Enchanting Perspectives: Revitalizing Sentiment Analysis Approaches for Textual Resonance

[1]D Roja Ramani, [2]Faraz Shahjahan, [3]Raja Srujana, [4]Varun R Kolur
[1]Associate Professor, [2,3,4]Student
Department of Computer Science and Engineering,
New Horizon College of Engineering, Bangalore, India

Abstract: This paper reviews the application of natural language processing in sentiment analysis. Sentiment analysis is an important task aimed at automatically identifying and inferring sentiment tendencies and sentiment intensity in texts. This paper first introduces the application areas of sentiment analysis, including practical applications of text sentiment analysis. Then, text pre-processing techniques such as word separation, deactivation removal and punctuation processing are discussed. Then, feature extraction and representation methods are explored, including bag-of-words model, TF-IDF, word embedding and Word2Vec, attention mechanism and Transformer. In addition, methods for sentiment analysis, such as sentiment dictionaries and rule-based methods, traditional machine learning methods, and deep learning-based methods, are presented. Finally, the application areas of sentiment analysis are discussed and conclusions are given. The review in this paper will help readers understand the current status and development trend of natural language processing applications in sentiment analysis, as well as the advantages and disadvantages of different methods in sentiment analysis.

Keywords: Natural Language Processing, Sentiment Analysis, Machine Learning, Deep Learning.

I. Introduction

Natural Language Processing (NLP) is an important research direction in computer science and artificial intelligence, aiming to enable computers to understand, process, and generate human language. Text sentiment analysis is an important task in NLP, which automatically identifies and infers the sentiment tendency and intensity in text by extracting information from user opinions to obtain key sentiment information[1]. With the increase of text data such as social media and online comments, sentiment analysis is crucial to understanding user needs, market dynamics and changes in public opinion[2]

II. Applications in Text Sentiment Analysis

2.1. Text Pre-Processing Technology

Before text sentiment analysis, text pre-processing is a key step that involves cleaning, normalizing, and transforming the raw text data for subsequent feature extraction and analysis[3].

The specific text preprocessing process is shown in Figure 1

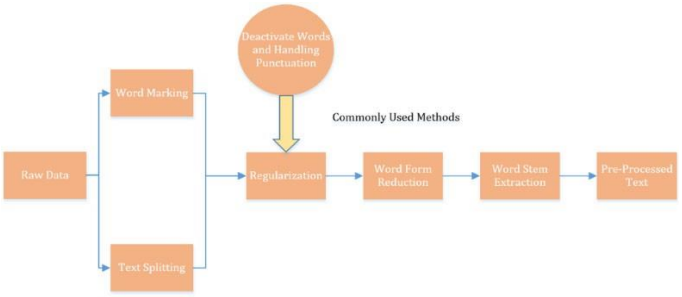

Figure 1. Data Pre-processing Flow Chart

2.1.1. Participle and Word Marking

Phrase splitting, the process of segmenting text into words, is crucial for analyzing sentiment tendencies in individual words[4]. Common techniques for word separation include lexical, rule-based, and statistical methods[5]. For Chinese text, tools like jibe and THULAC are commonly used, while English text can be split by spaces. Lexical annotation assigns linguistic properties like nouns and verbs to words, aiding in the identification of emotionally charged terms. Tools such as Hidden Markov Model, Maximum Entropy Model, and Conditional Random Field are employed for lexical annotation[8][9][10][11][12][13][14].

2.1.2. Deactivation and Punctuation Handling

Deactivation and punctuation handling are fundamental steps in the preprocessing phase of sentiment analysis, aimed at improving the accuracy and efficiency of sentiment analysis models.

- Deactivation:

Deactivation, also known as stop-word removal, involves the identification and removal of common, non-informative words from text data. These words, such as "of," "is," "the," and common pronouns, prepositions, and conjunctions, have little significance in sentiment analysis and can introduce noise into the analysis process. Their presence can increase the complexity of text processing and potentially interfere with the results of sentiment analysis. Removing these deactivated words is essential because it reduces noise and eliminates redundant information, ultimately enhancing the performance and effectiveness of sentiment analysis models. Deactivated word lists can be obtained from open-source libraries or custom vocabularies, and removal can be accomplished using simple rule-based methods or by matching against a list of deactivated words.

