Advances in Engineering Research, volume 114
International Conference on Advances in Materials, Machinery, Electrical Engineering (AMMEE 2017)

Study of Text Genre Classification and Sentiment Analysis Technology in the Information Age

Bing Fang, Longyin Du, Sheng Liang and Wei Huang
College of Mathematics and Information Science of University of Guiyang; Guiyang 550005 China

Keywords: information age; genre classification; emotion analysis; technical research.

Abstract. By applying emotion analysis on the n etwork text, enterprises can obtain the user experience and the Government can get public opinion, which is helpful for the two to take corresponding measures. This paper studies the technology of text genre classification first. A corpus needs to be built, so that the article genre can be distinguished according to the difference in sentence structure and vocabulary usage. And then, a research is given to the emotion analysis technology. Through collecting emotion words and establishing thesaurus for the emotion vocabulary, it compares the frequency of emotion words in the analysis object with the thesaurus to analyze its emotion orientation.

1. Introduction

The more efficient information transmission and larger transfer volume in network era brings difficulties to the classification of information. Moreover, it also poses some problems. For one thing, it is difficult for the government's regulatory authorities to supervise the overwhelming information due to the slow progress of the network real name system. Thus, there are always some people publish opinions which affect the building of a harmonious society and harm others. For another thing, it is also more difficult for users to get useful contents by screening the massive information. As a communication place of current social residents, network needs strong supervision. Therefore, it needs to be distinguished by the text genre classification technology and then given differentiated emotion analysis. On the one hand, the government can strengthen the control of the network public opinions to promote the construction of a harmonious society; On the other hand, the enterprise can get the users' emotion orientation to adjust its products or services accordingly; furthermore, the supervision department can also get some information about the credit behavior or product quality of enterprises, so as to enhance the efficiency of supervision. Therefore, we have a simple study of the genre classification technology and explore the emotion analysis technology in the context of the information age.

2. The Genre Classification of Network Text

2.1 The Definition And Types of The Genre

The genre of text is the specification and model which form the article. It also a distinction of the article style, which is generally divided into the following types: official style; political style; literary style; scientific style [1]. Different genres have different characteristics. It is the different content formed by the change of human ideological activities in the social development and it has certain recognition.

With the development of Internet technology, the network text genre becomes diversified while evolving fast. The boundaries between them are not so clear in the traditional articles and the boundaries even associate with each other. But, in general, genre still has its relatively stable characteristics. Taking poetry for example, it can be seen that it is always a kind of genre which shows strong emotion or magnificent imagination relying on the short refining language and good rhythm. The genre classification of article is conducive to promoting the efficiency of human communication, helping them find valuable information from tangled information [2].

2.2 The Automatic Classification of Network Text Genre

2.2.1 The Overview of Classification Principle

With the development of the article, styles are also changing. The genre has been very rich today. The frequency of different sentence structures and the use situation of some vocabularies are different in different genres. The initial simple distinction can be achieved according to these characteristics in the automatic classification. For example, the following Table 1 and Table 2 show the statistical comparison of the use of sentence structure and use situation of vocabulary in different genres.

Table 1. The comparison of sentence structure frequency in four styles of articles (frequency of occurrence in every 1000 words)

Genre	Relative clause	Complement clause	Adverb clause
Report	5.1	0.5	3.5
Official document	8.5	0.2	1.7
Conversation	2.5	3.4	4.2
Speech	7.9	1.7	7.7

Table 2. The statistics of the use nature of the word “de” (based on a 1.2-million-word corpus)

Genre	Literature and art	Science and technology	Political comment	Spoken language
Sentence with “de” structure	766	6	85	153
Transitive verb before “de”	265	3	49	83
Intransitive verb before “de”	258	2	18	53
Adjective before “de”	247	3	21	19

Based on the above tables, a primary conclusion can be received that, the genre can be distinguished according to the difference in sentence structure and use situation of word, and there are big differences between different genres. Therefore, in the genre classification of network text, it needs to collect various kinds of network text by big data and artificially building corpus and thesaurus. And then the targeted text can be compared with the text contents in the corpus, so as to realize its classification through data analysis. After completion of classification, the classified text can be brought in to the corpus to expand the corpus. There is a gap between the automatic text classification technology in China and such kind in western developed countries. Therefore, in addition to the research and development of corresponding software, attention should be paid to the technology innovation, combining with related knowledge of Chinese linguistics [3].

