Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Intelligent Grouping Method of Science and Technology Projects Based on Data Augmentation and SMOTE

Can Zhou, Mengting Li & Sha Yu

To cite this article: Can Zhou, Mengting Li & Sha Yu (2022) Intelligent Grouping Method of Science and Technology Projects Based on Data Augmentation and SMOTE, Applied Artificial Intelligence, 36:1, 2145637, DOI: 10.1080/08839514.2022.2145637

To link to this article: https://doi.org/10.1080/08839514.2022.2145637

Published online: 15 Nov 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2145637 (3506 pages)
https://doi.org/10.1080/08839514.2022.2145637

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS
Check for updates

Intelligent Grouping Method of Science and Technology Projects Based on Data Augmentation and SMOTE

Can Zhou[a], Mengting Li[a], and Sha Yu[b]

[a]School of Automation, Central South University, Changsha, China; [b]Special Management Department, China Science and Technology Exchange Center, Beijing, China

ABSTRACT
The current evaluation of science and technology projects is mainly completed by peer review, and in the process of evaluation, dividing projects into different groups is a crucial step. Project grouping is challenging due to the small amounts of data, sparsity of features, broad range of subject areas, and the seriously uneven distribution of categories. In this paper, we propose an intelligent automatic grouping method for science and technology projects based on keywords. We expanded the small dataset with samples generated by Paraphrasing, Mixup, and the GPT3 model. The text feature extraction techniques TF-IDF, Word2Vec, and TF-IDF weighted Word2Vec were utilized to pre-process the keywords of projects, and SVM and XGBoost as the classifier. Besides, we used SMOTE to process imbalanced data to alleviate model bias toward minority classes. Experiments show that the project grouping accuracy was substantially improved after introducing the data augmentation method and SMOTE. The combination of Paraphrasing, TF-IDF, SVM and SMOTE achieved the best performance, and the F1 score reached 96.78%, which proves the feasibility of the proposed method.

ARTICLE HISTORY
Received 3 July 2022
Revised 19 October 2022
Accepted 4 November 2022

Introduction

With the continuously expanding investment in scientific and technological innovation, the number of applications for scientific and technological projects is increasing. For instance, the number of applications funded by the National Natural Science Foundation of China in 2020 reached 281,170 (Zhao et al. 2021a), and in 2021, the number reached 287,323, an increase of 2.19% compared with 2020 (Hao et al. 2022). In addition, science and technology projects account for a large share in the allocation of science and technology resources, and the project approval, research process, and final achievements directly affect and promote the development of science and technology and society. Hence, the evaluation of science and technology projects is extremely important. To ensure the project review is smoothly carried out, it is necessary to group science and technology projects accurately at first, so as to provide

CONTACT Sha Yu yyyusha@163.com China Science and Technology Exchange Center, 54, Sanlihe Road, Beijing, China

references for recommending peer reviewers (Fang et al. 2022). At present, science and technology projects are mainly grouped manually by the disciplines selected by applicants when they declare projects, however, the chosen discipline may not be consistent with the practical field due to the broad range of subject areas and a high interdisciplinary degree. So, it is difficult to implement an accurate grouping of projects by merely relying on the disciplines selected. In addition, manual grouping is laborious and greatly influenced by subjective factors, which may pose threats to the fairness of the project evaluation. Therefore, it is of great significance to propose an intelligent grouping method to automatically group science and technology projects.

The science and technology grouping issue could be regarded as a short text classification issue. The science and technology projects usually have 3 to 5 independent keywords without contextual associations, and the project grouping issue aims to assign projects to corresponding fields based on the keywords contained in the projects. Therefore, the problem of project grouping is similar to short text classification. With the development of machine learning and deep learning, the short text classification task has achieved considerable performance (Alsmadi and Gan 2019; Deng, Cheng, and Wang 2021; Flisar et al. 2020; Yang et al. 2021). However, in addition to the inherent problems of short text classification, the science and technology grouping issue is also influenced by the small dataset and uneven distribution of categories. Motivated by the above factors, we propose an intelligent grouping method for science and technology projects to replace the laborious and time-consuming manual grouping method. To address the problem of sparse features and the absence of contextual semantics, we introduce the external Wikipedia corpus through the pre-trained Word2Vec model. We vectorize the keywords by Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and TF-IDF weighted Word2Vec, and use three different data augmentation methods to expand the small dataset. To tackle the uneven distribution of science and technology projects in different disciplines, this paper further utilizes Synthetic Minority Over-Sampling Technique (SMOTE) to deal with the unbalanced data and increase the Minority classes. In summary, our contributions are: to explore whether these three data augmentation methods can improve the classification performance of this small dataset, introduce SMOTE to process the imbalanced dataset to eliminate the impact of minority samples on the classification performance, and then propose a high-performance intelligent grouping method for science and technology projects.

Related Work

As this work involves the classification of imbalanced short text data with small samples, we review key related works in these areas.

Text representation is an important part of short text classification. The traditional bag-of-word (BOW) model such as One-hot Representation is simple and easy to implement, but it will cause the curse of dimensionality and ignore the original order and semantic relations between words, which is not effective in the short text classification issues (Sriram et al. 2010). In addition, although the TF-IDF model is widely used, and there are many improved models based on the TF-IDF model (Liu et al. 2018; Samant, Bhanu Murthy, and Malapati 2019), they just simply use word frequency without integrating semantics. The Word2Vec model (Mikolov et al. 2013a, 2013b) proposed by Google in 2013 is a low-dimensional word vector containing semantics, which is widely used in the field of natural language processing until now. Yilmaz and Toklu (2020) found that the use of different Word2Vec models has different impacts on the accuracy rate of different deep learning models in the question classification task. In 2014, the Glove model was proposed to generate word vectors by using characteristics such as word co-occurrence (Pennington, Socher, and Manning 2014). Furthermore, there are many innovative word representation methods based on the above methods. Many scholars combine Word2Vec and TF-IDF to form a novel text representation model (Liu et al. 2018; Zhu et al. 2016b; Zhu, Wang, and Zou 2016a), Rezaeinia et al. (2019) proposed a new text representation method by improving the Word2Vec model and the Glove model.