- 2. Punctuation Handling:

Punctuation handling in sentiment analysis involves the appropriate management of punctuation marks, including periods, commas, exclamation points, and question marks.

Punctuation marks can convey important information about the sentiment and emotional tone of a text. For example, exclamation points often indicate excitement or strong emotions, while question marks can suggest uncertainty. However, in some cases, excessive or inconsistent use of punctuation can hinder sentiment analysis algorithms. Therefore, it is essential to strike a balance between preserving the useful information conveyed by punctuation and avoiding its potential interference with the sentiment analysis process. This may involve removing certain types of punctuation or standardizing their usage to ensure consistent analysis results.

deactivation and punctuation handling are crucial preprocessing steps in sentiment analysis. Deactivation helps eliminate common, uninformative words that can add noise to the analysis, while punctuation handling ensures that punctuation marks are appropriately managed to preserve their meaningful contributions to sentiment while minimizing interference. These techniques collectively contribute to more accurate and reliable sentiment analysis results.

2.1.3. Morphological reduction and stemming

Word form reduction and stem extraction in sentiment analysis serve to reduce words to their original forms or to extract their stems in order to reduce variations in the surface forms of words and to capture core meaning[21]. Morphological reduction reduces words to their basic form by considering information such as root, lexical nature and context. Stem extraction, on the other hand, extracts word stems and reduces the dimensionality of the feature space. This better captures the true meaning of words and reduces redundancy. Commonly used techniques include dictionary-based and rule-based approaches[22]. Word form reduction and stemming extraction provide more accurate and consistent feature representation, eliminate feature redundancy and noise, and improve model accuracy and ability to capture sentiment tendencies. However, appropriate tools need to be selected, tuned, and optimized according to the task and dataset

2.2. Feature Extraction and Representation Methods

2.2.1. Feature extraction and representation methods are crucial in sentiment analysis,

It transforming raw text into computer-processable numerical features [23]. Bag-of-words model and TF-IDF are commonly used text feature extraction and representation methods for capturing the frequency and importance of words and applying them to sentiment analysis tasks[24]. The bag-of-words model treats text as a collection of words, ignoring order and grammatical structure and focusing only on the frequency of word occurrences[25]. The text is converted into a numerical representation by constructing a document-word matrix. The bag-of-words model assesses the importance of words by assuming that words with higher frequencies are given more weight in the representation. TF-IDF (word frequency-inverse document frequency) is a method that integrates word frequency and document frequency to evaluate the importance of words in text[26]. TF-IDF filters out common words and highlights the important words in a given document. Its formula is defined as follows:

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

$$TF - IDF = f_{ij} \times idf_i$$

The advantages of the bag-of-words model and TF-IDF are their simplicity and ease of implementation, and their ability to effectively represent lexical information of text. They are suitable for sentiment analysis, revealing users' opinions and emotional tendencies, capturing the frequency and importance of keywords. However, they also have limitations, such as the inability to capture word order and contextual information, and ignoring semantic relationships

2.2.2. Word embedding and Word2Vec

Word embeddings and Word2Vec play an important role in sentiment analysis. Word embedding converts text into a dense vector representation that better captures the semantic information of words[27]. This continuous vector representation provides better feature representation and semantic similarity computation, which helps to identify sentiment tendencies and expressions. Using the pre-trained Word2Vec model can extract the contextual relationships of words and improve the performance of sentiment analysis models[28]. Word embedding maps words to a continuous vector space and provides rich feature representation by learning to capture the semantic relationships of words, making semantically similar words closer together in the vector space. It uses language models or neural networks to learn word embedding models on large-scale text data, making similarly meaningful words closer together in vector space. Word2Vec is a commonly used word embedding algorithm based on a shallow neural network model trained on largescale text data to generate high-quality word vectors by predicting contextual information around words.