2.2.2 Feature Selection

Feature selection and comparison is also needed in the text classification of network information. The features are mainly cover the following aspects: the first one is the feature of syntactic structure, such as the number of the supplementary clauses; The second one is lexical feature, such as the using frequency of emotional words, using frequency of scientific and technical vocabulary, the number of written language, and so on; the third one is the symbolic feature, including the occurrence number of the question mark, frequency of space character, etc.; the fourth one is the feature of non-characters information, such as the content revealed by images, formula, hyperlinks, and so on; the fifth one is the feature of format, including the use of reference documents, introduction, summary, etc.; The sixth one is the feature of analysis level, such as the average length of sentences, the frequency of the English vocabulary, the use of rhetorical devices, and so on; the seventh one is the feature of derivative clues, such as the ratio and relation between the six features mentioned above. Feature selection process is shown in the figure below.

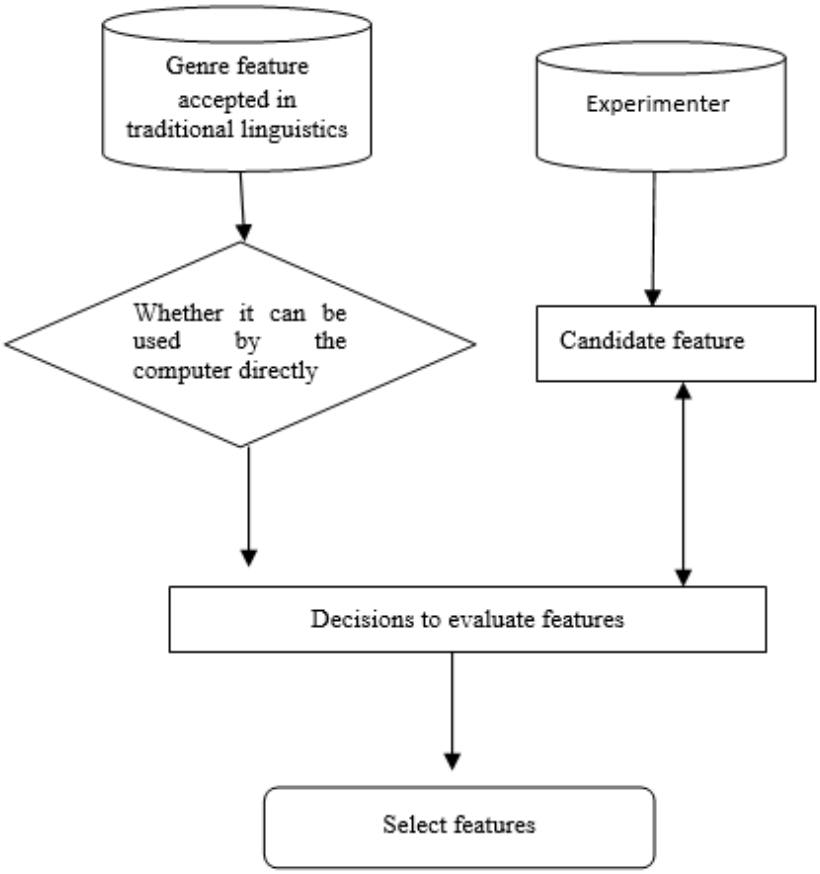

Figure 1. The flow chart of figure selection

By selecting these features to establish the comparison library, the computer system can achieve distinction automatically in the analysis [4]. When selecting features, the recognized feature of linguistics should be selected first; and then selection rules can be set artificially for the computer specific to the genre-related information which can be identified and used by the computer. On this basis, the computer can realize feature selection by automatic evaluation.