Traditional machine learning models for short text classification include Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), and so on (Hartmann et al. 2019). Deep learning models include the CNN model, the RNN model (Shen et al. 2018), and Transformer-based models such as Bert and GPT-3 (Brown et al. 2020; Devlin et al. 2018; Vaswani et al. 2017). Sharma and Shafiq (2022) used traditional machine learning models and deep learning models to classify and evaluate user intent in online reviews and social media. Noori (2021) used the DT model to classify customer reviews. Zhang, Zhao, and Lecun (2015) used character-level convolutional neural networks to classify text. Kim (2014) utilized convolutional neural networks to classify sentences. Liu and Guo (2019) utilized a bidirectional LSTM with an attention mechanism to classify text. Liu, Qiu, and Huang (2016) used recurrent neural networks to classify text. Chiu and Alexander (2021) used GPT-3 to identify hate speech and classify text as sexist or racist. Although word representation methods and text classification models are very mature, problems such as lacking sufficient contextual semantics, sparse features, and scarcity of text data will still affect the accuracy of short text classification. Liu, Li, and Hu (2022) proposed a CRFA model, which introduced the external knowledge base Probase in the embedding layer and then combined word vectors and corresponding entity vectors to alleviate the sparsity and ambiguity of short texts through multi-stage attention based on TCN. Flisar et al. (2020)

introduced DBpedia ontology which is structured data extracted from Wikipedia to perform feature extension on short texts. And some other works use their limited corpus to fine-tune the model pre-trained on a large corpus to append the semantic information, so as to tackle the problem of insufficient contextual semantics for short texts (Chang et al. 2020; Howard and Ruder 2018).

Data augmentation techniques are often used to alleviate the poor performance of text classification due to insufficient data. With the help of data augmentation, more data can be obtained to improve the classification effect, enhance the model generalization ability and improve the robustness of the model. But compared with computer vision, data augmentation in natural processing is more challenging since the text is discrete data, and inappropriate data augmentation may lead to text semantics changes (Li, Hou, and Che 2022). There are many text data augmentation methods, such as paraphrasing, adding noise, and sampling (Bayer, Kaufhold, and Reuter 2021; Liu et al. 2020; Shorten, Khoshgoftaar, and Furht 2021). Wei and Zou (2019) proposed Easy Data Augmentation (EDA) technique, which uses methods such as synonym replacement, random insertion, random exchange, and random deletion to augment text data. Chen, Yang, and Yang (2020) proposed the MixText model to generate text data through hidden layer interpolation. GPT-3 is also used in various NLP tasks with a few training data (Liu et al. 2021; Zhao et al. 2021b), and Balkus et al. (2022) used GPT-3 to classify whether a question is related to data science with the additional data generated by GPT-3.

Imbalanced data processing methods can be roughly divided into two categories, one is based on the data level including over-sampling methods, under-sampling methods, and hybrid sampling methods, and the other is based on the algorithm level including cost-sensitive method and ensemble learning method. Traditional methods such as random under-sampling and random over-sampling are widely used because of their simplicity, but they also may cause over-fitting, prolonged training time, and partial semantic information loss respectively. SMOTE is one of the basic over-sampling techniques used by scholars in handling class imbalanced issues (Chawla et al. 2002; Wang et al. 2019). Flores et al. (2018) combined SMOTE with SVM and Naïve Bayes for sentiment analysis. Sarakit, Theeramunkong, and Haruechaiyasak (2015) used SMOTE technique in an imbalanced YouTube dataset for emotion classification. The core of the cost-sensitive method is to assign different misclassification costs to different situations according to the cost matrix obtained by a great deal of prior knowledge. For the method is difficult to implement in practical problems, and tailored for specific problem, it is difficult to generalize to other problems. The ensemble learning methods have good classification effects and strong generalization ability, but the model is complex and sensitive to noise, and needs long training time (He and Garcia 2009; Wang et al. 2019; Xu, Chen, and Sun 2019).

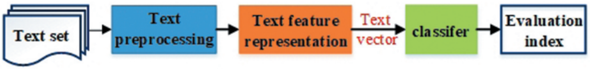


Figure 1. The process of text classification.

Proposed Methodology

Text classification is to assign text to the corresponding category according to the text content. As shown in Figure 1, text classification generally consists of text pre-processing, text feature representation, and training classifiers. Text pre-processing includes word segmentation, stop words filtering, and so on.

Figure 2 shows the overall architecture of the proposed method. We use three data augmentation methods to augment the small sample dataset, namely, Paraphrasing, Mixup, and generating data by Completion Endpoint of GPT-3. In the feature selection part, we compared the performance of TF-IDF, Word2Vec, and TF-IDF weighted Word2Vec model. And we used SMOTE to handle the minority classes. In the classification part, we chose SVM and XGBoost as our classifiers. SVM was proposed by Cortes and Vapnik (1995) and is widely used in the classification task. The learning strategy of SVM is to solve the separated line or hyperplane which can divide the training dataset correctly and has the maximum margin. The SVM also includes the kernel trick which allows it to be a non-linear classifier. XGBoost proposed by Chen and Guestrin (2016) is an open-source machine learning framework and one of the most popular algorithms for text classification and regression. It is an ensemble machine learning algorithm based on the decision tree, which adopts the idea of boosting, that is, gathering multiple classifiers to form a strong classifier. And we will explain the key methods above as follow.

TF-IDF

TF-IDF is a feature weighting technique commonly used in information retrieval and data mining (Kim and Gil 2019; Liu et al. 2018; Zhu et al. 2016b). The key idea of TF-IDF lies in that a word is not trivial to the text when it gets a high frequency in a text. Furthermore, if the word rarely or even does not appear in other texts except for the current text in the text set, the word has a strong ability to distinguish the current text and other texts. TF of TF-IDF is term frequency, which represents the frequency of occurrence of a word in the text. IDF of TF-IDF is the inverse document frequency, which means the lower word frequency is in other texts, the higher the IDF value is accordingly. The calculation formulation of the TF value of the word t_i in the text d_j is given as follows:

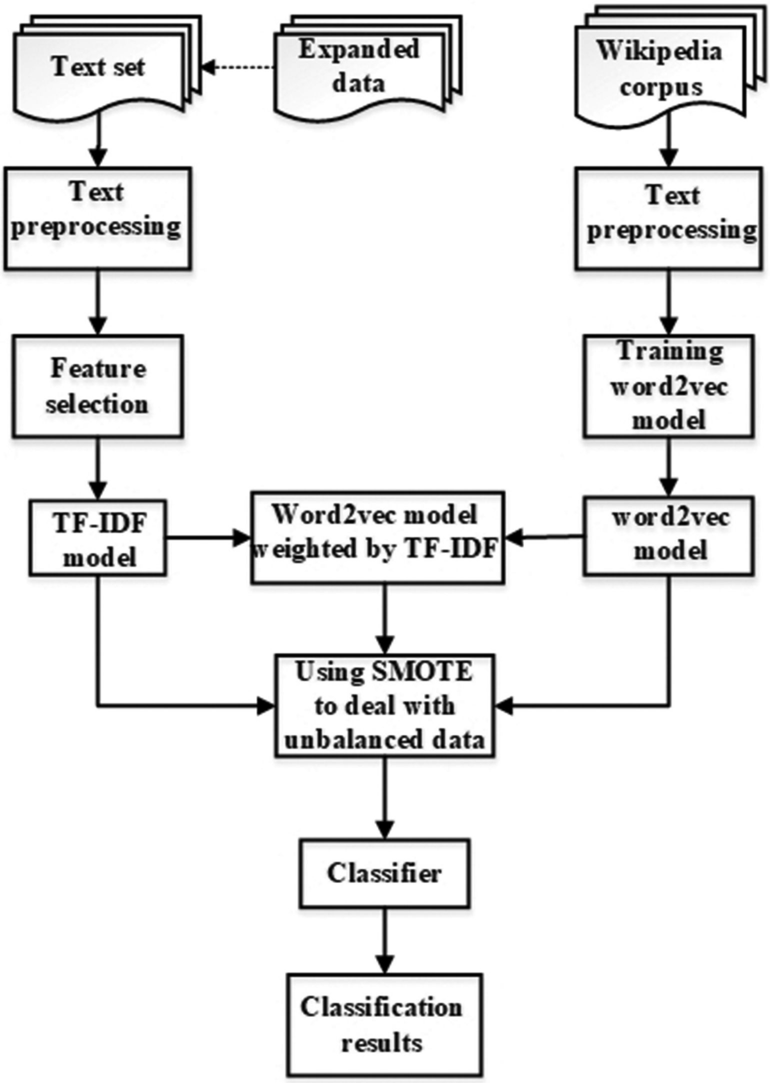


Figure 2. The overall model architecture.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where the numerator $n_{i,j}$ is the number of word t_i appearing in the text d_j, and the denominator $\sum_k n_{k,j}$ is the sum of words appearing in text d_j. The calculation formulation of the IDF value of the word t_i in the text d_j is given as follows:

$$idf_i = \log(\frac{D}{D(t_i)+1}) \tag{2}$$

where D is the total number of texts in the text set, and $D(t_{\mathrm{i}})$ represents the sum of texts that contain the word t_{i} in the text set. In case no texts containing the word t_{i} that will make the denominator of IDF become 0, so add 1 to the denominator. The calculation formulation of the TF-IDF value of the word t_{i} is given as follows:

$$TF-IDF_{i,j}=tf_{i,j}{}^{*}idf_{i} \tag{3}$$

The higher the TF-IDF value of the word, the stronger the text discrimination.

Word2vec

Word2Vec model based on the distribution representation is proposed by (Mikolov et al. 2013a, 2013b) in 2013. The Word2Vec model maps words to a low-dimensional vector of fixed length and evaluates the similarity between words by cosine distance. There are two learning algorithms in Word2Vec including Continuous Bag-of-Words (CBOW) and Skip-gram. CBOW predicts the current word by the context, and the length of the context is specified by the window size$k$. The mathematical expression is given as follows:

$$P(W_k|W_{t-k}, W_{t-k+1}, \ldots, W_{t+k-1}, W_{t+k}) \tag{4}$$

$W_{\mathrm{t\text{-}k}}$, $W_{\mathrm{t\text{-}k+1}}$, ..., $W_{\mathrm{t\text{-}k+1}}$, $W_{\mathrm{t+k}}$ represents the context, W_{k} represents the current word. Different from the CBOW model, the Skip-gram model uses the current word to predict the context, the mathematical expression is given as follows:

$$P(W_{t-k}, W_{t-k+1}, \ldots, W_{t+k-1}, W_{t+k}|W_k) \tag{5}$$

Generally, Skip-gram has a better effect than CBOW because it will train many times even low-frequency words, but the training time is inevitably longer than CBOW.

TF-IDF Weighted Word2vec

TF-IDF model can only represent the importance of words to the text, but does not contain the semantics of words. Although the Word2Vec model introduces semantics, it cannot distinguish which words are more important to the text. Therefore, using TF-IDF value to weight the vector after Word2Vec model, the attained vector of text d_{j} by TF-IDF weighted Word2Vec model is given as follows (Zhu, Wang, and Zou 2016a):

$$weight\_R(d_j) = \sum_{i=1}^{n} word2vec(w_i) \times tf-idf_{i,j} \tag{6}$$

n is the total number of words in the text d_j. Multiply the Word2Vec values by the corresponding TF-IDF values and add up to obtain the weighted Word2Vec values of the text.

Data Augmentation

Paraphrasing is one of the most widely used data augmentation methods which is easy to implement and can produce high-quality data (Li, Hou, and Che 2022). There are many companies such as Baidu and Google have opened translation interfaces due to the rapid development of machine translation. The procedure of paraphrasing is simple. Translating the text from the original language to another intermediate language, and then translating back to the original language to obtain the additional samples. Note that the intermediate language can be one or more. The label of augmented data is the same as the original text. Figure 3 shows an instance of paraphrasing, that translates the original text from Chinese to English and back to obtain the augmented text.

The Mixup augmentation method describes creating new augmented data by taking pairs of samples from the initial dataset and concatenating the words drawn from them together (Chen, Yang, and Yang 2020; Marivate and Sefara 2020). We randomly select two samples from the dataset, randomly sample n words from each of the two samples, and then mix them to generate new data. Where n is any integer from 1 to the number of words contained in the original text. Figure 4 shows an instance of Mixup. Three words are extracted from the original text above, while two words are extracted from the original

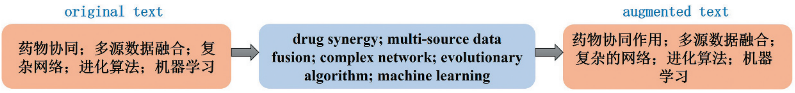

original text
药物协同；多源数据融合；复杂网络；进化算法；机器学习
drug synergy; multi-source data fusion; complex network; evolutionary algorithm; machine learning
augmented text
药物协同作用；多源数据融合；复杂的网络；进化算法；机器学习

Figure 3. Paraphrasing.