It consists of a continuous bag-of-words model (CBOW) and a Skip-gram model, both of which generate word embedding vectors that capture the semantic features of words. Its network model diagram is as follows:

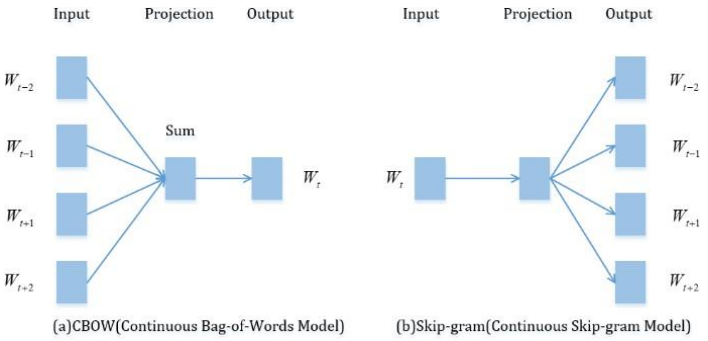

Figure 2. CBOW and Skip-gram Models

Word embeddings better capture the semantic information of words and usually perform well in sentiment analysis. word2Vec is an important feature extraction and representation method that captures semantic relations and

contextual information of words and provides rich feature representation, thus improving the accuracy and effectiveness of sentiment analysis

2.2.3. Attention mechanism and Transformer

Feature extraction and representation methods are pivotal in sentiment analysis. The Bag-of-Words (BoW) model and TF-IDF excel at capturing lexical information, while word embedding techniques like Word2Vec capture semantic relationships. However, the game-changer lies in the attention mechanism and the Transformer model. The attention mechanism mimics human focus, addressing issues in traditional networks, and the Transformer model, built entirely on attention, enhances sentiment analysis by capturing dependencies between global and local information. These methods combined improve sentiment analysis accuracy and effectiveness, enabling a deeper understanding of people's opinions and emotions in text data.

2.3. Emotional Analysis Method

2.3.1. Emotional lexicon and rules-based approach

The sentiment lexicon and rule-based approach is a commonly used sentiment classification technique. The method uses pre-constructed sentiment dictionaries and rules for sentiment classification, and the overall sentiment polarity is derived by matching and weighting the sentiment words in the text[31]. The sentiment lexicon contains vocabulary and corresponding sentiment tendencies, which can be constructed manually or obtained by automatic mining and machine learning. The process of sentiment lexicon analysis is as follows: word segmentation and lexical annotation segment the text into words or phrases and determine the lexical nature of each word. By matching the sentiment lexicon, the sentiment words present in the text are identified. The sentiment score of the text is calculated based on the number, position and sentiment polarity of the sentiment words. Common calculation methods include simple counting, weighted counting and rule-based matching [32].

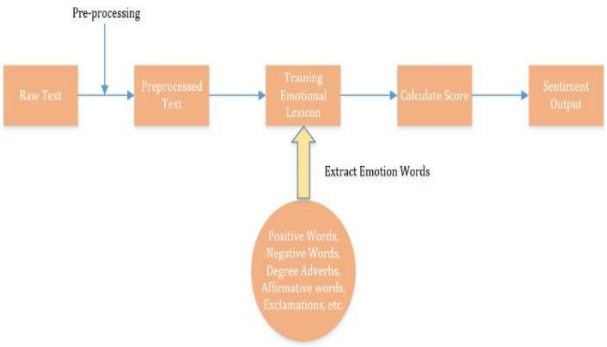

Figure 3: Flow Chart of Sentiment Analysis Method Based on Sentiment Dictionary

Rules can also be used to handle negation, degree adverbs, and context dependencies. Negation words can change the sentiment polarity of subsequent sentiment words, and degree adverbs can adjust the weights of sentiment words to reflect sentiment intensity. Rule matching and sentiment adjustment can be performed according to the context dependency of sentiment expressions. English sentiment dictionaries are more maturely developed, and the common ones are SentiWordNet and General Inquirer Opinion Lexicon[33]. Chinese sentiment dictionaries include HowNet, a Zhiwang dictionary, NTUSD of National Taiwan University and Dalian University of Technology's Chinese Sentiment Vocabulary Ontology Library[1]. The method based on sentiment lexicon and rules has the advantages of simplicity, intuitiveness and interpretability. It does not require a large amount of training data and is computationally efficient. However, the method also has limitations. The quality and coverage of sentiment dictionaries directly affect the classification accuracy, and the lack of domain-specific or new vocabulary sentiment information may lead to classification errors[34]. In addition, the rule-based approach has difficulty capturing complex semantic and contextual relationships.

