Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 4805537, 10 pages
https://doi.org/10.1155/2021/4805537

Research Article

Text Complexity Classification Data Mining Model Based on Dynamic Quantitative Relationship between Modality and English Context

Dan Zhang

Zhengzhou University of Industrial Technology, Zhengzhou, Henan 45000, China

Correspondence should be addressed to Dan Zhang; 201309020202@stu.sdnu.edu.cn

Received 24 November 2021; Accepted 18 December 2021; Published 30 December 2021

Academic Editor: Gengxin Sun

With the rapid development of mobile internet technology, there are a large number of unstructured data in dynamic data, such as text data, multimedia data, etc., so it is essential to analyze and process these unstructured data to obtain potentially valuable information. This article first starts with the theoretical research of text complexity analysis and analyzes the source of text complexity and its five characteristics of dynamic, complexity, concealment, sentiment, and ambiguity, combined with the expression of user needs in the network environment. Secondly, based on the specific process of text mining, namely, data collection, data processing, and data visualization, it is proposed to subdivide the user demand analysis into three stages of text complexity acquisition, recognition, and expression, to obtain a text complexity analysis based on text mining technology. After that, based on computational linguistics and mathematical-statistical analysis, combined with machine learning and information retrieval technology, the text in any format is converted into a content format that can be used for machine learning, and patterns or knowledge are derived from this content format. Then, through the comparison and research of text mining technology, combined with the text complexity analysis hierarchical structure model, a quantitative relationship complexity analysis framework based on text mining technology is proposed, which is embodied in the use of web crawler technology. Experimental results show that the collected quantitative relationship information is identified and expressed in order to realize the conversion of quantitative relationship information into product features. The market data and text data can be integrated to help improve the model performance and the use of text data can further improve predictions for accuracy.

1. Introduction

The meanings of modal verbs are highly uncertain, and their specific meanings are largely influenced by the context in which they appear [1], which brings great difficulties to natural language understanding and processing as well as machine translation. Therefore, it is very important to study the interaction limitation between the modal verb meaning and contextual features. It is of great significance to reveal the data structure relationship between the modal verb meaning and contextual features and to discover and identify the important contextual features of modal verb meaning, providing an important theoretical and practical basis for a modal verb meaning disambiguation and natural language processing. As modals have been a hot topic in linguistics, philosophy, and natural language processing, many scholars have paid much attention to them. Studies on modal meaning in traditional linguistics mainly focus on semantic classification and grammatical function description of modal verbs [2].

The interaction between modality and context has a great application prospect. In many intelligent systems, such as public opinion analysis, text filtering, recommendation system, and e-commerce, the interaction between modality and context can also play its role to make intelligent systems achieve better results. In the analysis of public opinion, the interactive relationship between modality and context can be used to find out whether people are for or against the event, and corresponding measures can be taken to control public opinion. In text filtering, when the sensitive information is filtered, judging whether the subject-sensitive information is

negative or positive through the interaction between modality and context can effectively reduce the error filtering of positive subject-sensitive text. In the recommendation system, the evaluation text of products can be analyzed, and the public's evaluation of corresponding products can be fed back to the user. In e-commerce, the interactive relationship between modality and context of online comments or information on websites can be used to grasp the market response of products in real-time and make timely countermeasures.

In recent years, scholars have paid more attention to the syntactic and semantic evolution of modal verbs [3], semantic and pragmatic functions [4], grammatical and semantic categories [5], semantic features and subjectivism [6], and the process of modal grammaticalization [7]. In terms of the relationship between word meaning and context, this paper studies the restrictive effect of news report context on the meaning of modal verbs. By analyzing the effect of context on modality, it is found that all factors constituting the context will affect the meaning of modal verbs, the meaning of discourse is restricted by its context, and context has a limiting effect on the generation and understanding of discourse [8]. This paper studies the meaning of modals in different contexts and points out that, the meaning of modals depends on the context in which they are located. As can be seen from the abovementioned studies, scholars have studied the modal meaning and the constraining effect of context on modal meaning from different perspectives. However, previous studies are mainly based on the analysis of the surface syntactic features of observed examples. With the deepening of modal semantics research, more attention is paid to the interaction between modality and context, and to the contextual features that have greater restrictions on modal semantics. In order to provide a theoretical and practical basis for the research of modal semantics, natural language processing, and intelligent semantic recognition feature selection, this paper uses the unique attribute feature extraction method to study the interactive relationship between the meaning of English modal verbs and multi-dimensional context features and reveals the interactive constraint relationship between them.

