International Journal of Information Technologies and Systems Approach
Volume 16 • Issue 2

Sentiment Distribution of Topic Discussion in Online English Learning:
An Approach Based on Clustering Algorithm and Improved CNN

Qiujuan Yang, Weinan Normal University, China*
Jiaxiao Zhang, Xidian University, China

ABSTRACT

Online English teaching resources have recently surged, highlighting the exigency for efficient organization and categorization. This manuscript introduces an innovative strategy to classify university-level English teaching resources, employing a sophisticated density clustering algorithm. Initially, student discourse was mined within a teaching platform comment section, and in-depth textual analysis was conducted. Subsequently, the term frequency-inverse document frequency (TF–IDF) feature extraction algorithm was enhanced, while emotive attributes were seamlessly integrated into the textual manifestation layer during the classification procedure. This enabled the distribution of topics and emotions to be acquired for each comment, facilitating subsequent analyses of emotion feature extraction and model training. An improved weight calculation was designed based on TF–IDF to evaluate the importance of feature items for each corpus file. The simulation results demonstrate the proposed scheme's effectiveness. The algorithm facilitates faster scholarly access to educational resource information and effectively classifies data for high research adaptability.

KEYWORDS

Density Clustering Algorithm, IF–IDF, Teaching Resource Classification, Text Mining, Word Frequency Analysis

INTRODUCTION

English is an essential component of China's basic education system. As international economic trade becomes more frequent, English is no longer simply a subject but a vital communication tool. Improving the effectiveness of English teaching in universities and enabling students to acquire English proficiency more efficiently is a topic of ongoing research. While traditional classroom teaching has the advantage of conveying knowledge quickly, it falls short in consolidating and reinforcing students' impressions, particularly regarding language learning, which benefits from exposure to a suitable language environment. The emergence of online English teaching, made possible by the rapid development of Internet technology, has revolutionized distance learning, breaking down time and space constraints and greatly enhancing the learning experience. As a result, online learning

DOI: 10.4018/IJITSA.325791 *Corresponding Author

has become a significant reform direction for university-level English teaching. Using the Internet, teachers can provide students with a wealth of diverse learning resources, encouraging their enthusiasm for learning. Furthermore, online learning platforms facilitate listening, reading, and conversation, improving the practicality of English and enhancing students' language application abilities.

As network technology advances and the information age emerges, the demand for online teaching resources has expanded significantly. The core purpose of educational resources is to provide relevant services for learners and maximize their utilization value through reasonable classification (Colangelo et al., 2018). Misclassifying teaching resources not only reduces their educational value but also results in the wastage of human and material resources. Hence, teachers and students in basic education urgently need scientifically organized and managed education and teaching resources. Moreover, the development of online English teaching has spurred the construction of educational resource platforms, which play a vital role in the education informatization process. Currently, most college and university teachers use online platforms to provide discussion areas where students can express their doubts and opinions. Teachers can dynamically adjust teaching content, plans, and focus based on content in their online platform's student comment section. However, although the relevant educational resource platforms have reached a certain level of development, many shortcomings remain regarding their application effects. The openness of teaching platforms leads to an uncontrollable level of user access, making it challenging for teachers to sort out and summarize comment content. In this context, analyzing teaching resources based on topic discussion and sentiment analysis can enable the fuller and more effective use of online teaching resources. This is important in promoting the comprehensive reform of basic education in China and enhancing its degree of informatization (Bustos et al., 2020; Newman & Joyner, 2018; Yadollahi et al., 2017).

Online pedagogical resources encompass text, audio-visual, and pictorial materials. As such, several conventional classification and management systems have been developed to organize these resources suitably. However, the current literature lacks a comprehensive analysis of the interrelatedness of these resources, which has led to low classification reliability. In particular, researchers have neglected to account for students' emotional tendencies during the learning process. This paper proposes a method for content extraction and emotional analysis based on student discussions in online English classes. The extracted content sequences were analyzed through contextual information recognition to classify teaching resources. Specifically, the content of student comments on an online English teaching platform was gathered, and the text of these comments was analyzed to improve the term frequency–inverse document frequency (TF–IDF) feature extraction algorithm. Density clustering algorithms were utilized to analyze the sentiment of the comments based on word frequency, fully considering the correlation between resources in neighboring grids. Weighted grids were constructed for each resource partition, and key-value indicators were set based on resource correlation. Additionally, a text expression layer, an attention mechanism layer, and a convolutional neural network (CNN) layer incorporating sentiment features were added to the classification process to achieve sentiment feature extraction. The model was then trained, and accurate classification of resource features was achieved by judging the range of key-value parameter values.