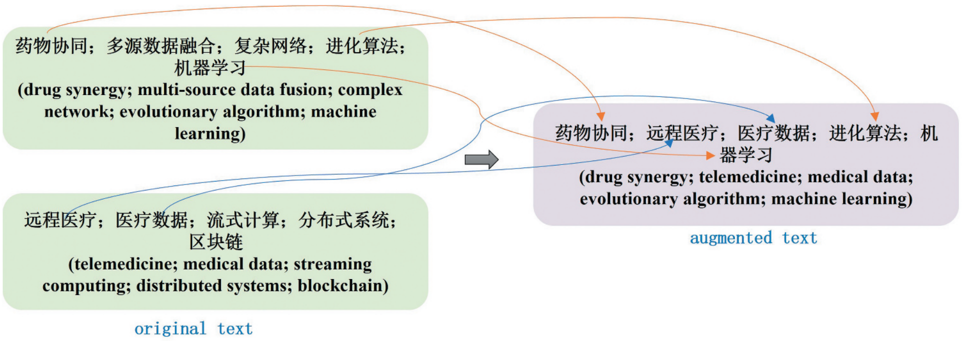
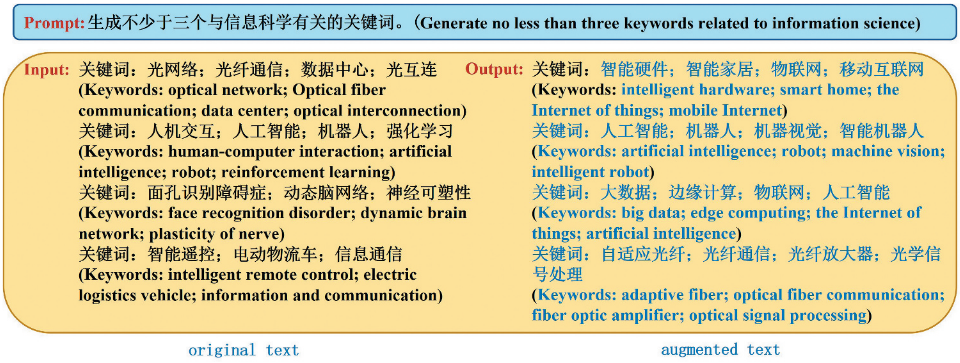

药物协同；多源数据融合；复杂网络；进化算法；机器学习
(drug synergy; multi-source data fusion; complex network; evolutionary algorithm; machine learning)
远程医疗；医疗数据；流式计算；分布式系统；区块链
(telemedicine; medical data; streaming computing; distributed systems; blockchain)
original text
药物协同；远程医疗；医疗数据；进化算法；机器学习
(drug synergy; telemedicine; medical data; evolutionary algorithm; machine learning)
augmented text

Figure 4. Mixup.



Prompt: 生成不少于三个与信息科学有关的关键词。(Generate no less than three keywords related to information science)

Input: 关键词：光网络；光纤通信；数据中心；光互连
(Keywords: optical network; Optical fiber communication; data center; optical interconnection)
关键词：人机交互；人工智能；机器人；强化学习
(Keywords: human-computer interaction; artificial intelligence; robot; reinforcement learning)
关键词：面孔识别障碍症；动态脑网络；神经可塑性
(Keywords: face recognition disorder; dynamic brain network; plasticity of nerve)
关键词：智能遥控；电动物流车；信息通信
(Keywords: intelligent remote control; electric logistics vehicle; information and communication)

Output: 关键词：智能硬件；智能家居；物联网；移动互联网
(Keywords: intelligent hardware; smart home; the Internet of things; mobile Internet)
关键词：人工智能；机器人；机器视觉；智能机器人
(Keywords: artificial intelligence; robot; machine vision; intelligent robot)
关键词：大数据；边缘计算；物联网；人工智能
(Keywords: big data; edge computing; the Internet of things; artificial intelligence)
关键词：自适应光纤；光纤通信；光纤放大器；光学信号处理
(Keywords: adaptive fiber; optical fiber communication; fiber optic amplifier; optical signal processing)

Figure 5. Generating data by GPT-3.

text below, and the label of augmented text is consistent with the original text with a larger n.

GPT-3 is a new milestone in the Natural Language Processing field. GPT-3 has a strong few-shot learning capability, which requires only a few training examples to teach the model to perform numerous tasks. The GPT-3 API provides the Completion Endpoint method so that the GPT-3 model can generate additional training samples with a few text-label pairs input as prompt (Balkus et al. 2022). Note that the quality of the generated data is related to the prompt format, the training examples, and even the order of the training examples (Zhao et al. 2021b). Figure 5 shows an instance of generating data by GPT-3. Provide the GPT-3 Completion Endpoint with the prompt "Generate no less than three keywords related to information science." followed by a few in-context examples. The tokens in black are the in-context examples fed to the model to ensure that the generation of the desired data, and the tokens in blue are the data generated by the model.

SMOTE

SMOTE (Chawla et al. 2002) calculates the Euclidean distance from every sample of the minority class to each remaining sample of the minority class and does not conduct any processing of the majority class data. Then randomly select a number between 0 and 1 to multiply the obtained Euclidean distance for interpolation to obtain a new minority sample, and the number of insertions is determined by specific demand. The mathematical expression is given as follows:

$$x_{new} = x + rand(0,1) \times (x_k - x) \tag{7}$$

x is a sample of a minority class, x_k is a sample selected from k nearest neighbors, and x_new is the newly synthesized sample.

Experiments

Experimental Data

The experimental data set comes from the projects of special project 2020 annual guide of the national key research and development project between the government. And the participant provided us with 858 pieces of text data containing information such as the group category of the project, the key field in the guide corresponding to the project, the research field of the project leader, and the project keywords. We extracted the project keywords and the corresponding project category as our experimental dataset. The project categories include 154 categories such as semiconductor, public management, clinical medicine, control theory, nanomaterials, and environmental engineering, and each project contains 3 to 5 keywords. Table 1 shows part of the experimental data.