2.3.2. Traditional machine learning based approach

In sentiment analysis, traditional machine learning-based approaches use labeled sentiment datasets for model training and prediction. The approach converts text features into numerical features and performs sentiment classification using classification algorithms, such as plain Bayes[35], support vector machines[36], decision trees[37] and random forests[38]. Feature selection and model tuning are key steps, where important features can be selected by information gain and mutual information, and parameters can be tuned using cross-validation, grid search, etc. Sentiment analysis methods based on traditional machine learning have the advantages of being explanatory and suitable for small-scale datasets. They can provide explanatory relationships between features and classification results, and can incorporate domain knowledge for feature engineering. However, the performance of traditional machine learning methods may be limited when dealing with large-scale datasets and complex contexts. In addition, feature engineering requires manual involvement and high requirements for domain knowledge.

2.3.3. Deep learning based sentiment analysis method

Deep learning-based sentiment analysis methods use multilayer neural networks to learn abstract feature representations. Trained with large-scale datasets, these methods are able to automatically extract key semantic features of text and implement sentiment classification tasks. At the core of deep learning methods are neural network models, as shown in Figure 4. In sentiment analysis, commonly used neural network models include convolutional neural network (CNN)[39], recurrent neural network (RNN) and long and short term memory network (LSTM)[40], Transformer[30], etc.

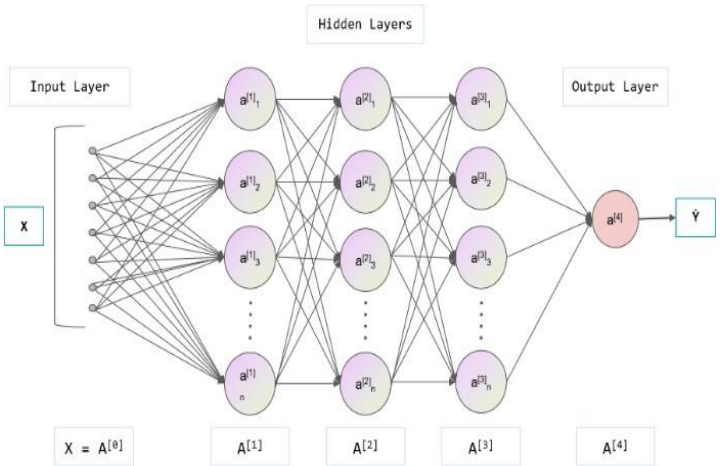

Figure 4. Neural Network Model

CNN (Convolutional Neural Network) suffers from the lack of sequence modeling capability in sentiment analysis 61 and has limited ability to model text sequences due to the limitation of local perceptual fields and weight sharing to capture long-range dependencies.

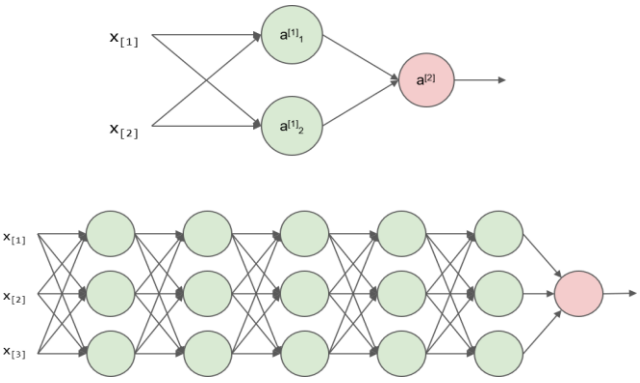

Figure 5. Convolutional Neural Networks

On the other hand, RNN (Recurrent Neural Network) is able to retain contextual information and model sequence dependencies but is prone to gradient disappearance and gradient explosion problems, and has limited memory capacity when dealing with long sequences, which affects the accuracy of sentiment classification[41].