The organizational structure of this paper is as follows. The second section introduces the research status of text mining technology and sentiment analysis technology. Based on these two technologies, the third section introduces the research focus of this paper, proposing an improved weight calculation method based on TF–IDF and using the density clustering method based on a weighted network to classify online teaching resources. Finally, the fourth section discusses the experimental data processing and experimental results.

RELATED RESEARCH

Information resource classification aims to differentiate and arrange information resources based on their attributes or characteristics, creating a systematic and orderly classification system. Currently,

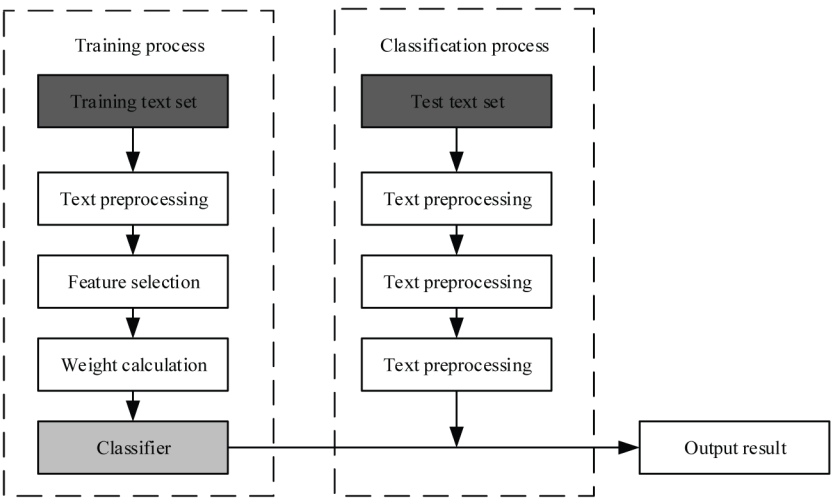
information resources are primarily categorized into traditional literature or paper resources and network information resources based on the attributes of the classified objects. The classification of teaching resources falls under network classification, in which web developers organize and integrate collected network resources using search methods for information queries. In this process, the selection of classification criteria is crucial. Figure 1 depicts a text classification flowchart comprising the two critical steps of the text classification process: training and testing. The most important step involves the training of text data or the mining of text for hidden information. At the same time as the training stage, the text needs to be classified; this step is based on emotion analysis. Thus, the critical aspects include text mining and sentiment analysis of student discussions on topics related to the teaching process.

Text Mining

As elucidated by Kumar et al. (2021), text mining constitutes a pivotal domain within the natural language processing field, encompassing the deconstruction of textual data to extract pragmatic, valuable, and substantive information dispersed within the text. This extracted content enhances information organization, rendering it more structured and discernible. The widespread proliferation of Internet usage has engendered an unprecedented influx of user-generated text disseminated across the digital landscape, catapulting text mining to the forefront of research endeavors in natural language processing. Scholarly investigations have delved into text mining across diverse domains, encompassing review data, policy texts, and literary compositions, thereby paving the way for applying text mining techniques to resource amalgamation, classification, and intelligent recommendation systems.

Scholars studying online comment text mining mainly focus on topic extraction and sentiment analysis. Extracting the main ideas from comment information is crucial for text content topic mining. To achieve this, scholars typically utilize cluster analysis techniques, such as k-means clustering (Syakur et al., 2018), hierarchical clustering (Murtagh & Contreras, 2012), and the latent Dirichlet distribution (LDA) topic model (Guleria & Sood, 2018). Text clustering is a research method that determines the subject content of a segment of text based on text similarity, with the content of texts on similar subject matters being more alike. For instance, Liu et al. (2018) deactivated words in the text, formed a feature word sample matrix using the TF–IDF (Nguyen et al., 2016) feature extraction

Figure 1. Text classification flow chart



method, and then clustered subject words using the bisecting k-means clustering algorithm (Zhang & Wang, 2020) to extract textual subject matter. In another study, Shafqat and Byun (2019) used an LDA topic model to analyze crowdfunding review topics, providing valuable insights for developing and promoting crowdfunding projects. Rashid et al. (2022) proposed a one-way clustering algorithm based on the LDA topic model to extract key feature information from text, enabling deeper mining of the information and exploration of hot topics in the text. Finally, Wu et al. (2020) proposed a biterm topic model to extract microblog opinion topics by calculating improved TF–IDF weight values to identify features of short microblog texts, effectively solving the high-dimensional sparse problem encountered in modeling short texts and achieving quality hot topic extraction. The proposed method makes up for the poor model construction in the case of sparse text words in the LDA topic model.