Considering the data size is very small but the category is extra diverse, this will lead to poor training effect, so we regrouped the original 157 categories into 8 categories according to the classification criteria of The National Natural Science Foundation of China (NSFC). The National Natural Science Foundation of China Catalogue is a mature classification system in the field of science and technology programs. It divides projects into 8 categories including mathematical science, chemical science, life science, earth science, engineering and materials science, information science, management science, and medical science. The NSFC Catalogue will be updated and applied in the project classification in due course and can be used as an important reference source (Zeng, Jia, and Wu 2018). Therefore, we regrouped the science and technology projects into 8 categories according to the above criteria as shown in Figure 6. Table 2 shows some text data corresponding to the changed labels shown in Table 1.

Table 1. Partial experimental data set.

category	keywords
畜牧-饲料 (livestock - feed)	旱生牧草; 选育; 退化草地; 改良; 应用示范 (xerophyte forage; selective breeding; degraded grassland; improvement; application demonstration)
食品包装与储藏 (food packaging and storage)	复合材料; 蛋白质; 聚吡咯; 晶体结构; 光电性能 (composite material; protein; polypyrrole; crystal structure; optoelectronic properties)
计算机软件与计算机应用 (computer software and computer applications)	老龄人口主动健康; 数据可信管理; 健康状态评估; 智能服务导航; 云端协同 (active health of the aging population; data trusted management; health status assessment; intelligent service navigation; cloud collaboration)
微生物学 (microbiology)	食品有机废物; 乳酸; 微生物菌肥; 全组分资源化利用 (food organic waste; lactic acid; microbial fertilizer; full-component resource utilization)

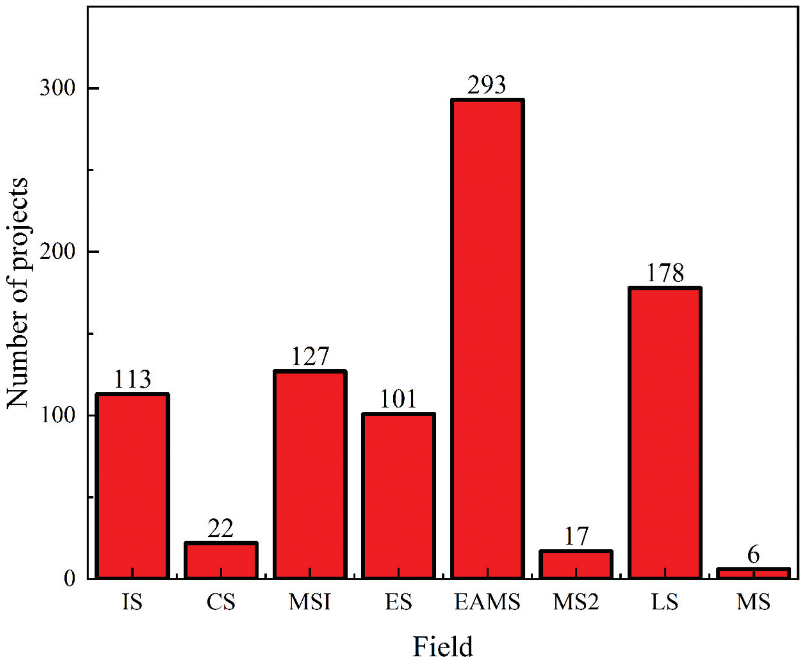


Figure 6. The reduced categories and corresponding quantity. (IS: Information Science, CS: Chemical Science, MS1: Medical Science, ES: Earth Science, EMAS: Engineering and Materials Science, MS2: Mathematical Science, LS: Life Science, MS: Management Science).

Table 2. Partial experimental data set after merging categories.

category	keywords
生命科学(Life Sciences)	旱生牧草; 选育; 退化草地; 改良; 应用示范(Xerophyte forage; selective breeding; degraded grassland; improvement; application demonstration)
工程与材料科学(Engineering and Materials Science)	复合材料; 蛋白质; 聚吡咯; 晶体结构; 光电性能(composite material; protein; polypyrrole; crystal structure; optoelectronic properties)
信息科学 (Information Science)	老龄人口主动健康; 数据可信管理; 健康状态评估; 智能服务导航; 云端协同(Active health of the aging population; data trusted management; health status assessment; intelligent service navigation; cloud collaboration)
生命科学 (Life Sciences)	食品有机废物; 乳酸; 微生物菌肥; 全组分资源化利用(Food organic waste; lactic acid; microbial fertilizer; full-component resource utilization)

Table 3. Confusion matrix of the classification result.

	predict value is positive	predict value is negative
the actual value is positive	TP	FN
the actual value is negative	FP	TN

Experimental Settings

We adopted the jieba to divide the keywords and removed the stop words such as ";" from the divided word after segmentation. We cleaned the external corpus Wikipedia and did the same pre-processing to the corpus. Utilizing the pre-processed corpus to train the Word2Vec model based on the Skip-gram algorithm, the size of the generated vector is 250. Considering the number in keywords in each science and technology project and of words after segmentation is different, we added up all the word vectors obtained in each text and took the average of the summation separately to obtain the vector of each text when using the Word2Vec model and TF-IDF weighted Word2Vec model for feature extraction of the experimental dataset.

We adopted three different data augmentation techniques: Paraphrasing, Mixup, and generating data by GPT-3. Paraphrasing extends data with English as an intermediate language by googletrans package. We used the Davinci engine in GPT-3 model and set the hyperparameter temperature to 0.5. Each data augmentation method generated the same amount of additional data as the original data with the unchanged class distribution.

We used linear kernel as the kernel function of SVM. And we used grid search to find the best hyperparameter combinations for the XGBoost model. The final XGBoost hyperparameters max_depth, learning_rate, gamma, reg_lambda and scale_pos_weight was 5, 0.15, 1.0, 1.0 and 1 respectively. We artificially synthesized instances to increase the number of all categories except Engineering and Materials Science to 293 through SMOTE, which was equal to the number of Engineering and Materials Science, the category with the most samples. And we split the dataset into 80% training and 20% testing dataset.

Experimental Index

We adopt the precision, recall, and F1 score to evaluate the performance of the text classification. The Table 3 presents the confusion matrix of the classification result.

