, have been widely applied to the derivation of various graphic model structures. The basic idea of Gibbs sampling is to create mutually independent variables for each piece of observed data, correct problematic variables in accordance with the observed value, and merge the observed data into the sampling process. For observed data, the probability distribution of the remaining variables is a posterior probability.

In this chapter, collapsed Gibbs sampling, an improved version of Gibbs sampling, is adopted. The value of latent variable E and z can be efficiently predicted if proportional variable is regarded as an intermediate result when this algorithm is used to predict latent variable E and z.

Latent variable value can be derived in accordance with the given probability distribution hypothesis and the observed value in the training data. Since E_{dsk} depends on other variables and Bayesian network knowledge, each sentence emotion orientation classifier E_{dsk} is directly proportional to the prior probability and likelihood probability, as shown in Formula 4.8.

$$p(E_{dsk} \mid w, z, E_{-dsk} : \alpha, \beta, \gamma) \propto p(E_{dsk} \mid E_{-dsk} : \gamma)\, p(w_{ds} \mid w_{-ds}, z, E : \alpha, \beta, \gamma) \tag{4.8}$$

Here, E_{-dsk} represents the collection of all sentence emotion orientation classifier but E_{dsk}, and $p(E_{dsk} \mid E_{-dsk})$ represents the prior probability of E_{dsk}. This probability can also be described with the *kth* entity of proportional variable η. According to condition hypothesis, variable η follows Beta distribution, as shown in Formula 4.9.

$$\eta \mid E_{-dsk} \sim \begin{cases} Beta(\mathrm{n}^1{}_1+\gamma^1{}_1,\cdots,\mathrm{n}^1{}_K+\gamma^1{}_K) & E_{dsk}=1 \\ Beta(\mathrm{n}^0{}_1+\gamma^0{}_1,\cdots,\mathrm{n}^0{}_K+\gamma^0{}_K) & E_{dsk}=0 \end{cases} \tag{4.9}$$

$n^1{}_k$ represents the number of the sentences containing emotion orientation k in the training data set, while $n^0{}_k$ represents the number of the sentences not containing emotion orientation k in the training data set. The factorization of the likelihood function of word w_{dsi} in sentence is as shown in Formula 4.10.

$$p(w_{ds} \mid w_{-ds}, z, E; \alpha, \beta, \gamma) = \prod_{i \in W_{ds}} p(w_{dsi} \mid w_{-dsi}, z, E; \alpha, \beta, \gamma) \tag{4.10}$$

According to the probability hypothesis, the probability of proportional variable φ follows Dirichlet distribution, as shown in Formula 4.11 and 4.12.

$$\varphi^1 \mid w_{-ds}, z, E \sim Dirichlet(n^1{}_{kz_{di}1} + \beta^1{}_{kz_{di}1}, \cdots, n^1{}_{kz_{di}N} + \beta^1{}_{kz_{di}N}) \quad E_{dsk} = 1 \tag{4.11}$$

$$\varphi^0 \mid w_{-ds}, z, E \sim Dirichlet(n^0{}_{kz_{di}1} + \beta^0{}_{kz_{di}1}, \cdots, n^0{}_{kz_{di}N} + \beta^0{}_{kz_{di}N}) \quad E_{dsk} = 0 \tag{4.12}$$

$n^1{}_{kz_{di}t}$ and $n^0{}_{kz_{di}t}$ are a pair of observed values. $n^1{}_{kz_{di}t}$ represents the number of the words with topic z_{di} and sentence emotion orientation k, while $n^0{}_{kz_{di}t}$ represents the number of the words with topic z_{di} but no sentence emotion orientation k. So, Formula 5.8 can be further derived and expressed as Formula 4.13.

$$p(E_{dsk} \mid w, z, E_{-dsk}; \alpha, \beta, \gamma) \propto \begin{cases} \dfrac{n^1{}_k + \gamma^1{}_k}{n^0{}_k + n^1{}_k + \gamma^0{}_k + \gamma^1{}_k} \exp\left(\sum_{i \in W_{ds}} \log \dfrac{n^1{}_{kz_{di}w_{dsi}} + \beta^1{}_{kz_{di}w_{dsi}}}{\sum_t n^1{}_{kz_{di}t} + \beta^1{}_{kz_{di}t}}\right) & E_{dsk} = 1 \\ \dfrac{n^0{}_k + \gamma^0{}_k}{n^0{}_k + n^1{}_k + \gamma^0{}_k + \gamma^1{}_k} \exp\left(\sum_{i \in W_{ds}} \log \dfrac{n^0{}_{kz_{di}w_{dsi}} + \beta^0{}_{kz_{di}w_{dsi}}}{\sum_t n^0{}_{kz_{di}t} + \beta^0{}_{kz_{di}t}}\right) & E_{dsk} = 0 \end{cases} \tag{4.13}$$

The posterior probability of topic z_{di} depends on other variables, and its probability is directly proportional to the prior probability and likelihood probability, as shown in Formula 4.14.

$$p(z_{di} \mid w, z_{-di}, E : \alpha, \beta, \gamma) \propto p(z_{di} \mid z_{-di} : \alpha) p(w_{di} \mid w_{-di}, z, E : \alpha, \beta, \gamma) \tag{4.14}$$

Probability $p(z_{di} \mid z_{-di})$ can be described with proportional variable θ, and variable θ follows Dirichlet distribution, as shown in Formula 4.15.

$$\theta \mid z_{-di} \sim Dirichlet(n_{d1} + \alpha_1, \cdots, n_{dJ} + \alpha_J) \tag{4.15}$$

n_{dj} represents the number of topic j in document d, so Formula 5.14 can evolve into Formula 4.16, as shown below:

$$p(z_{di} \mid \mathrm{w}, \mathrm{z}_{-\mathrm{di},} \mathrm{E}; \alpha, \beta, \gamma) \propto \frac{n_{dz_{di}} + \alpha_{z_{di}}}{W_d + \alpha^*} \times \prod_{k \in K^1{}_d} \frac{n^1{}_{kz_{di}w_{di}} + \beta^1{}_{kz_{di}w_{di}}}{\sum_t n^1{}_{kz_{di}t} + \beta^1{}_{kz_{di}t}} \times \prod_{k \in K^0{}_d} \frac{n^0{}_{kz_{di}w_{di}} + \beta^0{}_{kz_{di}w_{di}}}{\sum_t n_{0\,kz_{di}t} + \beta^0{}_{kz_{di}t}} \tag{4.16}$$

$K^1{}_d$ is a subset of the multi-label emotion classifiers with emotion k in document d, while $K^0{}_d$ is a complementary set of $K^1{}_d$. Collapsed Gibbs sampling is described below:

Algorithm 5.1 Collapsed Gibbs Sampling Algorithm of MLETM

Input: iterations of algorithm;

Output: conditional probability of sentence emotion orientations

Step1: carry out Gibbs sampling in line with the iterations set;

Step2: calculate $p(E_{dsk} \mid w, z, E_{-dsk}; \alpha, \beta, \gamma)$ in line with chapter number, sentence number and sentence emotion category number;

Step3: calculate $p(z_{di} \mid w, z_{-di}, E; \alpha, \beta, \gamma)$ in line with chapter number, sentence number and sentence emotion category number;

4.4 Experiments

4.4.1 Experimental Data

Here we mainly research the identification of multi-label sentiment orientations. All researches are carried out based on Ren-CECps, so the experimental data comes from Ren-CECps. 1000 blog articles containing 21225 sentences are selected randomly from Ren-CECps, and each sentence is labeled as a subset of 8 basic emotion orientations. 10-fold cross validation method is adopted for the experiment. In MLETM, the value of variables α, β and γ come from the training data.

Four experiment items are set, as shown below:

(1) The relation between sentence sentiment orientation and topic;

(2) Analysis of single topic orientation identification;

(3) Comparison between MLETM and naive Bayesian method (NB);

(4) Evaluation on the $Macro-precision$ and accuracy of MLETM;

4.4.2 Relations between Sentence Sentiment Orientation and Topic

The effect of latent topic feature on multi-label sentiment orientation identification is mainly analyzed in the experiment. First two evaluation criteria are defined, as shown below:

$$Accuracy_1 = \frac{the\ number\ of\ sentences\ at\ least\ one\ emotion\ matched}{the\ total\ number\ of\ sentences} \tag{4.17}$$

$$Accuracy_s = \frac{the\ number\ of\ correct\ \text{sin}\ gle\ emotion}{the\ total\ number\ of\ predicted\ \text{sin}\ gle\ emotion} \tag{4.18}$$

$Accuracy_1$ is the percentage of the sentences, at least one of whose multi-label emotion orientations is accurately identified by MLETM, while $Accuracy_s$ is the average percentage of the 8 basic emotion orientations accurately identified by MLETM. The relationship between latent topic feature and sentence sentiment orientation is shown in Figure.4.2.