2. Related Work

For modal verbs such as semantic complex are highly sensitive to the context in terms of parts of speech, only considering the co-occurrence of semantic and syntactic characteristics has great limitations, difficult to fully disclose and found that the nature of the interaction between the modal, semantic, and contextual relationship, therefore, there is a need to consider the multidimensional context characteristics, more comprehensive, also highlights the semantic disambiguation considering different characteristics of the importance of context in [7]. According to its input and output, relational extraction can be further divided into several similar sub-tasks. Among them, relation classification gives the context of two entities as inputs, with the purpose of learning a classification model and predicting the semantic relations of these two entities [9]. There are two main limitations of this task. Only a fixed number of relationships can be predicted, and all relationship categories need to be specified in advance. A large amount of manual annotation data is needed as input of the classifier. Open relation extraction overcomes the first limitation by automatically extracting "subject-predicate-object" structure from unmarked texts as candidate relation tuples [10]. In this series of algorithms, the categories of relationships between entities need not be specified in advance. In order to solve the second problem, remote supervision [11] uses the existing relationships in the knowledge graph to provide relationship annotation information for the unlabeled texts, thus greatly reducing the workload of manual annotation. In recent years, the research on relation extraction is gradually combined with deep learning algorithms, including adversarial learning and reinforcement learning. In addition to generic and lexical relation extraction based on unstructured texts, there are a few studies aimed at obtaining knowledge from shorter texts, such as open relation extraction based on nouns [11], relation extraction based on search engine query records [12], etc. Among these short texts, Wikipedia categories frequently receive academic attention because the language is more regular and generally describes the entity's information, e.g., [13]. The grammar and word-formation features based on Wikipedia categories are designed, and the corresponding relation extraction algorithm is proposed.

Information extraction of the interactive relationship between modality and context: it means extracting valuable emotional information from emotional texts. According to different extraction objects, it can be divided into extraction of emotional words, extraction of evaluation objects, extraction of opinion holders, and extraction of combined evaluation units. (1) Extraction of emotional words: It is mainly based on the existing corpus or dictionary to mine the statistical features or semantic relationships between words. According to the rule that adjectives connected by conjunctions have strong emotional relevance, a large number of descriptive evaluation words have been excavated from the corpus Wall Street Journal [14]. According to the similarity principle, the word clustering method is used to extract, for noun evaluation words, seed evaluation words and rule templates are manually selected, and noun evaluation words are extracted by the iterative method [15]. The method of point mutual information is proposed to evaluate the judgment of words. Some researchers first manually select some evaluation words as seed words, and then expand the seed words through dictionaries and annotations in dictionaries. The point mutual information proposed by Wu is used to judge the evaluated words by calculating the correlation degree between the adjectives and seed words in the dictionary. (2) Extraction of evaluation objects: Templates can be made manually, and rules are often related to the characteristics of language. For example, the parts of speech rule of progressive grade [16]; methods of association rule mining [17] or based on the results of syntactic analysis; some scholars try to judge the correlation between words and domain indicators by selecting the domain indicators [18]. With the rise of the topic model, the evaluation object is judged by the topic model. The evaluation objects of products were extracted by multi-grain topic models and similar clustering was carried out. (3) Extraction of opinion

holders: The opinion holders are identified by named entities and semantic role analysis, respectively [19].