Sentiment Analysis

Sentiment analysis—or scrutinizing textual content for sentiment patterns—utilizes various tools, such as artificial intelligence (AI), computer vision (CV), and natural language processing (NLP). Therefore, sentiment analysis methods typically utilize specific sentiment lexicons (Lu & Wu, 2019), machine learning algorithms for classification (Webb et al., 2010), and deep learning algorithms. For example, He (2022) presented a sentiment analysis approach that involves designing a sentiment score calculation based on WordNet and SentiWordNet sentiment lexicons validated on the BBC News domain dataset. However, this approach fails to consider domain-specific sentiment words and the polysemy of sentiment words, leading to biased analysis results. To address this issue, Xu et al. (2019) proposed a Chinese text sentiment analysis method based on a comprehensive sentiment dictionary that incorporates particular sentiment words and applies special treatment to polysemy sentiment words. Dias et al. (2020) used N-grams and introduced triads to develop machine learning–based text sentiment analysis schemes. This approach fuses text features from diverse domains as internal knowledge features while also discussing the impact of three weighting methods (term frequency (TF), TF–IDF, and binary) on support vector machine (SVM) classifiers. The sentiment score of a comment is then computed using a sentiment analysis package to create its sentiment score vector as an external knowledge feature. Additionally, Pan et al. (2021) proposed a bidirectional CNN–RNN algorithm model based on an attention mechanism to extract forward and backward feature information from the text; this information is then weighted using the attention mechanism to achieve sentiment analysis.

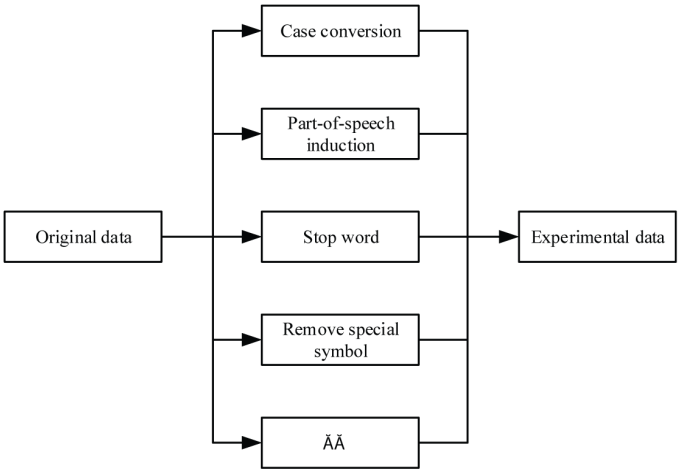
In their seminal work, Geraldi and Ghisi (2022) proposed the utilization of combinatorial neural network algorithms for clustering analysis of product attributes, complemented by the use of word2vec to compute semantic similarity, thereby constructing a sentiment lexicon based on product review data. On the other hand, Hassan et al. (2022) introduced a fusion model that integrates convolutional neural networks and bidirectional long- and short-term memory networks to effectively address the limitation of single convolutional neural networks in capturing contextual word meanings holistically. Conversely, Séin-Echaluc et al. (2015) leveraged a hybrid architecture consisting of a CNN and a recurrent neural network. Here, the CNN serves as the word vector layer, while the bidirectional long short-term memory (BiLSTM) network successfully mitigates the challenge of long-range word dependencies. However, this approach overlooks the crucial TF–IDF factor, thereby resulting in an incomplete analysis of the information pertaining to word frequencies.

Classification of Teaching Resources Based on Density Clustering Algorithm

Pre-Processing of Topic Discussion Content

The language used in student-posted topic discussions differs from conventional document text, as it constitutes short texts of varying length, containing more complete information than longer articles but with fewer characters. Consequently, extracting valuable information from such text is crucial. Chinese text presents a complicated sentence structure and lacks specific separators, thus necessitating the conversion of the review text into a language that computers can comprehend. As illustrated in Figure 2, the original

Figure 2. Text processing measures



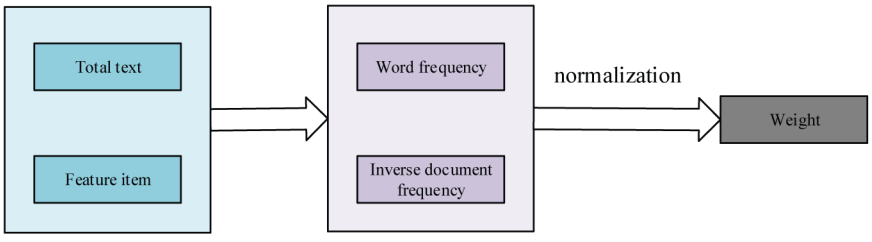
data was processed through case conversion, lexical summarization, deactivation, and the removal of special symbols during data processing. In the pre-processing of text content, each comment was split into different words since computers cannot comprehend the meaning of entire sentences. The initial dictionary only contains general common vocabulary, and the data used in this paper were sourced from the content released by the teaching platform. Therefore, we utilized the custom popular vocabulary dictionary as the foundation for word separation and added "English speaking practice" and "English tense" to the custom dictionary based on the subject matter of this study.