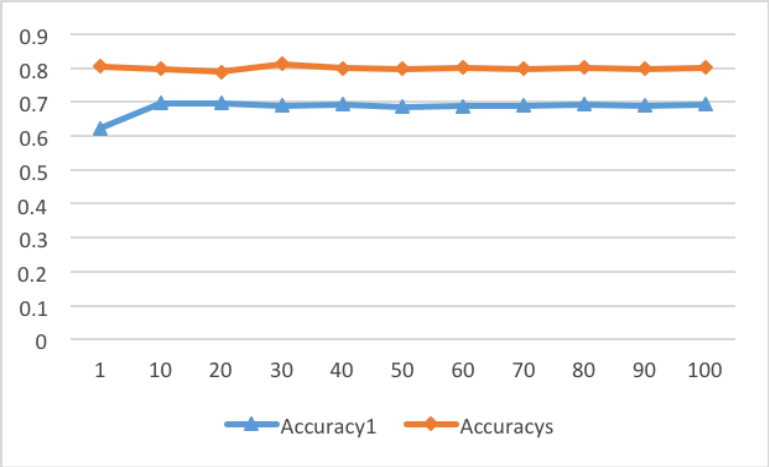

Figure.4.2 The Relationship between Topic and Sentence Emotion Orientation

Figure.4.2 shows that the identification accuracy of the multi-label sentiment orientations of sentences fluctuated with topic number increasing. When number increased from 1 to 10, the identification accuracy of single emotion orientation $Accuracy_1$ increases significantly, and reaches its maximum as the number increases to 10. For $Accuracy_s$, its value always changes with topic number increasing, and reaches its maximum when the number increased to 30. When topic number exceeded 30, both $Accuracy_s$ and $Accuracy_1$ begin to increase slowly.

In the identification process of multi-label sentiment orientations, topic number was an important factor to the complexity of MLETM.

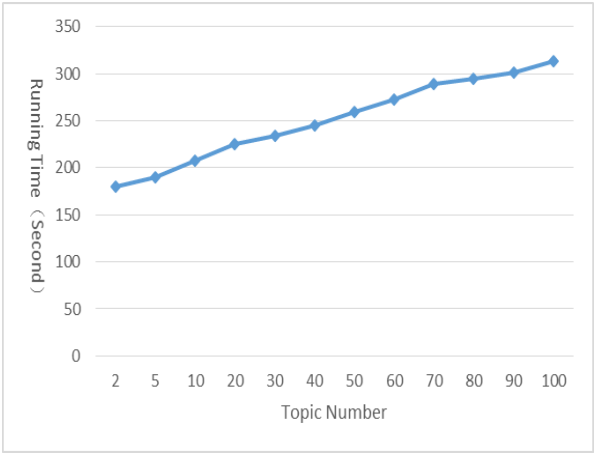

Figure.4.3 Change in the Running Time of MLETM

Figure.4.3 shows that during 1000 iterations performed for the test of 100 documents containing 2096 sentences, the running time of MLETM increases with topic number increasing, suggesting that MLETM is very sensitive to topic number in the identification process of sentence sentiment orientations.

4.4.3 Identification of Single Emotion Orientation

For the identification effect of the 8 basic emotion orientations in MLETM, the evaluation criteria are shown as follows:

$$AccuracyE_k = \frac{tp_k + tn_k}{tp_k + fp_k + fn_k + tn_k} \tag{4.19}$$

$AccuracyE_k$ evaluated each emotion orientation, and calculated the percentage of the sentences whose emotion orientations were correctly identified. The effect of MLETM on the identification of single emotion orientation is shown in Figure.4.4.

As can be seen in Figure.4.4, MLETM achieves a satisfactory effect on the identification of the 8 basic emotion orientations, and the identification accuracy of each emotion orientation exceeded 60%. But Figure.4.4 also reveals a problem: the identification accuracy of "surprise" reached 90%. An analysis of the data set shows that the main reason is due to the unbalanced distribution of emotion orientations in the data set. There are relatively few sentences containing "surprise", while "surprise" couldn't be found in the overwhelming majority of sentences, so tn_k is large in Formula 4.19. As a result, the identification accuracy of "surprise" reaches a super-high point. The corpus needs further improving in future, so that the emotion orientations in it should be distributed as uniformly as possible.

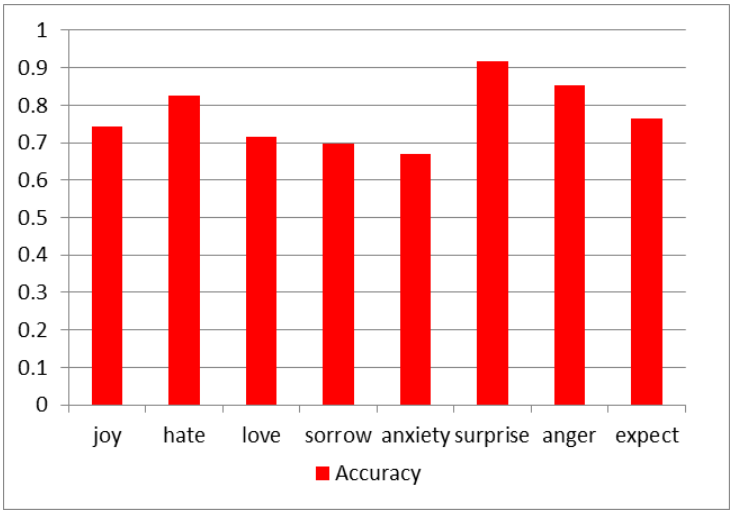

Figure.4.4 The Identification Accuracy of the 8 Basic Emotion Orientations

4.4.4 A Comparison between MLETM and NB

For the comparison experiment, two evaluation criteria were first defined, as shown in Formula 4.19 and 4.20.

$$Accuracy_2 = \frac{the\ number\ of\ sentences\ at\ least\ two\ emotion\ matched}{the\ total\ number\ of\ sentences} \tag{4.20}$$

$$Accuracy_a = \frac{the\ number\ of\ sentences\ all\ emotion\ matched}{the\ total\ number\ of\ sentences} \tag{4.21}$$

$Accuracy_2$ is the percentage of the sentences, at least two of whose emotion orientations accurately identified, while $Accuracy_a$ is the percentage of the sentences, all of whose emotion orientations are accurately identified. The effects of both methods on the identification of multi-label sentiment orientations arecompared by five evaluation standards, including $Accuracy_1$, $Accuracy_2$, $Accuracy_a$, $Macro-precision$ and $Macro-accuracy$ with NB as a reference classifier, as shown in Table.4.3.

Table.4.3 A Comparison between MLETM and NB

	NB	MLETM
$Accuracy_1$	0.040	0.687
$Accuracy_2$	0.029	0.327
$Accuracy_a$	0.022	0.154
$Macro-precision$	0.003	0.348
$Macro-accuracy$	0.807	0.798

As can be seen in Table.4.3, the $Macro-accuracy$ value of MLETM is slightly lower than that of NB, but the rest four evaluation standards are far superior to NB. $Macro-accuracy$ reflects the overall identification effect of sentence emotion orientations, involving the identification of not only the emotion orientations owned by sentences, but also the ones not owned by sentences. NB identifies the emotion orientations of most sentences as un-possession, while the sentences in the data set have fewer emotion orientations than those they didn't possess, so the $Macro-accuracy$ value of NB was slightly higher. This indirectly reflects that MLETM could correctly identify the emotion orientations of most sentences. Besides, $Accuracy_1$ was the highest, greater than 60%, while $Accuracy_a$ was the lowest, only

equal to 15%, and $Accuracy_2$ fell in between, suggesting that more emotion orientations mean more complicated emotions and lower identifiability.

4.4.5 Macro-averaging of MLETM

For evaluation of the macro identification effect of multi-label emotion orientations, two standards $Macro-accuracy$ and $Macro-precision$ are adopted, and the change rule of their iterations in Gibbs Sampling is analyzed, as shown in Figure.4.5.