English language complexity of the existing text classification method mostly rely on artificial feature extraction, the design characteristics of knowledge domain experts combined readability can often has a high correlation with the readability of text, the method can obtain a better evaluation result, but the characteristics of effective design often require researchers' repeated observation and experimentation, and it is very time-consuming [20]. With the rise of deep learning technology, researchers have found that a more efficient way of text representation is to automatically learn the text representation from data, also known as representation learning or automatic feature learning. In view of this, this paper studies the readability assessment based on representation learning. In the text representation learning technology, word representation, also known as word embedding, is generally learned first, and then the expression of sentences, paragraphs, and chapters based on word representation [21]. Through practical research, it is found that the word embedding directly obtained by using the general word embedding learning model cannot achieve satisfactory results in the readability assessment task, which may be because the general word embedding does not contain readability domain information, and the word embedding learning model needs to be further improved. Therefore, in the readability assessment task based on the representation learning framework, this paper carries out a more in-depth study on automatic feature learning technology [22, 23]. In the specific research, this paper measures the difficulty of words based on domain knowledge, and then transforms the word embedding model with text information, so that the word embedding can contain readable information [24]. The innovative points of this paper are as follows: (1) Combining mathematics, information science, and linguistics knowledge, using corresponding computer software, this paper studies the interaction between the modal verb meaning and context characteristics, and excavates hidden knowledge in the language data structure; (2) mining the relationship between the semantic conceptual structure data using the unique attribute feature extraction method based on formal concept analysis. Because this method can mine more concise classification rules and unique attribute features than other methods, it can reflect the interaction between modality and context more directly and clearly. (3) It overcomes the shortcomings of existing linguistic research methods and adopts formal semantic analysis method and computer technology to analyze the interactive restriction relationship between the modal verb meaning and context features. The research results are more scientific, reasonable, and effective.

3. Data Mining for the Interaction between Modality and Context in English Text Context Extraction

3.1. Semi-Supervised Extended Model of Upper–Lower Relation. The model-based scoring features refer to scores calculated according to existing readability models, such as readability formula features, statistical language model features, similarity model features, etc. For readability formula functionality, recalculation of formula coefficients is required by linear regression. For statistical language model features, a smooth unary model is constructed for each reading level, and then the probability of each document generated by the corresponding language model at each level is calculated. For the similarity feature, this work computes the cosine similarity between the target document and the known reading level document, each document represented by a vector <word, frequency>. Since these characteristics are highly correlated with reading levels, they can be useful.

Since the complexity text classification problem of the English language can be regarded as an ordered multiclassification problem, this work proposes an ordered classification method, which uses multiple dichotomous classifiers and votes the classification results to capture the sequential relationship between the reading levels. Its frame diagram is shown in Figure 1.

> Voter 1: For each target document, decisions are made along with $n-1$ results in ascending order from the first to the $n-1$. If the current i^{th} result is 0, the reading level is $L1$; otherwise, move to the i^{th} +1 result. If the result of the last classifier is 1, the final reading level of the document is Ln.

3.2. Text Context Classification Model Based on Modality and Context Interaction. The method framework of GRAW+ is drawn in Figure 2. GRAW+ takes auxiliary sentence corpus and target data set as input. The auxiliary sentence corpus consists of unmarked sentences and is mainly used to construct a word coupling matrix in reading difficulty. The target dataset is a readability document dataset that contains both tagged and untagged documents. The goal of GRAW+ is to predict the reading level of unlabeled documents based on labeled documents. GRAW x consists of two stages, feature representation and readability classification.

In the first stage, documents in the dataset are mapped to feature vectors, which can be extracted from two perspectives, from the perspective of the conjunction bag model and the perspective of language. From the perspective of the coupled-word bag model, the coupled-word bag model can be used as a feature, which is a variant of the basic word bag model applied to readability evaluation. From a linguistic point of view, it is possible to extract appropriate linguistic features, including those previously proposed and validated by many readability assessment researchers. By representing documents from both perspectives, the method measures both word-level difficulty distribution and document-level readability-related factors, which can provide more information for subsequent readability classifications.