Improved TF-IDF-Based Weight Calculation

TF-IDF is one of the most exceptional and prevalent weight calculation methods. It can effectively assess the significance of a feature item for each document in a corpus, thereby facilitating sentiment analysis. It is widely applied in the field of information retrieval or data mining. Figure 3 demonstrates the weight calculation process in the research scheme.

If a feature item appears more frequently in the comment area of one course and less frequently in the comment areas of other courses, then the feature item is considered to make a more significant contribution to the classification. W_{ik} denotes the weight value of the feature item T_r in the text D_i, which is the frequency of the feature item T_k in the text D_i. IDF_k is the inverse document frequency, at which time the TF–IDF is calculated as follows:

Figure 3. Weight calculation process



$$TFIDF(D_i, T_k) = W_{ik} = TF_{ik} \times IDF_k \tag{1}$$

According to this formula, if the number of texts containing the feature item T_r in the training set is high, then the feature item is weak in distinguishing texts. Conversely, if the number of texts containing the feature item T_k in the training set is low, then we consider that the feature item is strong in distinguishing texts. The specific calculation of TF_{ik} is shown in Equation 2:

$$TF_{ik} = \frac{N_{ik}}{\sum_{j=1}^{n} N_{ij}} \tag{2}$$

where N_{ik} represents the number of times the feature item T_r appears in the document D_i and $\sum_{j=1}^{n} N_{ij}$ represents the sum of the number of times all feature items appear in the text D_i. The higher the number of times the feature item T_r appears in the document D_i, the larger the value of TF_{ik}.

The specific calculation of IDF_k is as follows:

$$IDF_k = \log \frac{N}{N_k} \tag{3}$$

where *N* denotes the total number of texts in the training set, and N_k denotes the number of texts containing the feature T_r in the training set.

N represents the total number of texts contained in the training set, and N_k represents the number of texts containing feature items in the training set. When *N* is a fixed number, the larger N_k and the smaller IDF_k, the less important the feature item is. Next, we organize and transform Equation 1 to obtain:

$$IDF_k = \log \frac{N}{N_k} \; TFIDF(D_i, T_k) = W_{ik} = TF_{ik} \times IDF_k = \frac{N_{ik}}{\sum_{j=1}^{n} N_{ij}} \times \log \frac{N}{N_k} \tag{4}$$

Density Clustering Algorithm With Sentiment Analysis

After conducting a frequency analysis and feature extraction of the text, we considered the correlation among teaching resources. Therefore, we delineated the scope of the weighted network's operation as follows:

$$f(s) = \{N(G_i) \mid \forall s, 1 \le i \le m\} \tag{5}$$

where $N(G_i)$ is the weighted grid action range, *I* is the grid within the weighted grid action range, and *m* is the total number of constructed grids. To set the specific weights of the weighted

grid, we adopted the principle of grid boundary expansion. Assuming that English education resource *a* exists in G_i, if there exists $a \in G_i$ and $G_i \in N(G_i)$ G,ÎN(G), the location of English education resource a is considered to be the boundary between two grid objects, and the corresponding correlation relationship exists between the English education resources of the corresponding grids.

The subsequent clustering stage requires the use of the merging process. At this time, we set the weight value of the grid to 1. Otherwise, no corresponding association relationship exists between the corresponding grid's English education resources, and the grid's weight value is set to 0. On this basis, the density of any grid is:

$$p(i) = P(density(i) = x) = \frac{count(t)}{count(n)} \tag{6}$$

Here, *p(i)* represents the total number of English education resources within the corresponding weighted grid after the gridding process of English education resources, while *density(i)* denotes the size of the grid cells involved in the statistics. Additionally, *count(t)* and *count(n)* represent the number of grid cells with density t and the number of non-empty grid cells, respectively.

Having defined the scope of the weighted network, we proceeded to incorporate additional layers to enhance the accuracy of our model. Specifically, we added a text expression layer, an attention mechanism layer, and a CNN network layer incorporating sentiment features to the density clustering because the original density clustering method struggled to detect contextual information within word sequences.

The text expression model with fused sentiment features uses the word2vec tool to train word vectors containing semantic information and the topic distribution extracted from each comment $r \in \{1,2,...,D\}$ $\theta \sim Dirichlet(\alpha)$, which is combined with the sentiment dictionary and lexical annotation tools to abstractly represent the attribute information of words, to obtain the sentiment distribution $\beta \sim Dirichlet(\xi)$, the semantic and attribute word vectors are combined to form a novel text word vector that contains both types of information. This new vector is then fed into a Bi-LSTM network that integrates current, past, and future information to extract the contextual information of the word sequences during the training process. Subsequently, a feature vector that incorporates the words' contextual, semantic, sentiment, and lexical information is generated.