As can be seen in Figure.4.5, $Macro-accuracy$ and $Macro-precision$ changes with iterations increasing. When the iterations fluctuate between 10 and 1000, the macro-averaging fluctuates as well, and when the iterations exceed 1000, the macro-averaging tends to be stable. $Macro-accuracy$ exceeds 75%, and $Macro-precision$ also reaches around 35%, suggesting that MLETM achieves a satisfactory effect on the identification of multi-label sentiment orientations, and maintaines certain stability.

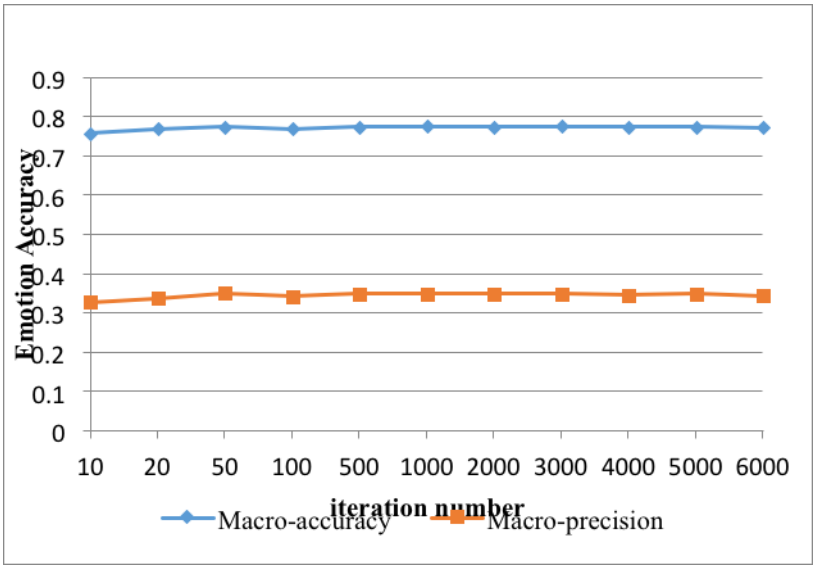

Figure.4.5 The Variation of $Macro-accuracy$ and $Macro-precision$

4.4.6 Analysis of Experimental Results

It can be seen from the above experimental results that MLETM could well identify the sentiment orientations of complex sentences: $Macro-accuracy$ reaches 75%, and the identification accuracy of all the 8 emotion orientations exceeds 60%. A comparison with NB also reveals the superiority of MLETM. However, we also see problems: with the increase in the number of emotion orientations contained in sentences, the identification ability of MLETM decreases, and $Accuracy_a$ is merely

slightly higher than 15%. This reflects not only the complexity of human emotion, but also the improvability of MLETM. The unbalanced distribution of emotion orientations in the corpus is also an important influencing factor to the identification ability of MLETM, so it is imperative to further enrich and optimize the emotional corpus in the future. Besides, multiple emotional corpuses, including voice corpus and face corpus, can be set up, so as to identify complex emotion orientations from the perspective of text, voice and face simultaneously.

4.5 Summary

We mainly analyze the identification of sentence sentiment orientations. By introducing LDA model, which is widely applied to text classification, into the research of multi-label sentiment orientation identification, a semi-supervised MLETM is proposed and applied to the identification of complex sentence sentiment orientations.

LDA model is a generative model, which helpes topic feature-based text classification by analyzing the probability distribution between document and words, as well as topic and words. We construct a generative MLETM by making an improvement in the LDA model, and then identified the complex sentiment orientations of sentences and analyzed the changes of emotion topics by extracting latent topics from the data. The main idea of this model is that there is a close relation between word topic and sentence sentiment orientation in the corpus. A combination of the priori knowledge of word and sentence emotion orientations with the corresponding probability distribution could well identify the sentiment orientations of new words. The experimental results also show that MLETM could well identify the multi-label sentiment orientations of sentences.

MLETM is also a semi-supervised learning method, and the essential data should be labeled in advance for model inference. In addition, the priori knowledge of Dirichlet and Beta should also be acquired from the training data. Besides, the latent topics generated in the model also affects the identification of sentence sentiment orientations. This is because the relative small scale of Ren-CECps make it impossible to carry on statistics from large-scale data to acquire more accurate parameters.

In the chapter, we primarily improve the LDA model and put forward a semi-supervised learning method known as MLETM, realizing the identification of multi-label sentiment orientations. The focus of future research is how to use MLETM

in text emotion analysis and other studies, and how to improve the accuracy of MLETM by combining other features of sentences, such as emotional lexicon.

Chapter 5

Document Sentiment Analysis based on Three-way Decisions

5.1 Introduction

The dramatic development of Web2.0 technology is rapidly changing the way of daily interpersonal communication. Most of the online information in texts tells something about users' personal view, attitude and emotion, including pleasure, anger, sorrow and joy, reflecting people's sentimental characteristics and sentimental changes [69-72]. Sentiment analysis of text is meant to make a judgment on the sentiment orientation of the words, sentences and text through mining and analyzing the opinions, views, emotions, and other subjective information revealed in the text. This chapter carries out sentiment analysis of Chinese text at these three levels based upon sentiment lexicon and sentiment orientation of sentences with topic feature, to identify the multi-label sentiment orientation of Chinese texts.

Sentiment analysis of document is a higher-grade form of expression in affective computing, for which there are two major research methods: supervised learning and unsupervised learning [73-77]. The method of unsupervised learning is to judge the sentiment category of document in accordance with the sentiment information of words or phrases in the document. Turkey [78] introduced a semantic orientation-based unsupervised method, and classified review articles according to the tendentious information of the commendation and derogation of word. Supervised learning method is a machine learning method by which texts are put under different sentiment categories. Pang et al [79] first applied machine learning method to the sentiment classification of a document and made a comparison in three classification models as NB, ME and SVM. Naive Bayes classifier is a simple probabilistic classifiers based on applying Bayes' theorem. ME states that the probability distribution which best represents the current state of knowledge is the one with largest entropy. SVM is a supervised learning model and can efficiently perform a non-linear classification. Jesus Serrano Guerrer et al [80] review and compare some free access web services, analyzing their capabilities to classify and score different pieces of text with respect to the sentiments.

In China, Xu Linhong et al [81] proposed an automatic identification mechanism that embraces semantic features and machine learning for Chinese text polarity. Xu Jun et al [82] researched the sentiment classification of news and comments using Naive Bayes Method and Maximum Entropy Method, and summed up the superiority and inferiority of each method through a series of experiments. Wang Suge [83] proposed a text vector representation model with strength of sentiment orientation by the use of the data representation model in rough set theory, constructing a weighted rough membership function, and applied it to sentiment classification of Chinese text. Fuji Ren et al [84,85,86] proposed some methods by using sentiment topic features to recognize the sentiment orientation of Chinese text over different level, such as words, sentences and documents.

In light of above problems, we bring forth a method for identifying sentiment orientations of documents based on three-way decisions. Sentiment orientations of documents are identified in two stages. First of all, they are classified into acceptance, rejection and delay according to affective characteristics of words. Next, sentiment orientations of documents delaying decision-making are further identified based on affective characteristics of sentences. Sentiment orientations of documents are judged by making full use of affective features of words and sentences, so as to reflect connections of these three levels (words, sentences anddocuments). The experimental result shows that sentiment analysis of document based on three-way decisions is a satisfactory one.

5.2 Three-way Decisions Model

Based on Pawlak's classical rough set theory, Yao[87,88] et al proposed the rough set theory of decision making, made a semantic explanation on the basis of Bayes' minimum-risk decision rule, depicted the probability domain by two thresholds and provided a practically effective method for threshold calculation. Based on decision acceptance or rejection, decision deferral is introduced into the three-way decisions model, thereby avoiding the losses from direct choice of decision acceptance or rejection.

For any object $x \in U$, it exists in two states, namely meeting or not satisfying given conditions. Thus, the set of objects U may be divided into two subsets, that is $U = X \bigcup \bar{X}$, where X is a set of objects meeting conditions and $\bar{X}$ is a set not in line with conditions. The action set $A = \{a_A, a_R, a_N\}$, which represents three actions

respectively, including decision acceptance, rejection and deferral. The actions may contribute to different losses, which are represented by λ_{AP} , λ_{RP} , λ_{NP} , for a_A, a_R, a_N taken by X respectively. λ_{AN} , λ_{RN} , λ_{NN} are used for representing losses of a_A, a_R, a_N when the objects don't belong to X , meeting two requirements as follows:

Requirement 1:

$$0 \le \lambda_{AP} \le \lambda_{NP} \le \lambda_{RP}, \quad 0 \le \lambda_{RN} \le \lambda_{NN} \le \lambda_{AN}$$

Requirement 2: $\dfrac{\lambda_{RP} - \lambda_{NP}}{\lambda_{NN} - \lambda_{RN}} > \dfrac{\lambda_{NP} - \lambda_{AP}}{\lambda_{AN} - \lambda_{NN}}$

For any object $x \in U$, the three-way decisions rules are as follows:

Rule A (Acceptance): If $P(X \mid [x]) \ge \alpha$, then $x \in POS(X)$.