The precision rate is the ratio of the instance number that the classifier correctly predicts as a positive instance. The formula is as follows:

$$precision{=}\frac{TP}{TP+FP} \tag{8}$$

The recall rate is the proportion of the instance number that correctly categorize positive instance. The formula is as follows:

$$Recall{=}\frac{TP}{TP+FN} \tag{9}$$

F1-score is a classification index that comprehensively considers accuracy rate and recall rate. The calculation formula is given as follows:

$$F1{-}\mathrm{score} = \frac{2^{*}precision^{*}recall}{precision+recall} \tag{10}$$

Table 4. Performance of the different models.

Model	Original	Paraphrasing	Mixup	Generating data by GPT-3
SVM+TF-IDF	47.52%	**79.64%**	**62.96%**	**69.96%**
XGBoost+TF-IDF	38.02%	63.38%	51.33%	61.12%
SVM+Word2Vec	**61.56%**	76.67%	56.34%	65.04%
XGBoost+Word2Vec	60.01%	75.27%	58.95%	69.52%
SVM+TF-IDF weighted Word2Vec	56.92%	71.92%	57.68%	67.91%
XGBoost+TF-IDF weighted Word2Vec	53.55%	71.04%	58.19%	69.40%

Results and Analysis

As can be seen in Table 4, there are different classification performances of different models before and after data augmentation. We use the F1 score to measure the quality of the classification effect.

As illustrated in Table 4, the dataset augmented by Paraphrasing had the best performance on the F1 score, which is 79.64% with the SVM model combined with the TF-IDF model. It indicates that the introduction of the data augmentation method can significantly improve the classification effect. Besides, the paraphrasing method has a high F1 score in comparison with alternative approaches, perhaps because it generates data more similar to the original data which makes it easier for the model to classify. From the perspective of classifiers, the SVM model generally outperforms the XGBoost model before and after data augmentation except for a few cases. We think there are several reasons why the SVM model performs better than XGBoost in this study. Although XGBoost is a very powerful machine learning algorithm and has achieved excellent performance in many tasks, it may be more suitable for processing small-size structural data or tabular data. Instead, SVM has a better effect on the text classification issue of such small datasets. SVM maps data to high-dimensional space, and it can classify the samples that cannot be linearly classified in the original space to reduce the probability of sample misclassification. And in many short text classification problems, many researchers chose SVM as the classifier instead of XGBoost which implies SVM might be more suitable for short text classification (Flisar et al. 2020; Luo 2021; Samant, Bhanu Murthy, and Malapati 2019). Nonetheless, the reason why SVM performs better than XGBoost needs to be further explored. From the perspective of the text representation model, the Word2Vec model performs best on the original data while the other three augmented datasets perform best with the TF-IDF model. Since the other three data augmentation techniques expand the dataset by adding similar samples, the TF-IDF model that uses the product of word frequency and inverse document word frequency as text representation is more advantageous. As Table 4 shows, the overall classification performance is not that ideal, the highest F1 score of the best-performing model is less than 80%. Note that uneven distribution data will greatly influence the performance of the classifier. As shown in Figure 6, we can see the number of projects varies greatly which results in poor classification effect of the classifier. Besides, as

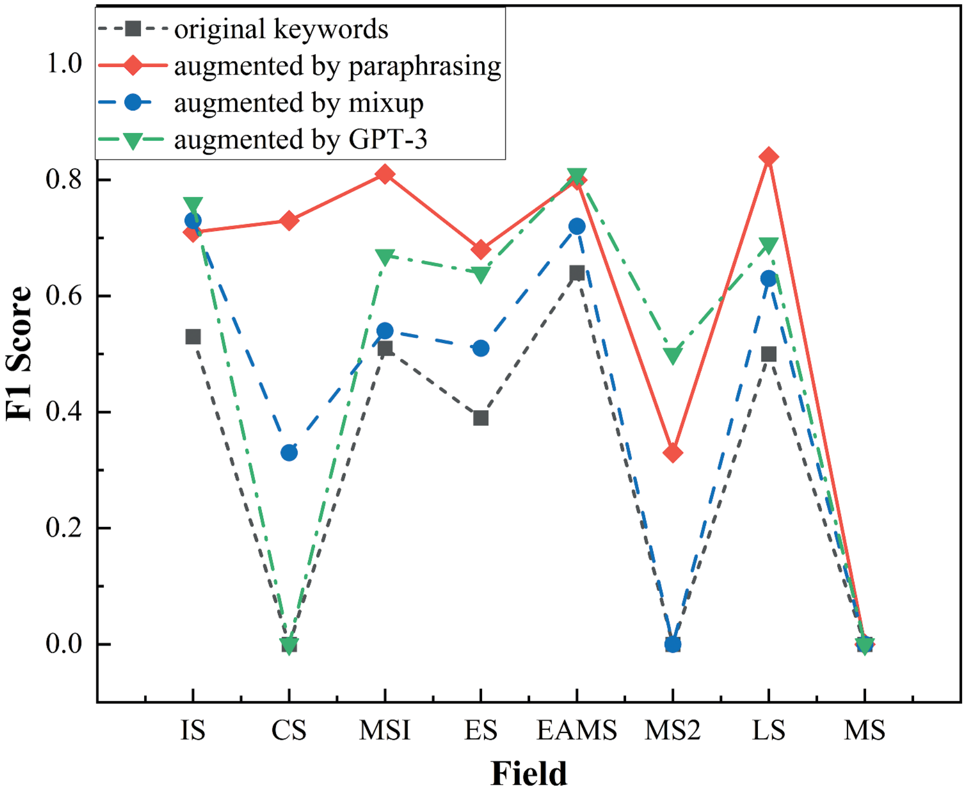


Figure 7. F1-score corresponding to each category. (IS: Information Science, CS: Chemical Science, MS1: Medical Science, ES: Earth Science, EMAS: Engineering and Materials Science, MS2: Mathematical Science, LS: Life Science, MS: Management Science).

shown in Figure 7, minority samples such as chemical science, mathematical science, and management science are hardly classified correctly. The low F1 score of the minority class samples pulls down the overall classification effect.

After the introduction of SMOTE to process the imbalanced data, The F1-score has risen significantly as can be seen in Table 5, and the F1-score of the best-performing model has risen to 96.78%. The SMOTE has greatly boosted the performance of the text classifiers.