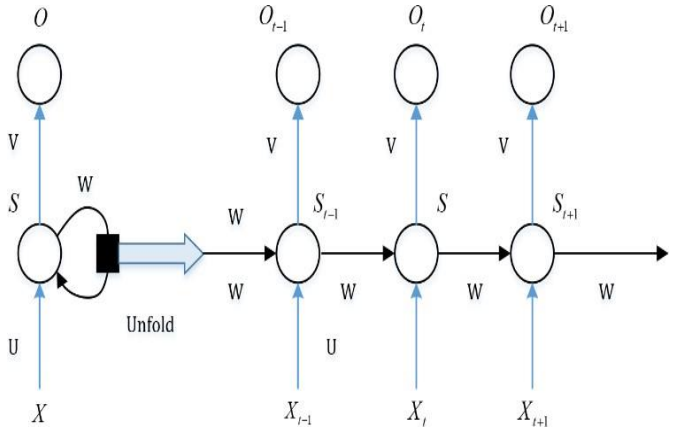

Figure 6. Recurrent Neural Networks

To overcome these problems, LSTM (long short-term memory network) introduces a gating mechanism and memory units to solve the gradient problem and long-term dependency problem and improve the performance of sentiment analysis.

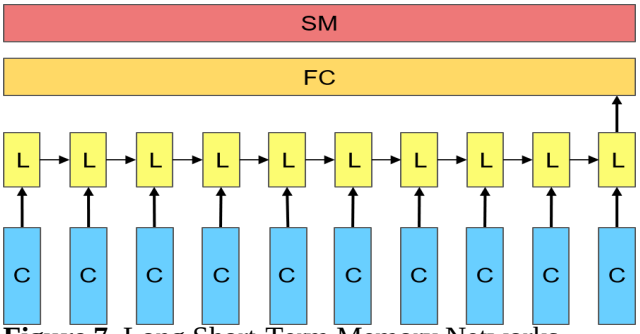

Figure 7. Long Short-Term Memory Networks

However, the computational complexity of LSTM is high and still has limitations for very long text sequences. When LSTM is used for text sequences of more than 200 words, it still has more serious long-term problems and generateshigher errors.

The transformer model is an emerging model after RNN and LTSM, which is widely used in text sentiment analysis. Based on the self-attention mechanism and the introduction of parallel computing capability, the processing and understanding of text sequences are more efficient, while the attention mechanism can solve the problem of long-distance dependence of sequences.

The essence of the attention mechanism is to selectively filter a small amount of important information from a large amount of information and focus on this important information, ignoring the unimportant information[42]. By weighting the weights at different positions in the input sequence, it enables the model to focus more on the important parts when processing the sequence data. Its structure diagram is shown in Figure

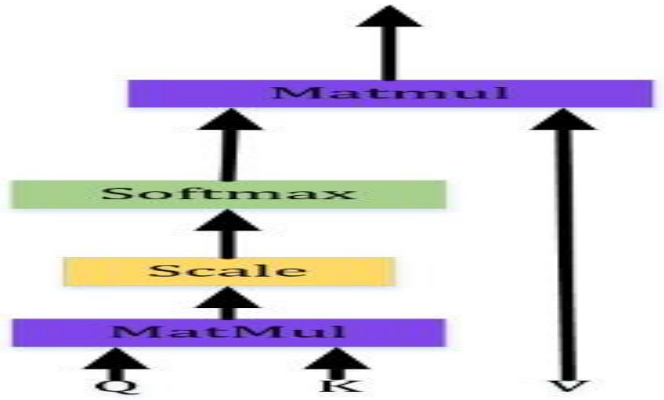

Figure 8. Attention Mechanism Module

In the Transformer model, the attention mechanism is applied to the Self-Attention mechanism (Self-Attention)[43]. The Self-Attention mechanism calculates the interactions of each position in the sequence with other positions and assignsa weight to each position based on the similarity. By weighting and summing the attention weights of each position, the contextual representation of each position and the degree of contribution can be obtained. The advantage of this attention mechanism is that it can fuse information from different positions in the sequence to establish global semantic associations and thus better understand the semantics and structure of the whole sequence.