Rule R (Rejection): If $P(X \mid [x]) \le \beta$, then $x \in NEG(X)$.

Rule N (Deferral): If $\beta < P(X \mid [x]) < \alpha$, then $x \in BND(X)$.

Where, threshold parameters α and β are calculated as follows:

$$\alpha = \frac{\lambda_{AN} - \lambda_{NN}}{(\lambda_{AN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{AP})} \qquad (5.1)$$

$$\beta = \frac{\lambda_{NN} - \lambda_{RN}}{(\lambda_{NN} - \lambda_{RN}) + (\lambda_{RP} - \lambda_{NP})} \qquad (5.2)$$

Concerning Rule A, when the probability of $x \in X$ is higher than α , acceptable rules may be created for the positive domain, and x may be included in the positive domain of X . According to Rule R, rejected rules will be developed for the negative domain when the probability of $x \in X$ is below β , and x will be categorized as a part of the negative domain in X . For Rule N, x will be included in the boundary domain of X in case that the probability of $x \in X$ ranges between α and β , while decision deferral means temporarily, no decisions are made.

Three-way decisions are made at the minimum costs. By calculating probability and thresholds of their categories, objects are categorized into positive, negative and boundary domains accordingly, which correspond to decision acceptance, rejection and deferral respectively. Being effective for processing and classifying data to reduce wrong decisions, the three-way decisions model may increase the accuracy of classification.

5.3 Document sentiment analysis

Sentiment orientation of texts is determined by sentiment characteristics of the basic elements included in text. In other words, sentiment information of words and sentences decide sentiment orientation of texts. Therefore, the core idea of this paper is that sentiment orientation of text is recognized from two different levels: words and sentences, which the three-way decisions model is applied. The detailed analysis procedure is shown here below.

5.3.1 Semantic Similarity

It is fundamental for analyzing sentiment orientations of texts by identifying sentiment orientations of words and their intensity. In this paper, a method based on Tongyi Cilin [32] and HowNet [28] to determine correlations between unknown words and seed words according to their synonymous relations and semantic similarities, in order to identify sentiment orientations of unknown words and their intensity.

5.3.1.1 Sentiment Lexion

In this paper, Ren_CECps Chinese sentiment corpus[89,90] is adopted in the experiment. After processing and labeling 1,487 Chinese blogs, 11,255 paragraphs, 35,096 sentences and 878,164 words are included in Ren-CECps.

In Ren_CECps, all language information of Chines texts associated with sentiment expressions are labeled by hands at three levels, including texts, sentences and words. The sentiment label at the level of words are essential for annotating the whole Chinese sentiment corpus. Specifically, orientation and intensity of sentiments as well as parts of speech are labeled for words and phrases.

All sentiments are divided into eight most basic categories, including surprise, sorrow, love, joy, hate, expect, anxiety and anger. Sentiment types and intensity of texts, sentences and words are represented by an 8-dimension sentiment vector as follows:

$$\vec{e} = (e^1, e^2, e^3, e^4, e^5, e^6, e^7, e^8) \qquad (5.3)$$

The value of e^i ranges from 0.1 to 1.0 and indicates sentimental intensity of a basic type of sentiments among the eight categories mentioned above. In this paper, sentiment words are extracted from training sets of Ren-CECps to make up a multi-label sentiment lexicon.

5.3.1.2 Calculation of Semantic Similarities

To analyze sentiment orientations of words, synonymous relations between unknown words and seed words of sentiment lexicon are determined by using Tongyi Cilin. If the unknown words are not included in Tongyi Cilin, the semantic similarities between unknown words and seed words shall be further measured by HowNet. In case that the unknown words exist in neither the Tongyi Cilin nor HowNet, the sentiment orientations of these words may be identified by the naive Bayes method introduced here below.

All words included in Tongyi Cilin are arranged in line with a tree-shaped hierarchical structure. In this dictionary, vocabularies are divided into three categories, including 12 in the division, 97 in the group and 1,400 in the class. The structure of Tongyi Cilin is shown in Figure.5. 1 as follows.

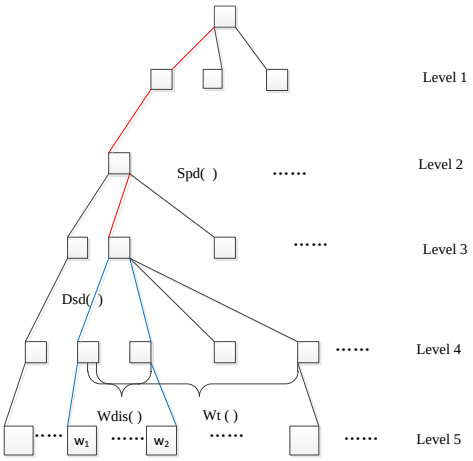

Figure.5.1 Structure of Tongyi Cilin

Definition 4 (Similarity): The path length of the common parent node of w_1 and w_2 in the hierarchical system of Tongyi Cilin is labeled as $Spd(w_1, w_2)$.

Definition 5 (Dissimilarity): This reveals that two words w_1 and w_2 gradually moves upwards along their separate parent nodes in the hierarchical system of Tongyi Cilin, until they reach a common parent node. In this case, the path they pass by is the shortest and labeled as $Dsd(w_1, w_2)$.

Based on above definitions, the semantic similarity of w_1 and w_2 may be conveyed by Formula 5.4 as follows.

$$SimC = \frac{2 \times Spd(w_1, w_2)}{Dsd(w_1, w_2) + 2Spd(w_1, w_2)} \qquad (5.4)$$

In case that two words w_1 and w_2 meet following requirements:

w_1 and w_2 belong to a common category $l(w_1) = l(w_2)$, the semantic similarity between w_1 and w_2 may be further represented by Formula 5.5 as follows:

$$SimC' = \frac{2 \times Spd(w_1, w_2) \times \alpha}{Dsd(w_1, w_2) + 2Spd(w_1, w_2) \times \alpha + Dnd(w_1, w_2)} \quad (5.5)$$

$$Dnd(w_1, w_2) = \frac{Wdis(w_1, w_2)}{Wt(Spd(w_1, w_2) + 1)} \times Cld(Spd(w_1, w_2) + 1) \quad (5.6)$$

Where, α is a control parameter, $Dnd(w_1, w_2)$ is used for measuring differences between w_1 and w_2, and $Cld(Spd(w_1, w_2) + 1)$ is an empirical parameter that indicates similarity of meaning items between w_1 and w_2 ($Cld(2) > Cld(3) > Cld(4) > Cld(5)$).

When w_1 and w_2 don't belong to a common category, or one of them isn't included in Tongyi Cilin, it will be inadvisable to calculate their semantic similarity by Formula 5.5, but by the HowNet-based semantic computation method, as shown in Formula 5.7.

$$Sim(w_1, w_2) = \max_{i=1 \cdots m, j=1 \cdots n} Sim(C_{1i}, C_{2i}) \quad (5.7)$$

5.3.1.3 Bayesian classification

Sentiment orientation of words may be determined by calculating semantic similarity of unknown and seed words. If the unknown words are not included in Tongyi Cilin and HowNet, it will be impossible to judge sentiment orientations and intensity of these words by measuring their semantic similarity. Under this situation, sentiment orientations of the unknown words may be identified by Bayes classifier.

Assuming that w is a sequence composed of several characters $(c_1, c_2, \cdots, c_n)$ and each character is characterized as a feature of the word, the probability of its sentiment orientation may be determined by the general expression for calculating the probability that all features belong to a type of sentiment orientations.