In the second stage, a two-view graph propagation method is proposed for the readability classification, which consists of three steps, graph construction, graph merging, and label propagation. In the first step, all documents (both labeled and unlabeled) are used for composition, each document is represented as a node, and their similarity in readability is represented by edge weights. In the second

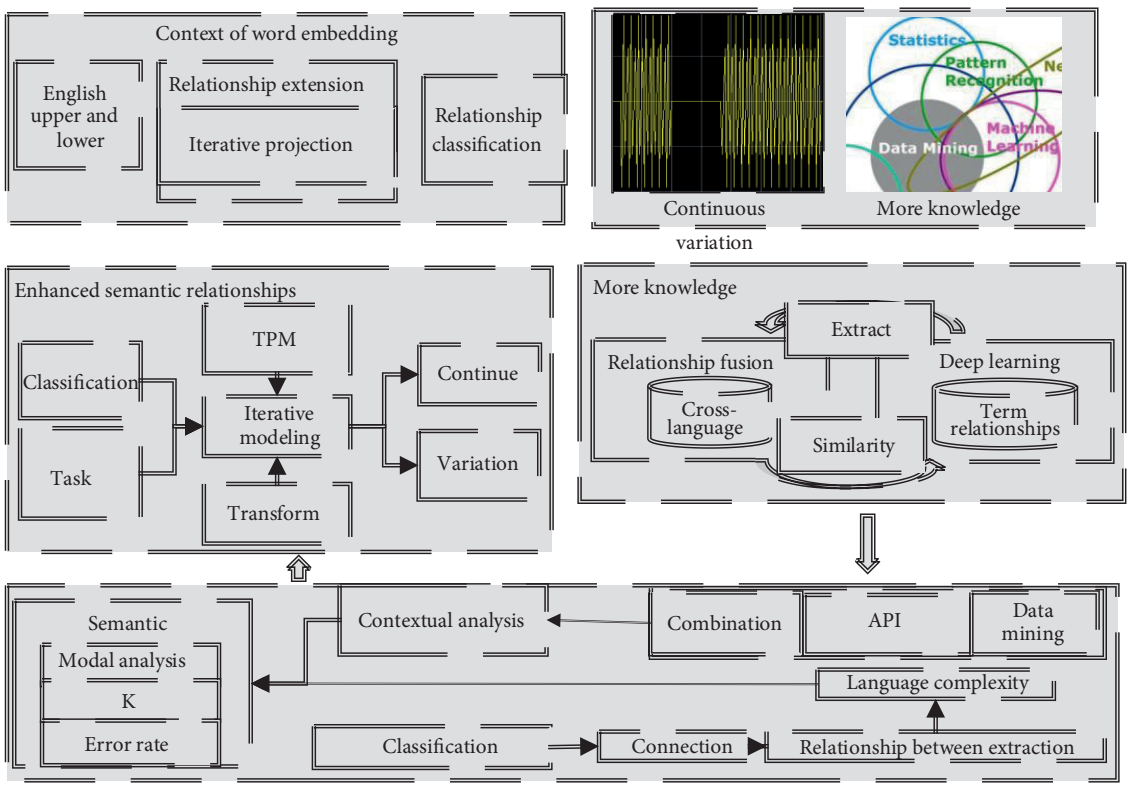

Figure 1: English language complexity text classification framework for modal and context data mining.

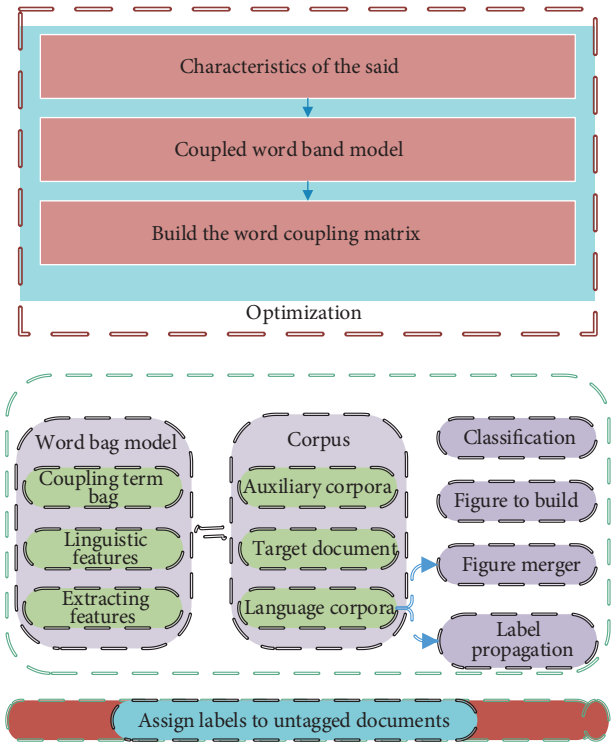

Figure 2: Grow Apriori+ framework.