After obtaining word vectors with fused sentiment features from the text expression layer, we generated sentence vectors through an attention mechanism layer. This mechanism acts as an adaptive selection process that continually optimizes attention parameters to distinguish primary and secondary positions of word sequence information in the text. The mechanism filters the word vectors with fused sentiment features with focus and sums up the sentence vectors, thereby removing redundancies and achieving the desired result.

Finally, the sentence vectors of the text are input to the CNN network in parallel to extract the spatial structure features of the text and achieve density clustering. Let the sentence vector be $(s_1, s_2, ..., s_L)$. Here, *L* represents the maximum number of sentences in the text. The CNN network leverages weight sharing to process the sentence vector in parallel, accelerating the processing speed. Moreover, based on the spatial height displacement invariance of the CNN network, the spatial structure information is extracted from the text, and the feature vector is extracted from the spatial structure information. Finally, the obtained feature vector is input into the softmax classifier to obtain the text emotion classification results.

EXPERIMENTS AND RESULTS

Data Processing

In this study, we obtained student comment content data from online English courses on the MOOC platform using Python web crawler technology. We conducted data preprocessing and Chinese text word separation processing to prepare the data for analysis. The web crawling process followed the procedure depicted in Figure 4, resulting in the acquisition of 48,181 pieces of student topic discussion data through the web crawler.

Analysis of Experimental Results

This research aimed to calculate the number of English education resources with the help of a map function centered on the target grid object and to obtain the corresponding key-value parameter values. Secondly, the computation results of the map function were combined and processed using the reduce function. In this stage, we used $key-value =< a, N(G_i), p(i) >\in \mu$ and $key-value =< a, N(G_i), p(i) >\notin \mu$ to calculate the density of each grid object after nesting it into the key-value parameter value, and we used the updated key-value parameter value as the new benchmark to calculate the state data of the next grid. In this way, we could classify the English education resources according to $key-value \in \mu$ and $key-value \notin \mu$.

To establish standardized criteria, our study encompassed four distinct categories of topic comments, with each category comprising 20 topics for comprehensive evaluation. The ensuing investigation primarily focused on comparing the efficacy of emotion recognition within the comments across varying topic features. Figure 5 visually depicts the outcomes derived from this comparative analysis, elucidating the discernible variations in emotion recognition effects associated with different topic attributes.

As shown in Figure 5, using the topic information extracted by our model and the topic-opponent sentiment information for comment topic detection led to a significant improvement in the detection effect. Furthermore, since topic discussions about teaching platforms often have a strong emotional tone, the detection effect of comments under our model for teaching platform comments was much better than for the other three topic discussions. These results demonstrate the effectiveness of the sentiment analysis model for the comment topics proposed in this paper.

To evaluate the effectiveness of the density clustering algorithm-based classification method for university English teaching resources proposed in this paper, we conducted comparative testing using

Figure 4. Crawl comments

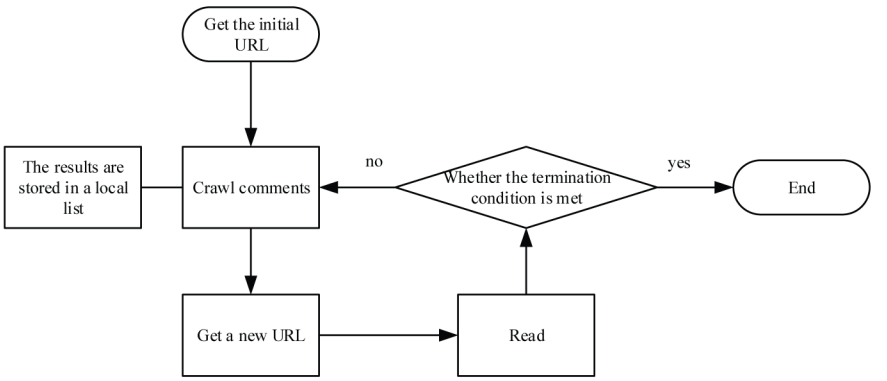
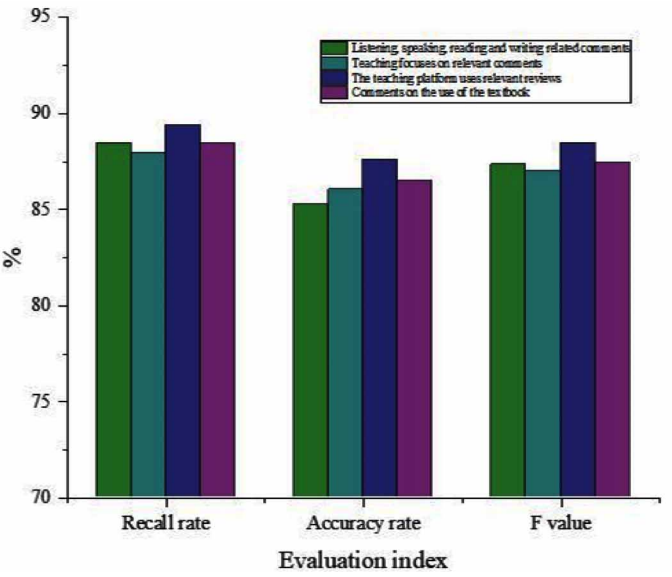