Definition 6: Prior probability of character features is defined as the probability of their appearance in different sentiment orientations, and calculated as follows:

$$P(c_i \mid e = k) = \frac{count(c_i = 1, e = k) + 1}{count(e = k) + 1} \quad (5.8)$$

Definition 7: When any unknown word appears, the probability for decision attributes to belong to different types of sentiment orientations turns into the posterior probability of combing features of characters, represented as follows:

$$P(e = k \mid w) = \frac{P(e = k) \prod_i P(c_i = 1 \mid e = k)}{\sum_k P(e = k) \prod_i P(c_i = 1 \mid e = k)} \quad (5.9)$$

Where, $P(e=k)$ may be calculated as follows:

$$P(e=k)=\frac{count(e=k)}{\sum_k count(e=k)} \quad (5.10)$$

Thus, the naive Bayes classifier may be conveyed as follows:

$$y=\arg\max_k \frac{P(e=k)\prod_i P(c_i=1|e=k)}{\sum_k P(e=k)\prod_i P(c_i=1|e=k)} \quad (5.11)$$

5.3.2 Document Sentiment Orientation

According to the "bag of words" hypothesis, a text is deemed as a set of sentiment words and phrases, which are weighted to identify sentiments. The vector space model of a text may be expressed as $D=\{w_1,w_2,\ldots,w_n\}$, where n represents number of sentiment words and phrases, and w_i is the ith sentiment word or phrase. w_i is denoted by a 8-dimension sentiment vector $e_i=(e_i^1,e_i^2,e_i^3,e_i^4,e_i^5,e_i^6,e_i^7,e_i^8)$, so the vector space model of sentiment may be further conveyed as follows for a text.

$$D=\{w_1,w_2,\ldots,w_n,e_1,e_2,\ldots,e_n\} \quad (5.12)$$

In order to facilitate weighting for identifying sentiment category of texts, Formula 5.12 is rewritten into Formula 5.13, which may be used for initially identifying sentiment category of text.

$$D=(\sum_{i=1}^{n}e_i^1\Big/n,\sum_{i=1}^{n}e_i^2\Big/n,\ldots,\sum_{i=1}^{n}e_i^7\Big/n,\sum_{i=1}^{n}e_i^8\Big/n) \quad (5.13)$$

The problem on how to distinguish multi-label sentiment orientations of texts is converted into a problem concerning identification of several binary sentiment orientations. Based on three-way decisions, an object d is categorized as certain sentiment category or not. It may be ascribed to that category or excluded from it. Hence, the state set is defined as: $\Omega=\{E_k,\neg E_k\}$, where E_k and $\neg E_k$ mean the x belongs to or is beyond E_k. The action set is defined as: $A=\{a_A,a_R,a_N\}$, which represent acceptance, rejection and deferral respectively. Based on experiences, the loss functions are shown in Table 5.1 as follows.

According to Table.5.1, formulas 5.1 and 5.2, a pair of thresholds are calculated, $\alpha=0.625$ and $\beta=0.286$. Thus, the decision rules are as follows for object d:

Rule A: If $P(E_k|d)\geq\alpha, d\in POS(E_k)$.

Rule R: If $P(E_k|d)\leq\beta, d\in NEG(E_k)$.

Rule N: If $\beta<P(E_k|d)<\alpha$,c

$P(E_k \mid d)$ is calculated as follows:

$$P(E_k \mid d) = \sum_{i=1}^{n} e_i^k \Big/ n \tag{5.14}$$

Table.5.1. Loss Functions for Decisions of Two States

Actions	Objective State	
	requirement P	requirement N
Acceptance: A	λ_{AP} : 0	λ_{AN} : 7u
Rejection: R	λ_{RP} : 8u	λ_{RN} : 0
Deferral: N	λ_{NP} : 3u	λ_{NN} : 2u

When the object d adopts deferral rules, it means the object d is impossible to directly judge the emotion E_k or not by weighing sentiment words. Then sentiment information of sentences in a text may be acquired by MLETM, and sentiment orientations of texts may be further discriminated based on sentiment characteristics of sentences.

For the object d delaying decision-making, a threshold θ is set according to sentiment characteristics of sentences, and handled as follows:

(1) If the emotional equivalence class proportion of the sentences with Emotion E_k in Text x is equal or greater than θ, we judge that Text x has Emotion E_k.

(2) If the emotional equivalence class proportion of the sentences with Emotion E_k in Text x is less than θ, we judge that Text x doesn't have Emotion E_k.

The multi-label sentiment analysis framework of Chinese texts is illustrated in Figure.5.2. Training process is in the left and testing process is in the right side. Moreover, sentiment lexicon, MLETM model and three-way decisions model are used to identify the multi-label sentiment orientation of Chinese texts. The judgment of the sentiment orientation of text falls into 6 steps, with more detail as follows:

Step 1: Select 1000 files from Ren_CECps Chinese emotion corpus as experimental data, to constitute training corpus and test corpus;

Step 2: Pre-process training corpus and test corpus respectively, to eliminate a few sentences without any emotion, and remove the stop words out of the corpuses in accordance with the stop word lexicon, to establish a sentiment lexicon.

Step 3: Based on training data set, conduct training to acquire all parameters needed by MLETM, and compute the sorts of the sentences included in the training data set;

Step 4: According to the training data set, first use the sentiment lexicon to identify the emotion polarity of texts in the test data set;

Step 5: Re-identify the multi-label sentiment polarity of texts using three-way decisions method and MLETM if the emotional recognition fails to be identified clearly;

Step 6: Evaluate the recognition result;

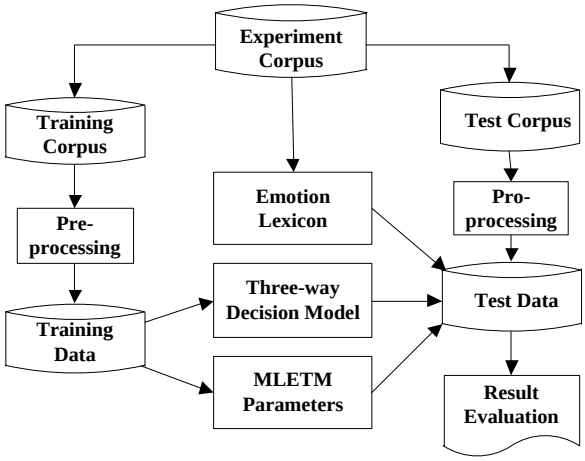

Figure.5.2 Multi-label Document Sentiment Analysis Framework

5.4 Experiments

5.4.1 Experiment Data

In this experiment, 1000 blogs are randomly selected from Ren_CECps as experimental dataset, where each blog is tagged as a subset of 8 categories of sentiment (including surprise, sorrow, love, joy, hate, expect, anxiety and anger). The distribution of sentiment orientations of texts is shown in Table.5.2.

Pre-processing of the data set: 1) remove a few sentences without any emotion out of data set; 2) Remove the stop words out of all the sentences. 3) 800 documents of the dataset make up a training set, while the testing set is composed of the remained 200 documents.

Table.5.2. Distribution of Sentiment Orientations for Texts

Sentiment Orientation	Number of Texts	Percent (%)
Surprise	70	7.0
Expect	392	39.2
Joy	356	35.6
Sorrow	427	42.7
Hate	191	19.1

Anxiety	456	45.6
Love	564	56.4
Anger	120	12.0

5.4.2 Standard of Experiment Evaluation

The experiment in the paper is aimed at recognizing the multi-label sentiment orientations of texts, and the experiment result is evaluated with a label-based evaluation method [91,92]. For some single label k, formula $M(tp_k, tn_k, fp_k, fn_k)$ is used to evaluate the classification result. In the formula, tp_k denotes correct identification of the number of the texts with emotion label k, tn_k denotes correct identification of the number of the texts without emotion label k, fp_k denotes false identification of the number of the texts with emotion label k, and fn_k denotes false identification of the number of the texts without emotion label k. The macro-average and micro-average formulas of multi-labeled classification are listed as follows:

$$M_{macro} = \frac{1}{|K|}\sum_{k=1}^{|K|} M(tp_k, fp_k, tn_k, fn_k) \tag{5.15}$$

$$M_{micro} = M(\sum_{k=1}^{|K|} tp_k, \sum_{k=1}^{|K|} fp_k, \sum_{k=1}^{|K|} tn_k, \sum_{k=1}^{|K|} fn_k) \tag{5.16}$$

5.4.3 Experiment Result

A method is developed for identifying sentiment orientations of texts according to affective features of words and sentences in combination with the thoughts of three-way decisions. The parameters of three-way decisions models may be set as: $\alpha = 0.625, \beta = 0.286$ and $\theta = 0.5$. All the above parameters are empirical values obtained from the training data set.