step, the intra-view merge is used to merge homogeneous graphs from the same view, while the inter-view merge is used to merge heterogeneous learning from different views.

In the third step, the label propagation algorithm is used to label propagation on the merged graph and obtain the reading level of the unlabeled document.

Step 1. Estimate the difficulty of reading sentences. Accurate estimation of sentence-level readability is a difficult problem, which has attracted the attention of many researchers in recent years. For efficiency, some heuristics are used here to make rough estimates. Specifically, consider a rough estimate of the difficulty of a text using features that reflect its readability from different angles. Eight heuristic functions {Len, ANS, ANC, LV, ATR, NTR, PTH, and ANP} are constructed from the classical features designed for the document-level readability assessment, and the most commonly used features suitable for sentence-level measurement are selected. These features can be used to measure sentence readability from three aspects, as shown in Table 1.

Step 2. Word difficulty distribution estimation calculates the difficulty distribution of each word based on the sentence-level reading difficulty. Since each sentence contains several words, and each word appears in several sentences, the difficulty distribution of words can be estimated from the frequency of words appearing in sentences of different difficulty levels.

$$t_i(i) = \frac{1}{t_n} \sum_r \beta(n \in t) * \delta(t) = i. \quad (1)$$

Step 3. Word coupling matrix construction. Given word set V, the coupling matrix is defined as C, and the

Table 1: Heuristic functions for estimating sentence reading difficulty.

Terms	Function	Description
Surface	Len(t)	The length of the sentence T
	Ans(t)	The average number of words (Chinese characters) and syllables (strokes) of the sentence T
	Anc(t)	The average number of word letters (Chinese characters) of the sentence T
Lexical	Lv(t)	The average number of parts-of-speech categories of the sentence T
	Art(t)	The proportion of adjectives in the sentence T
	Ntr(t)	The noun ratio of the sentence T
Grammar	Pth(t)	The syntax tree height of the sentence T
	Anp(t)	The average number of phrases (noun phrases, verb phrases, and prepositional phrases) in the sentence T

elements of the matrix represent the correlation between words. The correlation of each pair of words can be measured by the similarity of the word's difficulty distribution.

Considering that a large number of word sets makes the construction of word conjunction matrix time-consuming, a word filtering strategy is designed here to filter out words with less information based on the word difficulty distribution. The filtering indicator is the information state of the words. The words are arranged in ascending order according to the information state of the words, and the final proportion of the words will be filtered out.

$$E(t) = \sum_{i} p_t(i) \log\ p_i(i). \tag{2}$$

The situation awareness model is divided into three parts: situation awareness, situation understanding, and situation prediction.

(1) Situational awareness is to perceive the attributes, states, and dynamics of relevant elements in the environment, identify the key elements or events and combine the information obtained from the environment as the knowledge reserve of the environment.

(2) Situation understanding based on situation awareness, understands the elements of the perceived target according to its mission purpose, and form a global situation map, including the understanding of events and objects.

(3) Situation prediction is based on the understanding of the elements of the perceived target and the whole situation, forecasts the development of the situation.