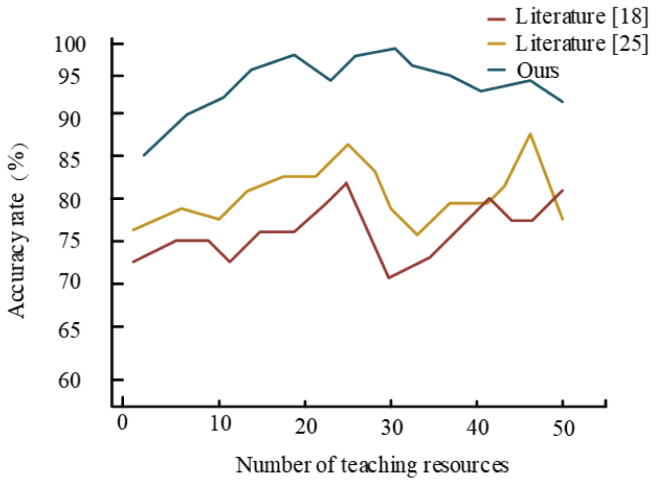
Figure 5. Comparison of the effect of emotion detection on different topic discussions



different classification methods. In the control group, we used the classification algorithm based on the BTM topic model and the clustering algorithm based on the combined neural network model (Geraldi & Ghisi, 2022) to demonstrate the performance of the TF–IDF improvement proposed in this paper and the performance of the clustering algorithm through comparison.

Figure 6 illustrates the classification accuracies of the various methods. As seen in the graph, while all three methods displayed some degree of fluctuation in their classification accuracy concerning the number of resources to be classified, the specific classification results varied significantly. Among these methods, the classification outcomes obtained by the method described in Geraldi and Ghisi

Figure 6. Classification accuracy of different methods



(2022) exhibited a relatively high degree of overall stability, with the classification accuracy for different categories of educational resources ranging from 76.0% to 87.0%. However, this method still has room for improvement. In contrast, the test results for the method outlined in Wu et al. (2020) displayed more noticeable fluctuations in the classification accuracy for different categories of educational resources, with the maximum value reaching 82.07% and the minimum value plummeting to 69.48%. By comparison, the method designed in this paper exhibited higher accuracy and stability as the number of teaching resources increased, with the corresponding parameter results attaining a higher level and the minimum and maximum values reaching 80.40% and 98.52%, respectively. The test results convincingly demonstrate that the proposed English teaching resources classification method achieves precise classification for various resource numbers.

Figure 7 shows that the F1 values for the different categories of English educational resources exhibit substantial disparities across the three classification methods. Notably, the F1 values obtained by the method described by Wu et al. (2020) ranged from 0.74 to 0.82, with the maximum value being attained when the number of teaching resources reached 32. Subsequently, the F1 values declined as the number of teaching resources increased. In contrast, the F1 values obtained through the method detailed by Geraldi and Ghisi (2022) displayed a more pronounced fluctuation range, with minimum and maximum values of 0.76 and 0.86, respectively. By comparison, the F1 values achieved through the method proposed in this paper were more impressive, with minimum and maximum values of 0.84 and 0.93, respectively. When modeling short texts, if the high-dimensional sparse problem is encountered using the method proposed in this paper, it can be effectively solved using an improved TF–IDF weight value to achieve the quality of hot topic extraction. These results affirm the effectiveness of the proposed method in accurately classifying various numbers of resources.

Figure 8 clearly illustrates that the recall rate of the algorithm proposed in this paper has increased to some extent in each case, except for a slight decrease in the recall rate when the number of categories is 8 or 23. By contrast, the comparison schemes outlined in Wu et al. (2020) and Geraldi and Ghisi (2022) exhibit more noticeable fluctuations in their recall rates, which are significantly lower in value than those of the scheme proposed in this paper because they do not take into account the weighting problem in word frequency analysis. While a slight downward trend in the recall rate occurs after the number of classifications reaches 43, when considered in conjunction with the accuracy and F1 values presented in Figures 6 and 7, it is evident that the improved TF–IDF algorithm enhances the classification performance by an overall three percentage points. These findings suggest that the