A multi-label sentiment recognition experiment on text is compared with Naive Bayes Method, sentiment lexicon method and the method based on three-way decisions respectively. Table.5.3 illustrates the macro-average and micro-average value in the three methods. The experimental result in Table.5.3 fully shows the superiority of the method based on three-way decisions in sentiment identification of texts.

Table.5.3 Comparison of Multi-label Sentiment Recognition

	Naive Bayes	Sentiment Lexicon Method	Three-way Decisions Method
Macro-precision	0.655	0.509	0.712

Macro-accuracy	0.521	0.613	0.751
Micro-precision	0.655	0.508	0.728
Micro-accuracy	0.554	0.636	0.778

As shown in Figure.5.3, the accuracy for identifying six basic sentiment orientations (including love, sorrow, anxiety, surprise, angry and expect) is higher than the accuracy for identifying joy and hate. The accuracy for identifying hate is low, fewer tests with sentiment orientation of hate are collected in corpus. As a consequence, models are not adequately trained, thereby impacting the accuracy of identifying hate.

8 categories of basic sentiment orientations are identified by three different methods and corresponding experimental results are also shown in Figure.5.3, from which it may be found that the three-way decisions method is advantageous in identifying a majority of sentiment orientations.

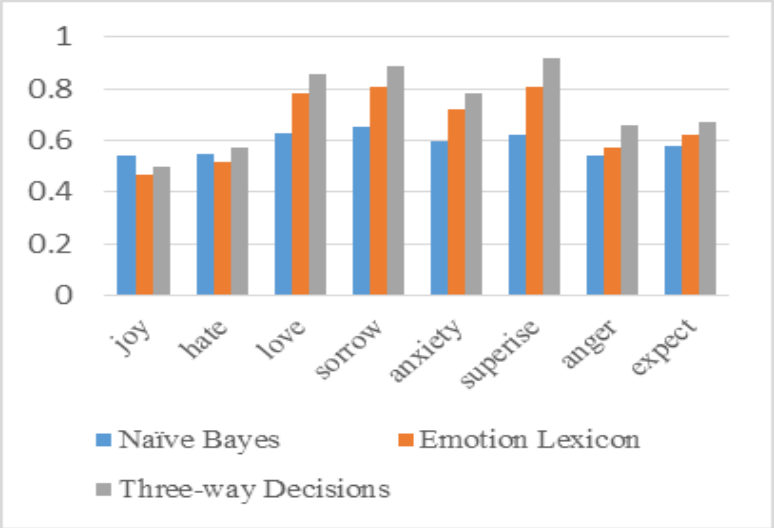

Figure.5.3 A Comparison of 8 Basic Sentiment Orientation

5.4.4 Discussion

In this section, a discussion is made to evaluate the results of our experiments and find the factors which influence the result of the multi-label sentiment recognition of Chinese texts.

The good results of the experiments show that our method in recognizing the multi-label sentiment orientation of texts is better than Naïve Bayes and sentiment lexicon method in Table.5.3 and the accuracy of the single-label sentiment orientation of 8 basic category is over 50 percent in Figure.5.4. But all of the above shows that the emotion of human beings is so complicated and the performance of experiments

also have space to improve in the future. It inspires us to study new methods and find more meaning sentimental information to improve the performance of our model. It is another important factor that sentiment distribution is not balance in corpus, so that it influences the sentiment recognition of text and some sentiment orientations of text can't be identified precisely, as shown in Figure.5.4. It is a difficult and tough task for us in the future.

5.5 Summary

We made an intensive study on the problem of multi-label sentiment analysis and proposed a three-level (word, sentence and text) multi-label sentiment analysis method based on three-way decisions model. Ren_CECps Chinese emotion corpus is adopted for the experiment in the paper.

The method proposed in the paper for analyzing sentiment orientations of Chinese texts based on three-way decision identifies the sentiment orientations by fully taking advantage of affective features of words and sentences. The experimental results prove the superiority of this method. The method for analyzing sentiment orientations of texts based on three-way decisions takes affective features of words and sentences into account to jointly discriminate the orientations, which solve the problem regarding loss of sentiment information between levels. The risk cost is introduced into the decision making, so it is necessary to seek appropriate thresholds for making decisions, in order to adopt different decision making rules for texts.

At present, WeChat has become the most common online social media with great value for research and application in identifying sentiment orientations of texts. WeChat texts are mostly concise with fewer words and sentences, but rich sentiments, flexible structures and often special sentiment symbols of the internet. Therefore, it is somewhat challenging to identify diversified and complex sentiment orientations of WeChat texts by efficiently and accurately exacting affective features from them.

Chapter 6

Summary and Future work

The research on Chinese text multi-label sentiment orientation identification is an uncertainty analysis of sentiment orientations. With granular computing thought as theoretical guidance and an analysis tool, multi-granular and multi-label Chinese text sentiment orientation analysis can not only enrich and promote the research of Chinese text sentiment analysis, but also extend the practical application of granular computing theory. We will first summarize the whole paper, and then point out the shortcomings of this paper, as well as the future research direction.

6.1 Overall Summary

The hierarchical research thought of granular computing is applied to the research of Chinese text sentiment orientations. From a perspective of linguistic granularity, text sentiment orientation research falls broadly into three levels: word, sentence and document. By full use of granular computing theory, an effective tool for uncertain information analysis and processing, we carries out an analysis on context-based word sentiment orientation identification, multi-feature fusion-based sentence sentiment orientation identification and complex, diverse document sentiment orientation identification, thus implementing research on the identification of multi-granular and multi-label Chinese text sentiment orientations. The main work of this paper is summarized as follows:

(1) Most of the previous studies of Chinese text sentiment orientations are focused on commendatory and derogatory meanings. Due to the complexity and diversity of text emotions, a multi-label emotion analysis method is proposed and applied to word-level, sentence-level and document-level Chinese sentiment orientation identification, thus effectively promoting the development of Chinese text sentiment analysis technology.

(2) The uncertainties of collocational word emotion orientations are studied. Based on the maximum entropy model and different collocational word features, we propose a maximum entropy-based word emotion orientation analysis method. First, around

unknown words, we extract multiple language features from the text, constituting a feature function for maximum entropy model. Second, under multiple feature constraint conditions, we select the sentiment orientation with the maximum entropy from all sentiment orientations and solve the problem of feature sparseness using smoothing technique. Finally, according to the emotional connection between words and the sentence concerned, we further eliminate the uncertainty of word sentiment orientations. Based on the experimental results of hotel review corpus and Ren-CECps, it proves the superiority of this method in multi-label Chinese word sentiment orientation identification.

(3) The uncertainty of sentence sentiment orientations caused by topic feature and multiple special words is researched. Since topic information has a good auxiliary effect on sentence sentiment orientation identification, we put forward a multi-feature fusion-based sentence sentiment orientation analysis method. First of all, it generated a topic-word probability distribution by means of the latent Dirichlet distribution to increase or decrease the emotion orientation intensity of the emotion words in sentences, and then calculated the sentiment orientation and intensity of sentences. This method is superior to some of the existing statistical methods. Then, it applied a few word features, such as negative word, conjunction and degree adverb, to the identification of sentence sentiment orientations, further eliminating the uncertainty of sentence emotion orientations.

(4) In the supervised text sentiment orientation analysis, large-scale and fine-grained labeled samples are not easy to acquire. So, sentence emotion feature is introduced into the LDA model, putting forward a new and semi-supervised MLETM. A new sentence emotion topic layer is added to the three-layer structure of the LDA model, to label each sentence with emotion orientation and each word in each sentence with topic, improving the probability distribution of emotion-topic. So, each word is related to topic and sentence emotion orientation, thus realizing the identification of sentence sentiment orientations.

(5) The uncertainty of document sentiment orientations is researched. Document emotion has abstraction and complexity, so direct identification of document emotion by the emotional information included in words often pays a big price and causes great deviation of sentiment orientation identification. However, three-way decisions method can well solve this problem, and especially, delayed decision alternatives is introduced on the basis of acceptance or rejection. First, an emotional lexicon is

constructed, and meanwhile thesaurus and HowNet are adopted for emotion similarity calculation. A new emotion similarity measurement method is adopted for word emotion orientation identification, and document set is divided into positive region, negative region and boundary region by statistical method and then set decision-making threshold value according to the emotion orientation intensity of document. Then, the emotion orientation and intensity of sentences are calculated in accordance with the MLETM, and afterwards a final judgment is made on the document sentiment orientations decided tardily.