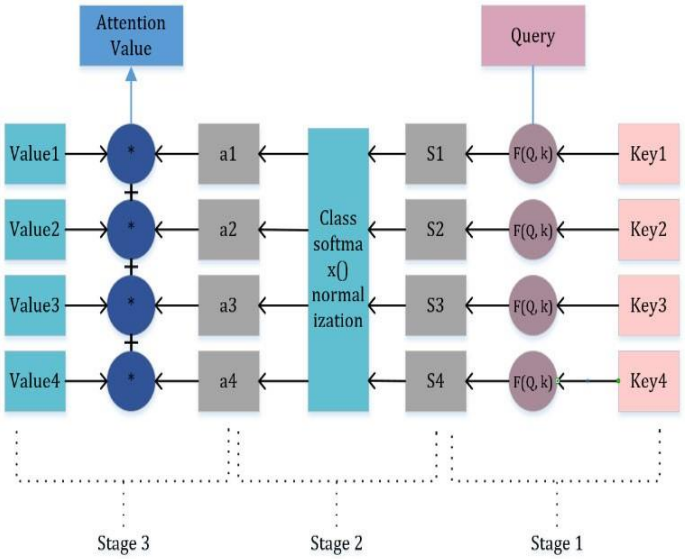

Figure 9. Self-Attention Mechanism Module

The difference between the self-attention mechanism and the attention mechanism lies in the fact that Q, K, and V of the self-attention mechanism are homologous, i.e., they all come from the same word vector.

Multi-head Self Attention[44] is an extended self-attentivemechanism model in a Transformer, i.e., the superposition of multiple multi-head self-attentive mechanisms constitutes theencoder-decoder structure in a Transformer. It slices the input into eight independent attention heads to compute the self-attention in parallel, each with different K, V, and Q, which is equivalent to eight linear transformations of the original self-attention, so that different representation spaces can be learned.

Transformer[30], proposed by Vaswani et al. is a neural network integrated with a multi-headed self-attentive mechanism. The model adopts an encoder-decoder structure,which solves problems such as the inability to compute in parallel with long-distance dependence in RNN and LTSM. In sentiment analysis tasks, the Transformer model possesses an extremely powerful context modeling capability[45]. It canencode and decode text sequences, capture semantic information between keywords and contexts, and effectively construct dependencies in sequences. The overlay computation by a multi-headed self-attentive mechanism enables the model to better understand the semantic and sentiment expressions in sentences and extract the features that contribute to sentiment classification. Meanwhile Transformer introduces operations such as residual module and batch normalization, which effectively avoid the problems of gradient disappearance and gradient explosion during network training while improving the training efficiency. Its schematic diagram is shown in Figure 10.

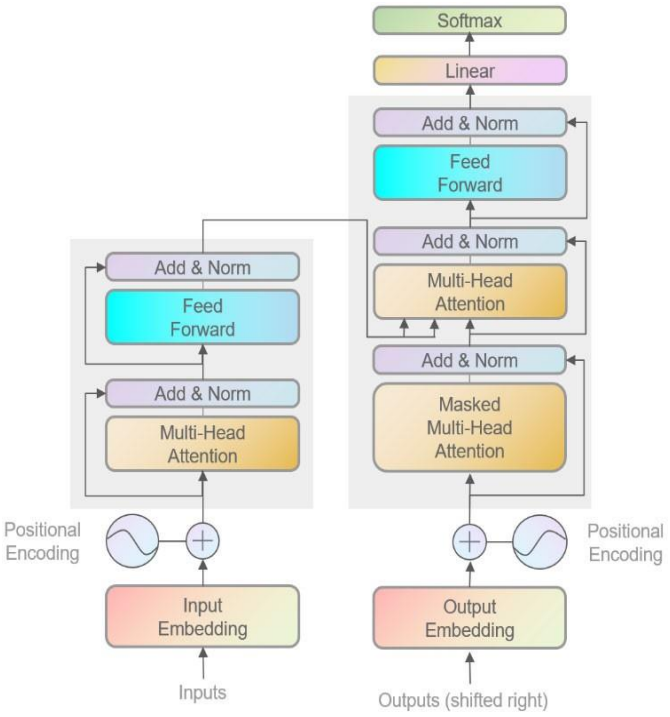

Figure 10. Transformer Module

However, deep learning-based sentiment analysis methods also have some challenges. First, deep learning methods have a high demand for data and require a large amount of labeled data for model training. Second, deep learning models tend to be more complex and require more computational resources for training and inference, and also lack interpretability to explain the prediction results of the models.

In summary, sentiment lexicon and rule-based methods are simple and intuitive, and are suitable for small-scale data and simple sentiment analysis tasks. Traditional machine learning methods have good explanations and interpretability and perform better on small data sets. Deep learning methods have achieved remarkable results on large-scale data and complex contexts by learning abstract semantic feature representations. Future research can explore the fusion of different methods to

improve the accuracy and robustness of sentiment classification and combine domain knowledge and contextual information to achieve more refined and personalized sentiment analysis applications.