It can be seen from Figure 8, the F1-score of the minority samples has been improved after artificially increasing the samples, which makes the overall classification effect improved. Since the original minority class data is limited

Table 5. Performance of the different models.

Model	Original	Paraphrasing	Mixup	Generating data by GPT-3
SVM+TF-IDF	77.82%	**96.78%**	**91.83%**	**94.65%**
XGBoost+TF-IDF	77.18%	90.82%	86.01%	89.85%
SVM+Word2Vec	**91.76%**	95.08%	84.43%	91.84%
XGBoost+Word2Vec	89.46%	93.91%	86.17%	91.45%
SVM+TF-IDF weighted Word2Vec	82.42%	94.07%	85.20%	90.95%
XGBoost+TF-IDF weighted Word2Vec	81.62%	92.61%	86.33%	93.28%

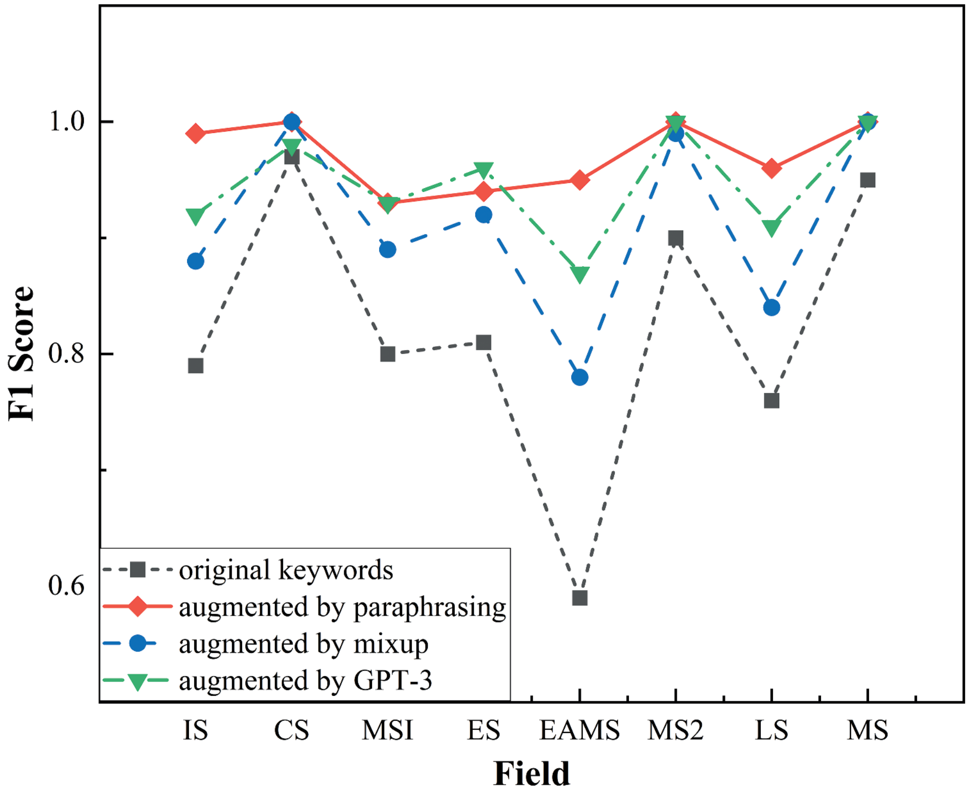


Figure 8. F1-score corresponding to each category after introducing SMOTE. (IS: Information Science, CS: Chemical Science, MS1: Medical Science, ES: Earth Science, EMAS: Engineering and Materials Science, MS2: Mathematical Science, LS: Life Science, MS: Management Science).

and the distribution is concentrated, which makes it easy to classify after SMOTE, the F1 score has been significantly improved to almost 100%, which has led to an increase in the overall classification effect.

Conclusion

This paper proposes an intelligent projects grouping method based on data augmentation and SMOTE. Considering the difficulty in obtaining training data, we use the data augmentation techniques to expand the data. By applying SMOTE technology for imbalanced experimental samples and artificially synthesizing minority sample data, the number of minority samples increases and the classification effect is greatly improved. Due to the small number of samples, the application of deep learning models will cause overfitting problems, and the classification effect is often unsatisfactory. Therefore, traditional machine learning models such as DT, NB, SVM, and XGBoost are more suitable for small datasets, and the SVM model is the most frequently chosen one among them.

Since the word segmentation is inaccurate and the external Wikipedia corpus does not contain all the words of our experimental data set, the pre-trained Word2Vec model cannot accurately vectorize all the words which affects the classification effect. Furthermore, the current group categories are only divided into eight categories, and the judgment of interdisciplinary projects is not accurate enough. Therefore, in the following research, we will further ameliorate the existing project grouping method to obtain a better project grouping model in response to the above problems.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Alsmadi, I., and K. H. Gan. 2019. Review of short-text classification. *International Journal of Web Information Systems* 15 (2):155–82. doi:10.1108/IJWIS-12-2017-0083.

Balkus, S., D. Yan, J. M. Shikany, S. V. Balkus, J. Rumbut, H. Ngo, H. Wang, J. J. Allison, and L. M. Steffen. 2022. A review of harmonization methods for studying dietary patterns. *Smart health (Amsterdam, Netherlands) arXiv preprint arXiv:2205.10981*. doi: 10.1016/j.smhl.2021.100263.

Bayer, M., M.-A. Kaufhold, and C. Reuter. 2021. A survey on data augmentation for text classification. *ACM Computing Surveys arXiv preprint arXiv:2107.03158*. doi: 10.1145/3544558.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33:1877–901.

Chang, W.-C., H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* San Francisco, California, USA, 3163–71.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research* 16:321–57. doi:10.1613/jair.953.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* San Francisco, California, USA, 785–94.

Chen, J. A., Z. C. Yang, and D. Y. Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *58th Annual Meeting of the Association for Computational Linguistics (Acl* Seattle, Washington, USA *2020*): 2147–57.

Chiu, K.-L., and R. Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3):273–97. doi:10.1007/BF00994018.

Deng, J., L. Cheng, and Z. Wang. 2021. Attention-based bilstm fused cnn with gating mechanism model for Chinese long text classification. *Computer Speech & Language* 68:101182. doi:10.1016/j.csl.2020.101182.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fang, Z., Z. Yang, L. Li, and T. Li. 2022. Multi-view project text classification based on cross-attention. *Journal of Chinese Information Processing* 36 (7):123–31.