6.2 Future Work

We tried to use granular computing thought in the research of Chinese text multi-label sentiment orientation analysis, and put forward multiple granular computing-based multi-label text sentiment orientation research methods in accordance with the existing main problems. Although the explorative research in this paper achieved some preliminary results, the integral research is still at the stage of theoretical analysis and experimental validation, and the defects of the model found in practical application need to be further solved. Besides, with the emergence of massive information on the Internet, the content and structure of Chinese texts will become more complex. Undoubtedly, this is going to pose a new challenge to the research of Chinese text sentiment orientations. The development of big data theory and methodology has also provided a new research idea and method for the analysis of Chinese text sentiment orientations. At the end of this paper, we further summed up the problems found in the research process, and prospected the future research work.

(1) Multi-label Text sentiment Orientation Analysis

We do research into word, sentence and document emotion orientation identified from a perspective of multi-label learning. It is easy to identify one or a few more of multi-label emotion orientations, but it is very difficult to precisely identify all emotion orientations. The unbalanced distribution of emotion orientations in the corpus data also affects the identification of multi-label emotion orientations. So, some scholars have begun to research interdisciplinary and cross-linguistic emotion orientation analysis.

(2) Sentence Emotion Orientation Analysis

Chapter 3 and 4 analyzes sentence sentiment orientations, but all this involved the identification of the sentiment orientations of subjective sentences. At present, some

scholars have been aware that objective sentences also contain emotions, so it has become a research topic as to how to identify the sentiment orientations of objective sentences. In addition, some special sentences, such as conditional sentence, comparative sentence and disjunctive sentence, have unique features. In recent years, some scholars have specially proposed to analyze the orientations of conditional sentence based on product attributes and research the identification of Chinese comparative sentences. So, it is also a future research direction about to identify the emotion orientations of special sentences according to their features.

(3) Chapter Emotion Orientation Analysis

Chapter 5 proposes a three-way decisions method for document sentiment orientation analysis, and integrates the emotional information of words and sentences into the identification process of document sentiment orientations, thus integrating the information of these three granularities together. Some scholars applied conditional random field model and multi-layer structure model to the identification of document sentiment orientations. However, none of these methods can enable the full use of the emotional information of words and sentences, while there are also lots of useful features needing further analyzing and extracting, and should be applied to document sentiment orientation analysis.

Bibliography

[1] Zhao Yanyan, Qin Bing, Liu Ting. A Survey of Sentiment Analysis. Journal of Software, 2010, 21(8): 1834-1848.

[2] V Hatzivassiloglou, KR Mckeown. Predicting the semantic orientation of adjectives. Processings of the EACL'97, Morristown: ACL, 1997, 174−181.

[3] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: ACL, 2006, 355-363.

[4] Choi Y, Cardie C. Adapting a polarity lexicon using iteger linear programming for domain-specific sentiment classification. Proceedings of the 2009 Conferences on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: ACL, 2009, 590-598.

[5] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 2009, 35(3): 399-433.

[6] Lu Y, Castellanos M, Dayal U, et al. Automatic construction of a context-aware sentiment lexicon: an optimization approach. Proceedings of the 20th International Conference on World Wide Web, New York, NY, USA: ACM, 2011, 347-356.

[7] Feldman R. Techniques and applications for sentiment analysis. Communications of the ACM, 2013, 56(4): 82-89.

[8] Huang Xuanqing, Zhao Jun. An analysis of Chinese text sentiment tendency. China Computer Society Communication, 2008, 4(2): 41-46.

[9] R W Picard. Affective Computing. MIT Press, Mass, 1997.

[10] Lin T Y. Granular computing. Announcement of the BISC Special Interest Group on Granular Computing, 1997.

[11] Miao Duoqian, Wang Guoyin, Liu Qin. Granular computing: past, present and future. Science Press, Beijing, 2007

[12] Zadeh L A. Fuzzy logic = Computing with words. IEEE Transactions on Fuzzy Systems, 1996, Vol.4 (2): 103-111.

[13] Pawlak Z. Rough sets. International Journal of Information and Computer Science, 1982, Vol.11 (5): 314-356.

[14] Xu L H, Lin H F, Zhao J. Construction and analysis of emotional corpus. Journal of Chinese Information Processing, 2008, 22(1):116−122.

[15] Finn A, Ku Shmerrick N, Smyth B. Genre classification and domain transfer for information filtering. Proceedings of the 24th BCS-IRSG European Colloquiumon Information Retrieval Research: Advances in Information Retrieval, UK: Springer, 2002, 353- 362.

[16] Yu H , H Atzivassiloglouv. Towards answering opinion questions: separating fact s from opinions and identifying the polarity of opinion sentences. Proceedings of the 2003 Conference on EMNLP, USA: ACL, 2003: 129-136.

[17] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Morristown, NJ, USA: ACL, 2004: 271- 278.

[18] Ye Qiang, Zhang Zhiqiong, Luo Zhengxiong. A study on Chinese subjective automatic discriminant analysis for sentiment analysis based on Internet. China Journal of Information Systems, 2007, 1(1) : 79- 91.

[19] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion observer: analyzing and comparing opinions on the web. Proceedings of the 14th International Conference on World Wide Web, New York: ACM, 2005: 342- 351.

[20] Yao Tianfang, Nie Qingyang, Li Jianchao. An opinion mining system for Chinese automobile reviews. The 25 annual conference of Chinese information society. Tsinghua University Press: Beijing, 2006: 260- 281.

[21] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on Human Language Technology and empirical methods in Natural language Processing. 2005, 347-354.

[22] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the International conference on Language Resources and Evaluation, 2010, 2200-2204.

[23] Pennebaker J W, Booth R J, Francis M E. Linguistic inquiry and word count. LIWC 2007, Austin, TX: LIWC, 2007.

[24] Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2004, 168–177.

[25] Kamps J, Marx M, Mokken RJ. Using WordNet to measure semantic orientation of adjectives. Proceedings of the LREC, 2004, 1115−1118.

[26] Kim S M, Hovy E. Determining the sentiment of opinion. Proceedings of the 20th International Conference on Computational Linguistics, 2004, 1367-1373.

[27] Esuli A, Sebastiani F. Determining the semantic orientation of terms through gloss classification. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005, 617-624.

[28] Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: sentiment tag extraction from WordNet glosses. Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006, 209-216.

[29] Liu Qun, Li Shujian. Semantic similarity calculation based on HowNet. Computational linguistics in Chinese, 2002, 7(2): 59–76.

[30] Li Feng, Li Fang. An new approach measuring semantic similarity in HowNet 2000. Journal of Chinese Information Processing, 2007, 21(3): 99–105.

[31] Jiang Min, Xiao Shibin, Wang Hongwei. An improved word similarity computing method based on HowNet. Journal of Chinese Information Processing, 2008, 22(5): 84–89.

[32] Lin D. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, 296–304.

[33] Zhu Yanlan, Min Jin, Zhou Yaqian. Semantic orientation computing based on HowNet. Journal of Chinese Information Processing, 2006, 20(1): 14–20.

[34] Cheng Chuangpeng, Wang Hailong. Research on selection of paradigm words in the judgment of emotional tendency. CAAI Transaction on Intelligent Systems, 2013, 8(4): 349–355.

