Analyse the data tendency in the public opinion monitoring system

Jian Li*

Luoyang Campus of the PLA Information Engineering University, Luoyang, China

*Corresponding author e-mail: maomaotnfntf@163.com

Abstract. With the rise of the network era of big data, many of a colour view of public opinion data emerge in endlessly, network also drives the various views in the different public opinion direction, so the tendency of public opinion data analysis is becoming more and more attention by many scholars and government officials, public opinion is most critical text content, data is mixed of text content. This paper analyses the tendency of network comment data, introduces language processing, subject extraction and other related knowledge, and achieves the purpose of improving the accuracy of tendency classification.

1. Introduction

With the rapid development of Internet technology, people are more inclined to express their opinions on the Internet, such as blogs, BBS, WeChat, live broadcast, twitter, etc., or spread events and participate in comments through news, current events, social channels and so on. So public opinion information has been widely spread on the Internet in various media, many media, business organizations, such as e-commerce platform will use public opinion to implement the regulation, promotion, marketing and other purposes, for example, when consumer is buying a product or service, will refer to the comment below has the user information, to achieve direct understanding of the product or service. Merchants collect these reviews, optimize customer products and after-sales services, and implement better production and marketing strategies. The national government analyzes public opinion data, understands social needs, provides better services for the people, analyzes information sources for emergencies, finds the most critical and popular problems, reasonably guides and solves them, and maintains social stability.

Data trend analysis has an important position in the field of public opinion monitoring, big data in the Internet environment, using the analysis of mining technology to judge subjective emotional color text, text data to mixed results, a process that involves the different level of word, word, sentence discourse data content, and produces different tendency analysis method, including the data mining technology, machine learning techniques [1], [2] language processing technology, information retrieval technology [3] and other related domain knowledge, and analysis of existing complexity. In addition, the application of these technologies in the field of trend analysis has expanded the scope of public opinion analysis system, new technical problems have been emerging, and numerous researchers have been solving new problems, which has also made the subject of trend analysis increasingly perfect in the field of public opinion monitoring.

Nowadays at home and abroad, this paper study tendency analysis method, summarized based on machine learning method based on semantic judgment method and mode of different characteristics, focus on the solution produced by different population level view differences of misjudgment effects on the fact that, in accordance with point of view of people's difference and the formation process of the
theme idea, adopting reasonable subject building model and feature selection method to extraction point topic. In view of the massive network public opinion data, a reasonable solution is found. By building a combinatorial classifier model on the distributed platform, the trend analysis results of public opinion data can be obtained quickly and accurately. The research in this paper can reduce the impact of subjective crowd thinking differences on the accuracy of tendency classification, and achieve a rapid and efficient analysis of public opinion tendency.

2. Data tendency classification method

2.1. Feature selection technology

Feature selection is an important part of data tendency classification. The feature item should fully reflect the information carried by the text data. Therefore, the more reasonable the feature item is selected, the more accurate the classification result will be.

2.1.1 Word frequency method, document frequency method

The word frequency method (TF) counts the frequency of characteristic words appearing in the text, and involves the technical content of some statistics. The size of the threshold is used to determine the topic words, and the statistical results and the range of the threshold are compared, which will exceed the threshold interval. The vocabulary outside the scope is eliminated, which not only extracts the high-frequency vocabulary, but also reduces the text dimension, which provides convenience for the subsequent classification by the classifier.

The document frequency method (DF) is developed based on the word frequency method. First, a feature item is defined. The feature item is placed in a specified database corpus, and the number of the feature items in the data set is counted and calculated. It accounts for the ratio of the total dataset corpus. The formula is as follows [4]

\[
DF(t_i) = \frac{N_{ti}}{N_{all}}
\]

In formula (1), \(N_{ti}\) is a feature item, \(N_{all}\) is the number of documents contained in feature item i, and \(N_{all}\) is the number of documents in all data sets. The ratio of the feature items in the corpus is obtained by the method, so that the theme is extracted for the feature item, that is, the word frequency method is used again. Use the previously specified threshold range to filter the topic, exclude words that do not meet the range of the threshold, or use the size of the DF value to judge. If the value is small, the frequency of a document appears to be small, then the document matches. The feature item does not adequately reflect the subject matter of the document. Conversely, when the value is large, it indicates that the document appears more frequently, and the matching feature item has clear content recognition ability. The center of the document word frequency method is reflected in the vocabulary labeling, matching documents, and judging the devaluation. It is an early widely used topic extraction method, but there are many uncertainties depending on the number of words.