Figure 7. F1 values for different methods

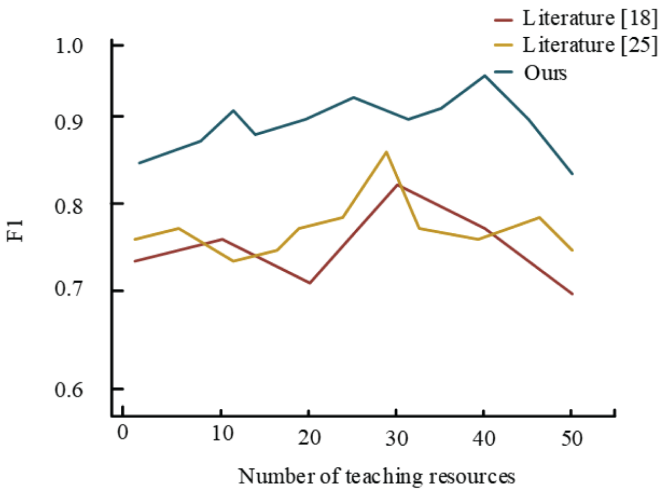
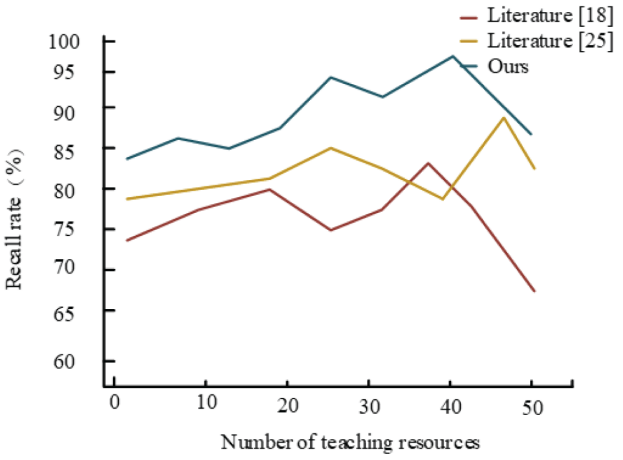


Figure 8. Recall rate values for different methods



improved weight calculation method, TF–IDF, has better weight assignment ability and superior sentiment analysis and classification capabilities than traditional TF–IDF.

Discussion

With the rapid development of information technology and the widespread use of the Internet, education has entered a new era of digitalization, in which teachers place new and higher demands on online teaching resources and services. In this context, analyzing the emotional content of students' topic discussions on online English teaching platforms is a critical area of research. The results of our experiment demonstrate that the proposed scheme can effectively classify educational resources, achieving an average classification accuracy of 91%, an F1 evaluation index of 0.9, and an average recall rate of 87%. These findings are of great significance as they optimize the organizational framework and classification system design of online teaching resource libraries, providing teachers with greater convenience and speed in their teaching applications.

CONCLUSION

As living standards continue to rise, individuals are increasingly emphasizing foundational education. Effectively and efficiently classifying the intricate text-based teaching resources available online presents a formidable challenge for educators. This study presents an enhanced approach to the conventional TF–IDF algorithm, employing automated text classification technology. Our improved algorithm incorporates parameters to ascertain the intra-class and inter-class distribution of feature terms, thus augmenting the precision of the weighted classification outcomes. Furthermore, we integrated a weighted network within the density clustering algorithm and leveraged a fusion of topic and sentiment distribution derived from each topic comment. The attention mechanism and CNN network layers were harnessed to facilitate sentiment feature extraction and model training for classification purposes. Ultimately, we conducted a comparative analysis of sentiment extraction across various topic discussions and performed longitudinal scheme comparison experiments to underscore the efficacy of our classification approach.

Our research underscores the substantial enhancement of both sentiment analysis and resource classification achieved through our algorithm. Nevertheless, it is imperative to acknowledge that the

current algorithm primarily addresses the cropping of text within high-density areas without adequately addressing the processing of low-density text segments. We also recognize the importance of vertically classifying school EFL resources. Thus, we propose establishing multiple classification systems that transcend disciplinary boundaries, fostering a cross-disciplinary classification methodology aligned with the competence index system. Our future endeavors entail exploring methods to effectively address low-density areas and implement vertical classification for teaching resources.

ACKNOWLEDGMENT

I want to thank the anonymous reviewers whose comments and suggestions helped improve this manuscript.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

FUNDING AGENCY

This work was not funded.

REFERENCES

Bustos López, M., Alor-Hernández, G., Sánchez-Cervantes, J. L., Paredes-Valverde, M. A., & Salas-Zárate, M. D. P. (2020). EduRecomSys: An educational resource recommender system based on collaborative filtering and emotion detection. Interacting with Computers, 32, 407–432. doi:10.1093/iwc/iwab001

Colangelo, P., Nasiri, N., Nurvitadhi, E., Mishra, A., Margala, M., & Nealis, K. (2018). Exploration of low numeric precision deep learning inference using Intel® FPGAs. In *IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)* (pp. 73–80). IEEE.