[35] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 2003, 21(4): 315–346.
[36] Wiebe J. Learning subjective adjectives from corpora. Proceedings of the 17th National Conference on Artificial Intelligence, Menlo Park, CA, USA: AAAI Press, 2000, 735–740.
[37] Rao D, Ravichandran D. Semi-supervised polarity lexicon induction. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA: ACL, 2009, 675–682.
[38] Velikovich L, Blair Goldensohn S, Hannan K, et al. The viability of web-derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the European Chapter of the Association for Computational Linguistics, 2010, 777-785.
[39] Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation. Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009, 1199-1204.
[40] Zhang L, Liu B. Identifying noun product features that imply opinions. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: ACL, 2011, 575–580.
[41] Zhang L, Liu B. Extracting resource terms for sentiment analysis. Proceedings of the 5th International Joint Conference on Natural Language Processing, 2011, 1171-1179.
[42] Jaynes T. Information Theory and Statistical Mechanics. Physics Reviews, 1957(106): 620-630.
[43] Martin S C, Ney H, Zaplo J. Smoothing methods in maximum entropy language modeling. IEEE International Conference on Acoustics, Speech and Signal Processing, Icassp, 1999, 1(3): 545-548.
[44] Tan songbo, Wang Suge, Liao Xiangwen. The fifth Chinese Orientation Analysis and evaluation report. Proceedings of the fifth Chinese Orientation Analysis and Evaluation Conference, 2013. 5–34.
[45] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Oinion extraction and semantic classification of product reviews. Proceedings of the 12th International Conference on World Wide Web, New York, NY, USA: ACM, 2003. 519–528.
[46] Yang Bishan, Cardie Claire. Context-aware learning for sentence-level sentiment analysis with posterior regularization. Proceedings of the ACL 2014, Baltimore, ACL, 2014, 325-335.
[47] Narayanan R, Liu B, Choudhary A. Sentiment analysis of conditional sentences. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: ACL, 2009, 180–189.
[48] Khan A, Baharudin B, Khan K. Sentence based sentiment classification from online customer reviews. Proceedings of the International Conference on FIT, Islamabad: ACL, 2010, 25-29.
[49] Liu Xiaohua, Zhou Ming. Sentence-level sentiment-analysis via sequence modeling. Proceedings of the ICAIC 2011. Germany: Springer Verlag, 2011, 337-343.
[50] Zhao Yanyan, Qin Bing, Liu Ting. Integrating Intra- and Inter- document Evidences for Improving Sentence Sentiment Classification. ACTA AUTOMATICA SINICA, 2010, 36(10): 1417-1425.
[51] Song Rui，Lin Hongfei，Chang Fuyang. Chinese comparative sentence identification and comparative relations extraction. Journal of Chinese Information Processing，2009，23(2)：

102-107.

[52] Bhowmick P K. Reader perspective emotion analysis in text through ensemble based multi-label classification framework. Computer and Information Science, 2009, 2(4): 12-23.

[53] Bhowmick P K, Basu A, Mitra P. Classifying emotion in news sentences: when machine classification meets human classification. International Journal on Computer Science and Engineering, 2010, 2(1): 98-108.

[54] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

[55] Griffths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228–5235.

[56] Charles N, Sandra A Thompson. Mandarin Chinese: A functional reference grammar. University of California Press, Berkeley, Los Angeles, 1981.

[57] Lilliane Haegeman. The Syntax of Negation. Cambridge Press, New York, 1995.

[58] Chen Li, Li Baolun, Pan Haihua. The syntactic status of Chinese negative word "No". Linguistic Sciences, 2013, 12(4): 337-348.

[59] Ren Fuji, Quan changqin. Automatic annotation of word emotion in sentences based on ren-cecps. Proceedings of the 9th International Conference on Language Resources and Evaluation, Malta, 2010.

[60] Ren Fuji, Quan changqin. An exploration of features for recognizing word emotion. Proceedings of COLIN 2010, Beijing, China, 2010, 922-930.

[61] Titov I, McDonald R. Modeling online reviews with multi-grain topic models. Proceeding of WWW'08, New York: ACM, 2008, 111–120.

[62] Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization. Proceedings of ACL 2008 HLT, Stroudsburg: ACL, 2008, 308–316.

[63] Zhao X, Jiang J, Yan H F, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg: ACL, 2010: 56–65.

[64] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews. Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL, Stroudsburg: ACL, 2010, 804–812.

[65] Jo Y, Oh A. Aspect and sentiment unification mode for online review analysis. Proceedings of the 4th ACM International conference on Web search and data mining, New York: ACM, 2011, 815–824.

[66] Lin C H, He Y L. Joint sentiment/topic model for sentiment analysis. Proceeding of the 18th ACM conference on Information and knowledge management, New York: ACM, 2009, 375–384.

[67] Mei Q Z, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs. Proceeding of WWW'07, New York: ACM, 2007, 171–180.

[68] Lei Wang, Fuji Ren, Duoqian Miao. Multi-Label Emotion Recognition of Weblog Sentence Based on Bayesian Networks. IEEE Transactions on Electrical and Electronic Engineering, 2016, 11(2): 178-184.

[69] Ren Fuji: "Affective information processing and recognizing human emotion", Journal of Electronic notes in Theoretical computer science, pp.39-50, 2009.

[70] Ji Li and Fuji Ren: "Emotion Recognition of Weblog Sentences Based on an Ensemble Algorithm of Multi-label Classification and Word Emotions", IEEJ Transactions on Electronics, Information and Systems, Vol.132,pp.1362-1375, 2012.

[71] Ren Fuji: "From cloud computing to language engineering, affective computing and advanced intelligence", Journal of Advanced Intelligence, Vol.2, No.1, pp.1-14, 2010

[72] Ren Fuji and Changqin Quan: "Linguistic-Based Emotion Analysis and Recognition for Measuring Consumer Satisfaction - An Application of Affective computing", Information Technology and Management, Vol.13, No.4, pp.321-332, 2012

[73] Taboada M, Brooke J, and Tofiloski M: "Lexicon-based Methods for Sentiment Analysis", Journal of Computational Linguistics, Vol.37, No.2, pp.267-307, 2011

[74] C.Strapparava and R. Mihalcea: "Learning to identify emotion in text", Proceedings of the

[75] Kumar Ravi,Vadlamani Ravi: "A survey on opinion mining and sentiment analysis: Tasks, Approaches and Applications", Journal of knowledge-based systems, Vol 89, pp.14-46, 2015.

[76] Vidisha M Pradhan, Jay Vala and Prem Balani: "A survey on sentiment analysis algorithms for opinion mining", Journal of Computer Application, Vol133(9), pp.7-11,January 2016.

[77] Doaa Mohey El-Din Mohamed Hussein: "A survey on sentiment analysis challenges", Journal of King Saud University Engineering Sciences, April 2016.

[78] Peter D Turney: "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.417-424, 2002.

[79] Pang B, Lee L, Vaithyanathan S: "Thumbs up? Sentiment classification using machine learning techniques", Proc. of the EMNLP 2002. Morristown: ACL, pp.79-86, 2002.

[80] J. Serrano-Guerrero, J.A. Olivas, F.P. Romero and E. Herrera-Viedma: "Sentiment Analysis: A Review and Comparative Analysis of Web Service", Journal of Information Science, pp.18-38, 2015.

[81] Xu Linhong,Lin Hongfei,Yang Zhihao: "Text Orientation Identification Based on Semantic Comprehension[J]", Journal of Chinese Information Processing, pp.96-100,2007.

[82] Xu Jun,Ding Yuxin,Wang Xiaolong: "Sentiment Classification for Chinese News Using Machine Learning Methods[J]", Journal of Chinese Information Processing, pp.95-101,2007.

[83] Wang Suge, Research of sentiment classification Based on the Web Comments, 2008.

[84] Fuji Ren, Kazuyuki Matsumoto: "Semi-automatic Creation of Youth Slang Corpus and Its Application to Affective Computing", IEEE Transactions on Affective Computing, 2015,DOI: 10.1109/TAFFC.2015.2457915

[85] Fuji Ren, Xin Kang, Changqin Quan: "Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model", IEEE Journal of Biomedical and Health Informatics, DOI:10.1109/JBHI.2015.2459683.

[86] Changqin Quan, Fuji Ren: "Feature-level sentiment analysis by using comparative domain corpora", Enterprise Information Systems, 2014. DOI: 10.1080/17517575.2014.985613

[87] Yao Y Y: "An outline of a theory of three-way decisions", Proceeding of the RSCTC 2012, LNCS(LNAI)7413, pp.1-17, 2012.

[88] Yao Y Y: "Three-way Decisions with Probabilistic Rough Sets[J]", Journal of Information Science, Vol180, pp.:341-353, 2010.

[89] Changqin Quan and Fuji Ren: "A blog emotion corpus for emotional expression analysis in Chinese", Journal of Computer Speech & Language, Vol.24, No.4, pp.726-749, 2010

[90] Fuji Ren: Document for Ren-CECps 1.0, http://a1-www.is.tokushima-u.ac.jp/member /ren/Ren-CECps1.0/Ren-CECps1.0.html , 2009.

[91] G.Tsoumakas, I. Katakis: "Multi-Label Classification: An Overview[J]", International Journal of Data Warehousing and Mining, Vol3(3), pp.1-13, 2007.

[92] G.Tsoumakas, I.Vlahavas. Random K-Labelsets: "An Ensemble Method for Multilabel Classification", Proceedings of the 18th European Conference on Machine Learning(ECML2007), Warsaw, Poland,pp. pp.406-417, 2007.