2.1.2 Mutual information

Mutual information is used to analyze the relationship between two different random variables, that is, to measure whether the variables are related. It embodies the correlation between feature items and classifiers from the side, so that we can select the appropriate classification system based on this relationship. Different from the word frequency method, the feature item does not depend on the vocabulary, but is related to the tightness between the variables. The closer the relationship between the two random variables, the larger the feature item, the looser the association degree and the smaller the feature item. The essence is the degree of mutual correlation of feature items, and the expression of mutual information between category \(C\) and feature \(t\) is as follows

\[
MI(t_i, c_i) = \log \frac{p(t_i|c_i)}{p(t_i)p(c_i)}
\]

In formula (2), \(t_i\) is the feature item, \(c_i\) is the category to which the feature item belongs, \(MI\) is the relationship between a feature item and the belonging category, and \(p\) is the probability that the item
appears in the category. It can be seen that when M1 is 0, it indicates that the feature item has no relationship with the category, that is, two random events are independent of each other. If the value of M1 is larger, it indicates that the feature item is closely related to the category. By extracting topics from mutual information, the steps are cumbersome. It is necessary to label a certain category or a feature item first, so we must first use the classifier to do a classification operation, but this method directly categorizes the feature items and reduces the classifier. The influence between the topic extraction and the theme extraction is conducive to the subsequent classification model construction, and has a high degree of precision.

3. **Optimization Tendency Analysis Method**

3.1. **Feature Selection Technology Optimization Method**

When preprocessing the text data to be analyzed, if the vector feature is established directly by the word segmentation technique or the statistical word frequency to obtain the text feature, the text vector will have a huge dimension, the workload is very large, and the classifier classification is greatly reduced. Accuracy, so to ensure the core information of the text, simplify the workload, improve the classification accuracy, you need to adopt the most representative feature items, which is the focus of research feature selection technology. Common feature selection methods are Document Frequency Method (DF) and Information Gain Method (ICG). These methods all have their own selection criteria, but they also have their own shortcomings. This chapter analyzes the advantages and disadvantages of these feature selection methods, and improves the inadequacies. It adds word frequency parameters to reduce the interference factor and improves the accuracy of feature extraction.

3.1.1 **Document Frequency Method Improvement**

The document frequency method is simple to calculate. It only needs to consider the number of feature documents contained in the entire document set. It is necessary to calculate the weight value of each word in the data set and compare the set threshold value selection feature items. Feature items that are within the range of the threshold are retained, and feature items that are outside the range of the threshold are filtered out. This method is suitable for feature screening of large-scale document sets. However, the disadvantage is that there are many characteristic words with few occurrences, but these words contain important classification information. If it is directly filtered, it will have a great influence on the classification accuracy.

3.1.2 **Improvement of Information Gain Method**

The information gain is used to measure the difference between the feature item and the classification information. If a feature item can make the amount of information generated by the classifier larger, the more important the feature item is. Suppose there is a variable Y, and its value has n possibilities, namely $Y_1, \ldots, Y_N$ and each value probability is $p_1, \ldots, p_n$, then the information expression of the variable Y is:

$$H(Y) = -\sum_{i=1}^{n} p_i \times \log_2 p_i$$ (3)

It can be seen from formula (3) that the richer the number and type of variables, the larger the amount of information, indicating that the amount of information is related to the number of values and the probability of each number, and the value of the variable itself. The value does not matter much.