Dias Canedo, E., & Cordeiro Mendes, B. (2020). Software requirements classification using machine learning algorithms. *Entropy (Basel, Switzerland)*, *22*(9), 1057. doi:10.3390/e22091057 PMID:33286826

Geraldi, M. S., & Ghisi, E. (2022). Data-driven framework towards realistic bottom-up energy benchmarking using an Artificial Neural Network. *Applied Energy*, *306*, 117960. doi:10.1016/j.apenergy.2021.117960

Guleria, P., & Sood, M. (2018). Predictive data modeling: Educational data classification and comparative analysis of classifiers using Python. In *Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 740–746). IEEE. doi:10.1109/PDGC.2018.8745727

Hassan, S. U., Mohd Zahid, M. S., Abdullah, T. A., & Husain, K. (2022). Classification of cardiac arrhythmia using a convolutional neural network and bi-directional long short-term memory. *Digital Health*, *8*, 20552076221102766. doi:10.1177/20552076221102766 PMID:35656286

He, Q. (2022). Recent works for sentiment analysis using machine learning and lexicon based approaches. In *5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 422–426). IEEE. doi:10.1109/AEMCSE55572.2022.00090

Kambar, M. E. Z. N., Nahed, P., Cacho, J. R. F., Lee, G., Cummings, J., & Taghva, K. (2021). Clinical text classification of Alzheimer's drugs' mechanism of action. In Proceedings of Sixth International Congress on Information and Communication Technology: ICICT (Vol. 1, pp. 513–521). Springer Singapore.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(1), 100008. doi:10.1016/j.jjimei.2021.100008

Liu, C. Z., Sheng, Y. X., Wei, Z. Q., & Yang, Y. Q. (2018). Research of text classification based on improved TF-IDF algorithm. In IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE) (pp. 218–222). IEEE. doi:10.1109/IRCE.2018.8492945

Liu, Y., Ren, Z., Li, J., & Li, J. (2022). Design of informatization college and university teaching management system based on improved decision tree algorithm. Wireless Communications and Mobile Computing. doi:10.1155/2022/3127487

Lu, K., & Wu, J. (2019). Sentiment analysis of film review texts based on sentiment dictionary and SVM. In Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence. (pp. 73–77). ICIAI. doi:10.1145/3319921.3319966

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *2*(1), 86–97. doi:10.1002/widm.53

Newman, H., & Joyner, D. (2018). Sentiment analysis of student evaluations of teaching. In *Artificial Intelligence in Education: 19th International Conference. Proceedings, Part II* (pp. 246–250). Springer International Publishing. doi:10.1007/978-3-319-93846-2_45

Nguyen, K. L., Shin, B. J., & Yoo, S. J. (2016). Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In *International Conference on Big Data and Smart Computing (BigComp)* (pp. 223–230) IEEE. doi:10.1109/BIGCOMP.2016.7425917

Pan, M., Liu, A., Yu, Y., Wang, P., Li, J., Liu, Y., ... & Zhu, H. (2021). Radar HRRP target recognition model based on a stacked CNN-Bi-RNN with attention mechanism. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14.

Rashid, J., Kim, J., Hussain, A., Naseem, U., & Juneja, S. (2022). A novel multiple kernel fuzzy topic modeling technique for biomedical data. BMC Bioinformatics, 23, 275. doi:10.1186/s12859-022-04780-1 PMID:35820793

Séin-Echaluce, M. L., Fidalgo Blanco, Á., J. García-Peñalvo, F., & Conde, M. Á. (2015). A knowledge management system to classify social educational resources within a subject using teamwork techniques. In Learning and Collaboration Technologies: Second International Conference, Proceedings 1 (pp. 510–519). Springer International Publishing.

Shafqat, W., & Byun, Y. (2019). Identifying topics: Analysis of crowdfunding comments in scam campaigns. *Software Engineering, Artificial Intelligence, Networking and Parallel*. *Distributed Computing*, 137–148.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering, 336, 012017. IOP Publishing. doi:10.1088/1757-899X/336/1/012017

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of Machine Learning, 15, 713–714.

Wu, D., Zhang, M., Shen, C., Huang, Z., & Gu, M. (2020). BTM and GloVe similarity linear fusion-based short text clustering algorithm for microblog hot topic discovery. IEEE Access, 8, 32215–32225. doi:10.1109/ACCESS.2020.2973430

Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., & Wu, X. (2019). Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access, 7, 43749–43762. doi:10.1109/ACCESS.2019.2907772

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, *50*(2), 1–33. doi:10.1145/3057270

Zhang, F., & Wang, S. (2020). Detecting group shilling attacks in online recommender systems based on bisecting k-means clustering. IEEE Transactions on computational social systems, 7(5), 1189-1199.