III. Text Sentiment Analysis Application Areas

A large amount of user-generated text data on social media platforms contains rich sentiment information, and sentiment analysis can be used to analyze users' sentiment tendencies toward specific events, products, or topics, thus providing insights into users' emotional attitudes. For example, in microblog sentiment analysis, based on sentiment lexicon, Yuqing Li et al. proposed a bilingual lexicon sentiment analysis method based on KNN algorithm[46], andYanyan Zhao used the huge data volume of microblogs to build a mega sentiment lexicon[47], which has significantly improved the effect of sentiment classification on microblogs.

Sentiment analysis can also help companies understand consumers' emotional attitudes and emotional needs towards brands so that they can develop more targeted marketing strategies and brand management strategies. For example, Asghar[48] et al. produce sentiment dictionaries integrating emoticons and domain terms for hotel review data and obtainhotel review sentiment for improving hotel facilities to meet people's needs. Min Peng[49] et al. use LDA topic model to analyze potential attributes of goods and calculate users' emotional tendencies and user similarity to the attributes of goods, which enables merchants to better design more perfectgoods according to customers' emotional preferences. Jinmeng Li combines sentiment analysis algorithms with eye-tracking devices[50], where users wear eye-tracking devices to derive the area of gaze on the text, and finally the sentiment value of the user's area of interest is derived from the algorithmic model to determine the user's preference for the content .

IV. Conclusion

This review provides a comprehensive overview of the application of natural language processing in text sentiment analysis. Sentiment features can be effectively extracted from text by text pre-processing techniques and feature extraction methods. Different sentiment classification algorithms and sentiment intensity analysis methods can help to understand the sentiment tendency and intensity in texts. Moreover, multilingual sentiment analysis and case studies in different application domains demonstrate the wide application of sentiment analysis. However, there are still challenges such as data scarcity, model generalization capability and interpretability that need to be addressed. Future research can combine multimodal data and emerging technologies to improve the accuracy and efficiency of sentiment analysis and provide more valuable information for decision making in various industries.

References

1. Zhong Jiawa, Liu Wei, Wang Sili et al. Review of Methods and Applications of Text Sentiment Analysis[J]. Data Analysis and Knowledge Discovery,2021,5(06):1-13.
2. Wang YJ, Zhu JQ, Wang ZM et al. Review of Applications of natural language processing in text sentiment analysis [J/OL]. Computer Applications:1-12 [2023-07-03].
3. Zhao Jingsheng, Song Mengxue, Gao Xiang, et al. A study of text representation in natural language processing[J]. Journal of Software, 2021, 33(1): 102-128.
4. Li Yachao, Xiong Deyi, Zhang Min. A review of neural machine translation[J]. Journal of Computer Science, 2018, 41(12): 2734-2755.
5. Ao Sheng,Xu Lan,Ao Qingwen.Application of NLP Chinese word separation technology in bridge report data processing[J]. Traffic World,2020(17):3-5.DOI:10.16248/j.cnki.11-3723/u.2020.17.001.
6. Wang YY, Dong GW. Research on the problems and countermeasures of netroots libraries based on jieba subtext[J]. Jiangsu Science and Technology Information,2023,40(06):35-38.
7. Yin R, Wang Q, Li P, et al. Multi-granularity chinese word embedding[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 981- 986.
8. Zhao Pengfei,Zhao Chunjiang,Wu Huarui et al. BERT-based multi-feature fusion for named entity recognition in agriculture[J]. Journal of Agricultural Engineering, 2022, 38(03): 112-118.
9. Chen Zhiyan,Li Xiaojie,Zhu Shuhua et al. A two-way maximal matching word splitting method based on Hash structureddictionary[J]. Computer Science,2015,42(S2):49-54.
10. Li Ship. Statistics in Chinese word splitting[J]. China Statistics,2020(10):34-35.
11. Liu, X.Y.Y., Sheng, Y.H., Qin, J.R., et al. A semantic matchingmethod for spatio-temporal trajectories based on hidden Markov model[J]. Geography and Geographic Information Science,2023,39(03):1-6.
12. Liu, Yunzhong, Lin, Yaping, Chen, Zhiping. Hidden Markov model-based text information extraction[J]. Journal of SystemSimulation, 2004, 16(3): 507-510