3.2. **Subjective Extraction Technology Optimization Method**

Among the various news and commentary topics on the Internet, the topic has a wide range of topics, which may be aimed at different social events, government decisions, facts, or opinions. Some comments contain repetitive and diverse views, i.e., subject matter. Fuzzy, contradictory views, which will affect the screening of features, and finally cause the problem of reduced classification accuracy. Therefore, an excellent topic recognition method can extract information from a wide variety of topics, and achieve the most attention-grabbing content in mining online public opinion. Topic recognition is
divided into two aspects, perspective topic recognition and text topic recognition. There are connections between the two and there are big differences. The subject of opinion refers to the object of subjective emotion published in the text content. The text topic refers to the object expressed by the text content, that is, by whom to express and what to express. In a variety of online news or comments, most researchers only analyze text topics, and ignore the topic, which leads to overlap between the two, affecting the accuracy of text topic extraction. Therefore, for both text and perspective, it is necessary to find a reasonable topic recognition technology to improve the accuracy of topic recognition.

4. Based on the concept of machine learning to build a public opinion classification model

4.1. Target requirements
The majority of netizens use various online platforms to comment on news reports or social dynamics in the society, express their own opinions, and form word of mouth or public opinion. The social public opinion monitoring department and the majority of social survey scholars have obtained different emotional tendencies, derogatory or derogatory meanings of the masses through different methods of inclination analysis. Through different degrees of analysis, we can measure the weight of public opinion and judge the possible impact of public opinion events. However, the perspectives of different social groups are different, and the opinions expressed are different. They may violate the original truth of the facts and form a wrong thinking guide. Therefore, public opinion will have deviations in the process of communication, leading to the accuracy of the tendency to discriminate. Sexual decline, so in order to maintain a high degree of accuracy in the final analysis, it is necessary to consider how to use an efficient method to solve the impact of crowd differences.

4.2. Subjective thinking differences
The difference in subjective thinking is manifested in two aspects: the complexity of Chinese expression and the diversity of people in different fields.

(1) Part of speech ambiguity: Different emotional words have different meanings under different parts of speech, and thus the tendency of emotions is different, such as "you are a person who is easy to be proud" and "I am proud of him", in two Under the word, "pride" has different tendencies, adjectives mean derogatory, and nouns mean derogatory.

(2) Collocation ambiguity: In the process of text expression, some words have no emotional tendency, but when used in conjunction with other words, they have emotional tendencies, such as "high quality" and "high price". The former means derogatory, The latter represents derogatory. "Learning prominent" and "lumbar prominence", the former derogatory and the latter derogatory.

(3) Statement Ambiguity: The same emotional words show different emotional polarities under different contexts. For example, some textual content will use the derogatory words: “True is OK” and “You are extremely excellent”. "Eye" and other expressions convey a concept of irony and accusation.

(4) Ambiguity of the evaluation object: Generally, the content of the news report will describe and comment on the object of an evaluation in different language expressions. However, due to the complexity of the Chinese context, the evaluation object needs us to make a clear judgment. Inaccurate, it may have the opposite emotional tendency, such as "a good cook made a bad meal", "a rotten chef made a delicious dish", evaluation chefs or evaluation dishes, we need to determine the object of evaluation, otherwise it will Misleading emotional tendency results. It can be seen that the semantic complexity of Chinese is restricted by part of speech, collocation, text environment, etc. Therefore, it is necessary to do a lot of optimization processing when judging the tendency of text, helping us to eliminate the influence of these potential factors and make the structure of sentiment analysis more To be accurate.

4.3. Solution
(1) The tendency analysis of social public opinion topics is a large-scale data set analysis. In order to take into account the emotional differences caused by different fields and different groups of people, a
single classifier cannot meet the demand. Therefore, it is necessary to build a framework and adopt the pattern of multiple classifications requires reasonable selection of text extraction methods and classifiers to ensure that different classifiers can be combined reasonably to obtain accurate data trend analysis results.

(2) Extract the theme first, and then tend to classify it as the main idea. It aims to resolve the differences in thinking between different groups of people and to weaken the interaction between different thinking in communication. After extracting the theme, it is necessary to construct a classifier, fully consider the classification method, the classification object, the association and influence of the classification steps, realize the distinction between the population, and reduce the influence of subjective thinking on the objective lyric facts.

Finally, a large amount of data is processed, and the results of the classification are obtained, and the accuracy of the results is verified.