2.2.3 Operation of Classification System

The figure 2 shows the route chart of text genre classification

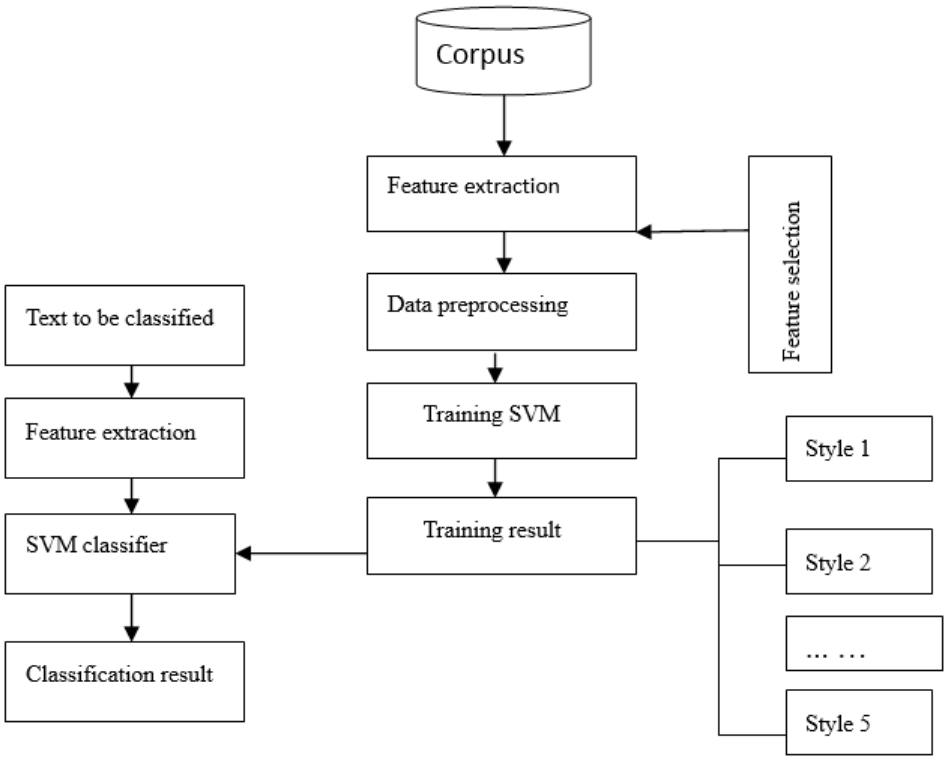

Figure 2. Route chart of the classification system

After selects the feature of text waiting for classification, the system gives it simple data preprocessing and inputs the data to the SVM classifier, and thereby receives the classification results. The core of this process is the SVM classifier. Through selecting features in the corpus, it makes SVM keep self-learning and supports manual intervention to correct the distinguishing way of information classification, and finally forms strict and scientific classification rules [5].

3. The Analysis Technology of Text Emotion

The emotion analysis technology is to analyze the emotion orientation of the text by making use of Internet to get inner emotion of the author and explore his/her emotional world. This technology can not only provide a reference for the evaluation of enterprise products, but also can help the government control the network of public opinion and clean up the network to promote the construction of a harmonious society. In addition, it can provide reference for the cultural industry. When using this technology, articles must be categorized (the categorization technology has been briefly described above) before their contents accept horizontal processing to determine the subjects that need to be analyzed. And then, the emotion of the article can be analyzed. Special attention should be given to the errors in analyzing, because a lot of words have different meanings in different environment and sometimes they even lead to opposite emotion direction. Therefore, the use of environment analysis is needed to the targeted text.

3.1 The Method Emotion Analysis

3.1.1 Data Preprocessing

First of all, data of the network text should be preprocessed, so that it can be converted into the language which can be recognized by the computer and some informal expressions can be standardized in this process. Moreover, the nature of each word in the article needs to be annotated and explained.

3.1.2 The Evaluation Object

The horizontal data investigation should be applied on the article which will accept emotion analysis to analyze the frequency of relevant words, so as to predict its possible emotional tone [6]. For example, in the search for “Obama”, the politics-related words show high frequency while

"economy", "diplomacy", "China" and other words appear frequently. These kinds of words should be attached importance in the emotion analysis. In addition, the form change of some words can also lead to the error of the computer analysis results, and thus it is necessary to apply problem detection on words. For example, the harmonic tone of some dialects needs to be transformed.

3.1.3 The Emotion Detection

Emotion adopts detection support vector machine (SVM). It can carry out self-learning to expand its store of emotion vocabulary and thus improves the accuracy of the emotion analysis. It is able to achieve the emotion distinction according to the characteristics of vocabulary. Table 3 shows the feature selection criteria of SVM.