Flisar, J., V. Podgorelec, C. Badica, M. Ivanovic, Y. Manolopoulos, R. Rosati, and P. Torroni. 2020. Improving short text classification using information from dbpedia ontology. *Fundamenta Informaticae* 172 (3):261–97. doi:10.3233/FI-2020-1905.

Flores, A. C., R. I. Icoy, C. F. Pena, and K. D. Gorro. 2018. An evaluation of svm and naive bayes with smote on sentiment analysis data set. In *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 1-4: IEEE.

Hao, H., Y. Zhao, Z. Zheng, H. Yang, Z. Gao, S. Zhang, Z. Li, C. Che, L. Yang, and C. Wang. 2022. Proposal application, peer review and funding of national natural science foundation of china in 2021:An overview. *Bulletin of National Natural Science Foundation of China* 36 (01):3–6.

Hartmann, J., J. Huppertz, C. Schamp, and M. Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing* 36 (1):20–38. doi:10.1016/j.ijresmar.2018.09.009.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21 (9):1263–84. doi:10.1109/TKDE.2008.239.

Howard, J., and S. Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kim, Y. 2014. *Convolutional neural networks for sentence classification, 1746-51*. Doha, Qatar: Association for Computational Linguistics.

Kim, S.-W., and J.-M. Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. *Human-Centric Computing and Information Sciences* 9 (1). doi:10.1186/s13673-019-0192-7.

Li, B., Y. Hou, and W. Che. 2022. *Data augmentation approaches in natural language processing: A survey*. China: *AI Open*.

Liu, G., and J. Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337:325–38. doi:10.1016/j.neucom.2019.01.078.

Liu, Y., P. Li, and X. Hu. 2022. Combining context-relevant features with multi-stage attention network for short text classification. *Computer Speech & Language* 71:101268. doi:10.1016/j.csl.2021.101268.

Liu, P., X. Qiu, and X. Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Liu, C. Z., Y. X. Sheng, Z. Q. Wei, and Y. Q. Yang. 2018. Research of text classification based on improved tf-idf algorithm. *2018 Ieee International Conference of Intelligent Robotics and Control Engineering (Irce)* Lanzhou, China: 218–22.

Liu, J., D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Liu, P., X. Wang, C. Xiang, and W. Meng. 2020. A survey of text data augmentation. Paper presentat at the 2020 International Conference on Computer Communication and Network Security (CCNS) Guilin, China.

Luo, X. 2021. Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal* 60 (3):3401–09. doi:10.1016/j.aej.2021.02.009.

Marivate, V., and T. Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 385-99: Springer.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 .

Noori, B. 2021. Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence* 35 (8):567–88. doi:10.1080/08839514.2021.1922843.

Pennington, J., R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* Doha, Qatar, 1532–43.

Rezaeinia, S. M., R. Rahmani, A. Ghodsi, and H. Veisi. 2019. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications* 117:139–47. doi:10.1016/j.eswa.2018.08.044.

Samant, S. S., N. L. Bhanu Murthy, and A. Malapati. 2019. Improving term weighting schemes for short text classification in vector space model. *IEEE Access* 7:166578–92. doi:10.1109/ACCESS.2019.2953918.

Sarakit, P., T. Theeramunkong, and C. Haruechaiyasak. 2015. Improving emotion classification in imbalanced youtube dataset using smote algorithm. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications* (*ICAICTA)*, 1-5: IEEE.

Sharma, A., and M. O. Shafiq. 2022. A comprehensive artificial intelligence based user intention assessment model from online reviews and social media. *Applied Artificial Intelligence* 36 (1):1–26. doi:10.1080/08839514.2021.2014193.

Shen, Y., Q. Zhang, J. Zhang, J. Huang, Y. Lu, and K. Lei. 2018. Improving medical short text classification with semantic expansion using word-cluster embedding. In *International Conference on Information Science and Applications*, 401-11: Springer.

Shorten, C., T. M. Khoshgoftaar, and B. Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data* 8 (1):101. doi:10.1186/s40537-021-00492-0.

Sriram, B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* Geneva, Switzerland, 841–42.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 .

Wang, Z., C. Wu, K. Zheng, X. Niu, and X. Wang. 2019. Smotetomek-based resampling for personality recognition. *IEEE Access* 7:129678–89. doi:10.1109/ACCESS.2019.2940061.

Wei, J., and K. Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Xu, X., W. Chen, and Y. Sun. 2019. Over-sampling algorithm for imbalanced data classification. *Journal of Systems Engineering and Electronics* 30 (6):1182–91. doi:10.21629/JSEE.2019.06.12.

Yang, T., L. Hu, C. Shi, H. Ji, X. Li, and L. Nie. 2021. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Transactions on Information Systems (TOIS)* 39 (3):1–29. doi:10.1145/3450352.

Yilmaz, S., and S. Toklu. 2020. A deep learning analysis on question classification task using word2vec representations. *Neural Computing & Applications* 32 (7):2909–28. doi:10.1007/s00521-020-04725-w.

Zeng, J., J. Jia, and W. Wu. 2018. Study of the construction of national technology program classification. *Journal of the China Society for Scientific andTechnical Information* 37 (8):796–804.

Zhang, X., J. Zhao, and Y. Lecun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems* 28 .

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh. 2021b. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 12697-706: PMLR.

Zhao, Y., Z. Zheng, H. Hao, Z. Gao, S. Zhang, Z. Li, C. Che, Y. Wang, and C. Wang. 2021a. Proposal application, peer review and funding of nsfc in 2020:An overview. *Bulletin of National Natural Science Foundation of China* 35 (01):12–15.

Zhu, L., G. J. Wang, and X. C. Zou. 2016a. A study of Chinese document representation and classification with word2vec. *Proceedings of 2016 9th International Symposium on Computational Intelligence and Design (Iscid)* Hangzhou, China, Vol *1*: 298–302.

Zhu, W., W. Zhang, G. Z. Li, C. He, and L. Zhang. 2016b. A study of damp-heat syndrome classification using word2vec and tf-idf. *2016 Ieee International Conference on Bioinformatics and Biomedicine (Bibm)* Shenzhen, China: 1415–20.