(3) Because the concept of propensity analysis of data content is closely related to the technology of data mining and machine learning related fields [4], the related concepts of data mining and machine learning are introduced, and reasonable mining methods and learning modes are analyzed and selected. And on the basis of this, improve the construction of the classifier, and provide reliable and efficient text content support for our subsequent text data trend analysis operation.

4.4. Machine learning and data mining concepts

Data mining refers to searching in a large way in a large number of data. This process of searching is divided into automatic or semi-automatic [5]. For relatively large-scale data groups, automatic or manual supervision can be used for mining, but artificially assisted. The semi-automatic method is extremely labor intensive, so it is necessary to implement the concept of automated mining by means of machine learning, and to find hidden or unknown potential information in a large amount of data. Machine learning refers to the use or design of some analysis algorithms, according to specific rules, allowing the computer to automatically complete the analysis of the text content, and to achieve regular judgment of the content of the unknown text, and obtain the prediction results. Machine learning is the core part of data mining. It is mainly divided into supervised and unsupervised learning [6], as shown in Figure 1:

![Figure 1: Data mining and machine learning flow chart](image)

Figure 1 shows the implementation process of machine learning and related algorithms, which are mainly divided into supervised learning and unsupervised learning:

1) Supervised learning process

Supervised learning is also called predictive learning. In the case of text-oriented classification, it uses the classified text data set as a training sample, the predicted text data as a test sample, the design learning algorithm to achieve classification, and the result and sample set. The content is compared and the algorithm is optimized. Including classification and prediction algorithms.

2) Unsupervised learning process

Unsupervised learning refers to the analysis of text data directly through the algorithm under the condition of the determined sample data set, realizing the prediction of unknown text data, mining valuable information, no comparison of training sets, and completely based on the prediction set itself. The algorithm includes two modes: cluster analysis and association analysis. The clustering algorithm
implements the subdivision of samples, finds the similarities of sample features and classifies them into one class. Association analysis finds the connection between things and is used to optimize the analysis results.

It can be seen that the unsupervised learning method mainly relies on data mining to achieve, but there is no clear measurement standard for the specific algorithm optimization concept, that is, it only realizes mining, so this paper uses supervised learning to measure and update the algorithm.

5. public opinion analysis experiment

5.1. Environment Construction
The experimental background uses the eclipse development environment, the graphics editor Origin, to complete the text data preprocessing work. Feature extraction and classifier construction are implemented using the weka platform to verify the feature selection optimization effect. The combined classifier was constructed using the Hadoop distributed environment [7] to verify the accuracy of the combined classifier.

5.2. Evaluation indicators
The classification evaluation index in this paper is also an important reference for measuring feature selection and theme refinement, including: precision rate, recall rate, and F value as evaluation indicators:

(1) Precision and recall rate:

- Precision: Precision is the ratio of the number of texts that are to be classified to \( N_{w\rightarrow r} \), and the number of texts with correct classification results from \( N_{r\rightarrow w} \), indicating whether the classification result is accurate, such as Formula (4):

  \[ p = \frac{N_{r\rightarrow w}}{N_{r\rightarrow w} + N_{w\rightarrow r}} \]  

  \( (4) \)

- Recall: The full rate refers to \( N_{r\rightarrow w} \) in all texts belonging to a category. This class is correctly judged. The proportion of the number from \( N_{r\rightarrow w} \) reflects whether the classification result is complete, as shown in formula (5):

  \[ R = \frac{N_{r\rightarrow w}}{N_{r\rightarrow w} + N_{r\rightarrow w}} \]  

  \( (5) \)

(2) F value:

- F is a comprehensive evaluation of the recall rate and precision rate:

  \[ F = \frac{2 \times P \times R}{P + R} \]  

  \( (6) \)

However, the precision ratio P and the recall rate R are inverse relations. The change of the precision and the recall rate show a reverse trend. It is necessary to use the F value to balance the direct relationship between the two, but in the complex network public opinion environment, accurate It is very difficult to calculate the recall rate. Therefore, under the premise of introducing the F value, this paper mainly evaluates the precision.

5.3. Feature Optimization Effect
According to the weka visualization interface, we can calculate the trend of the curve after optimization by different feature selection algorithms, and analyze the optimization effect through the graph.