Table 3. SVM feature criteria

Feature	Explanation
N-grams	Lexical feature, feature of binary word, feature of trigram word (unique)
Part-of-Speech	The part-of-speech tagging
Gazetteer	Vocabulary matching (comparing in the thesaurus)

3.2 The Establishment of Emotion Thesaurus

In the emotion analysis, the words in the article need to be matched with data in the thesaurus, so as to obtain the emotional orientation. Therefore, the emotion thesaurus is not only related to the comprehensiveness and accuracy of the emotion analysis, but also affects the efficiency of analysis. In addition to the same words, the matching relationship can also be used in the matching of vocabulary, such as coordinating relation, adversative relation, and so on.

The methods of thesaurus establishment can be divided into two kinds. The first one is to build thesaurus by artificial means. It makes use of a lot of man power to analyze and summarize emotion words, and then input them into the thesaurus after they are been classified. At present, this method can achieve PAD three-dimensional classification, evaluation of emotional effects and classification of self-related vocabulary. However, this method wastes much man power and money, which is not conducive to the effective promotion of emotion analysis. And sometimes, there will be a certain error due to the inconformity between lexical classification judged by human thinking and the computer language or classification system, which will destabilize the automatic classification system. And thus many researchers begin to use semi-automatic or full-automatic way to build the thesaurus -- this is the other method. This method enables the rapid updating of the emotion thesaurus and highly-efficient self-learning. By grading the objectifiability, positivity and negativity of a word, it obtains the vector of its feature, carries out vector computation and then classifies the word.

3.3 Psychological Research

The accuracy of emotion analysis presented by the technology which makes use of emotion thesaurus to carry out emotion match has yet to be improved [7] However, a large number of data analyses have proved that this technology is able to show the public's emotion changes at a macro level and thereby provides a reference for social psychology. In social psychology, for example, people with the same hobbies or the same characteristics always like to get together. In the analysis of the network text, the analysis of the people who embody the same hobby or the same characteristics shows that they also reveal some similar emotions in their articles. Thus, in the emotion analysis, information can be synthesized by correlating the relevant people to achieve an analysis more comprehensive. The implementation of this technology is difficult because it requires the support of general census. However, it can be combined with the credit system to achieve a full range of public information collection. Finally, some conclusions in social psychology can be used as a basis for the text emotion analysis in the information age. For example, psychology holds that the emotion of human beings reveals some regular changes at different time, such as the relatively negative emotion at noon. In the emotion analysis of an article, the corresponding analysis can be conducted according to the author' dispatch time.

4. Conclusion

In the information age, both the genre classification and emotion analysis of network text should be based on the establishment of a corpus and then carries out matching in the corpus. The matching of genre classification focuses on word frequency, sentence structure, article format and other text attributes. Emotion analysis attaches more importance on the matching of emotion words. Furthermore, it also analyzes the emotion orientation of the article by combining with the context and dispatch time of the network text, the related information of the author and various other factors. The genre classification of text based on classification technology can enhance the efficiency of government-related management. By accessing to the psychological feelings of commentators through analyzing the text, the Government can understand the emotional status of Chinese residents, which is beneficial for it to take targeted measures to promote the construction of a harmonious society. The author hopes that the research in this paper can help the contemporary relevant industries of information transmission and processing, so as to promote the construction and development of the harmonious society.

Acknowledgments

Science and Technology Project of Department of Science and Technology of Guizhou Province (China): “Research on Financial Text Genre Classification and Sentiment Analysis Based on the Web” (NO.QKHLKG [2013]45)

References

[1]. Fang Bing, Wu Jiang, Jin Yi. Research on the Genre Classification and Emotion Analysis Technology of Financial Text in the Information Age [J]. China Computer & Communication, 2014 (5).
[2]. Xu Jun. A study on the Technology of Genre Classification and Sentiment Analysis for Financial Information Retrieval [D]. Harbin Institute of Technology, 2011.
[3]. Zhou Wen. Emotion Tendency Analysis and Study of Chinese Network text based on EDT [D] .University Of South China, 2015.
[4]. Ye Xiangbin. The Study and Implementation of Network Text Sentiment Analysis [D]. Hunan University, 2015.
[5]. Fang Fengfei, Lin Hongfei, Yang Zhihao, et al. Automatic Classification Mechanism of Chinese Text Genre [J]. Journal of Chinese Information Processing, 2006, 20 (2): 24-32.
[6]. Ma Yuan. Research on the Emotion Analysis Technology of Short Text [D]. Chongqing university, 2011.
[7]. Chen Suqiong, Tao Juan. Overview of Text Classification Based on Depth Learning [J] .Research, 2015 (63): 72-72.