(1) The TF optimization effect map reflects the curve trend of the F value as the data volume changes, as shown in Figure 2:
It can be seen from Fig. 2 that the extraction effect of DF and DF/TF is analyzed, and the result curve is reflected by F value: DF/TF shows better accuracy as the number of feature items increases, and the F value gradually increases. High, indicating that the number of feature items does not affect the optimization method of adding TF, but the optimization effect is not obvious, and there is no optimization effect when the value is 4000 items.

(2) The IF optimization effect map reflects the trend of the F value as a function of the amount of data, as shown in Figure 3:

It can be seen from Fig. 3-7 that the extraction effect of IG and IF is analyzed, and the curve result is expressed by F value: it can be seen that IF has more obvious effect, but with the increase of characteristic phase, F value shows a downward trend. It shows that IG is not affected by TF, and feature items are more dependent on classification results. When the feature item reaches about 4500, the effect is best.

3) The MI optimization effect reflects the trend of the F value as the amount of data changes, as shown in Figure 4:
It can be seen from Fig. 4 that the mutual information M optimization effect is the most obvious among the three. When the TF is not considered, the F value gradually increases with the number of feature items, and when the TF factor is added, the F value is not affected by the feature. The number of items is affected.

6. Summary
In the context of massive public opinion data, this paper takes the tendency classification of the paper data as the main research content, and aims to find a text tendency classification method with high precision and fast recognition speed. Achieve accurate trend classification results for Internet news events or product reviews. This paper mainly considers the influence of subjective emotional color on classification accuracy, improves feature extraction algorithm, introduces topic recognition mode to optimize classification effect, and combines machine learning features to construct combined classifier to realize text data classification.

   (1) The feature selection algorithm is optimized in the feature selection stage. The TF word frequency correction parameters are added to the DF, MI, and IG feature extraction algorithms, aiming to optimize the classifier performance through a better feature selection algorithm.

   (2) proposes a solution to the subjective crowd's emotional color. Through the topic perspective extraction technology, and using the LDA topic extraction model and the derivative terminology, the classifier performance is optimized based on the feature extraction.

   (3) Based on the machine learning concept, this paper uses the integrated learning method to construct a combined classifier, and combines the clustering algorithm with the classification algorithm. Based on the optimal topic results, the classifier is used to derive the propensity results of different texts.

   From the final point of view of the optimization tendency classification results, this paper adopts a reasonable optimization method for each step of the inclination analysis, which not only effectively reduces the influence of emotional deviations at different levels of the population, but also makes the text data tend to analyze.

   The research on public opinion analysis and emotional judgment can't stay at the research stage, but it should be applied to many actual social statistics, such as government, enterprises, social organizations, etc., so that the statistical information is more accurate and practical. The reference value, although the public opinion monitoring can not reach the high level of foreign countries, but it must be in line with China's national conditions, and take advantage of the advantages of foreign monitoring methods, and steadily advance in the social environment.

References
[1] Dev S, Wen B, Lee Y H, et al. A Tutorial on Machine-Learning Techniques and Applications[J]. IEEE Geoscience&Remote Sensing Magazine, 2016, 4(2):79-93.

[2] Cui-Xia LI. Modern Computer Intelligent Recognition Technology of Natural Language Processing Research Progress and Application[J]. Science Technology and Engineering, 2012,12(7):51-53.

[3] Hammett I M. Study on the Cross-Language Information Retrieval Technology and Application[J]. Journal of Information, 2007, 34(1):85-110.

[4] Yang K F, Zhang Y K, Yan L I. Feature Selection Method Based on Document Frequency[J]. Computer Engineering, 2010, 36(17):33-35.

[5] Zhang W M, Jiang W U, Yuan X J. K-means teat clustering algorithm based on density and nearestneighbor[J]. Journal of Computer Applications, 2010, 30(7):1933-1935.

[6] Wang F, Lei B. Modeling and Analysis of Hadoop Distributed File System[J]. Tele communications Science, 2010:18(3),32-39.

[7] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves [C].International Conference on Machine Learning. ACM, 2006:233-240.