Starting from the three parts of situation awareness, situation understanding, and situation prediction of situation awareness model, combined with the idea of situation awareness, the text emotion analysis is divided into three parts: extraction of emotion information, the establishment of the emotion classification model, and the use of emotion classification model. Emotional information extraction is a basic method in which, the emotional analysis of the text sentiment analysis is simplified to the emotional information in text classification, this method is to extract the emotional comments that have emotional meaning in the text of the information unit, converting the original text of unstructured to the computer to identify the processing of structured text, used to train the sentiment analysis model. It is similar to identifying critical elements and events in situational awareness and providing material for subsequent work. The establishment of the emotion classification model uses the extracted and processed emotion information to establish the emotion text classification model by some method, which provides services for the direct classification of subsequent texts. This part corresponds to the situation understanding part of situation awareness, and the model training in this stage includes text understanding and analysis. Finally, the use of the classification model of English language complexity text classification can be seen as an interface for users to provide services, mainly for the prediction of texts with unknown emotional polarity. This part corresponds to the situation prediction part of situation awareness and is part of the interaction between the whole system and users.

4. Data Mining of Interaction between Modality and Context Based on Unique Attribute Features

Based on formal concept analysis theory and formal background unique attribute feature extraction method. Formal concept analysis is a branch of applied mathematics. Concept lattice is its core data structure. Each node of a concept lattice is called a formal concept, which consists of denotation and connotation. Denotation represents the collection of all objects belonging to this concept; the connotation represents the set of properties common to all of these objects. The formal background that reflects the relationship between the objects and attributes is the basis of generating concept lattices. The formal background is also an effective mathematical tool for knowledge discovery and data analysis. It has been successfully used in knowledge discovery, semantic disambiguation, and visualization. A flowchart for this approach is shown in Figure 3.

By comparing and analyzing the distribution of attribute features in these rules, we can find the interactive relationship between different meanings of the modal verb MUST and different contextual features. In the process of distinguishing the three kinds of meanings of MUST, four kinds of context features play an important role. They are the semantic feature, syntactic feature, pragmatic feature, and topic feature.. By comparing and analyzing the classification rules of three kinds of word meanings, it is found that

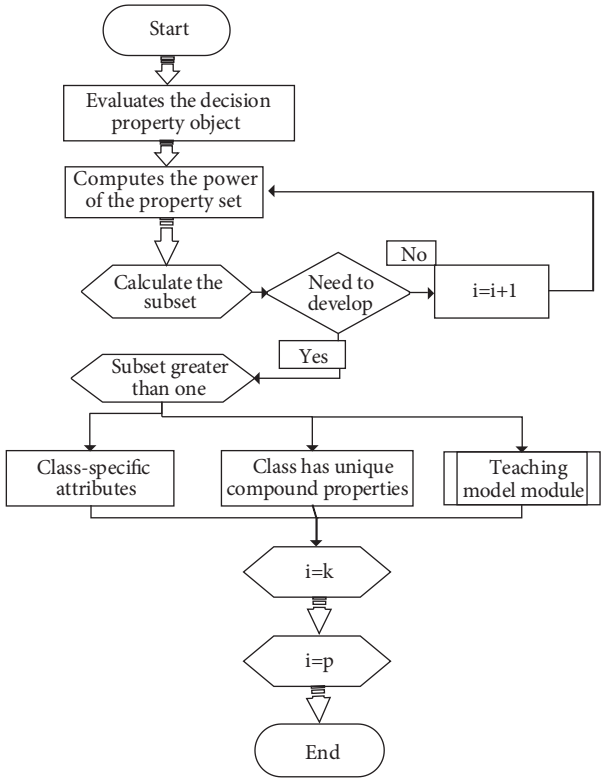

Figure 3: Flow chart of unique attribute calculation method.

(1) There are many unique attribute features in the three types of semantic classification rules of MUST. Where syntactic context feature A is a simple unique attribute feature, that is, the feature is only possessed by an object in the third class and is a class-unique attribute feature. The pragmatic features in context features are all unique attribute features of the first kind of sense must, that is, these features have a direct limiting effect on MUST. As long as one of these features appears in the sample examples, MUST must be the meaning of MUST. Similarly, topic feature (topic related to natural law) is the class unique attribute feature of the second kind of meaning for MUST. Syntactic features A19 (perfect tense) and A20 (continuous tense) are unique attributes of the third type of word, MUST. These class unique attribute features are the strong classification features of MUST because they limit the meaning of MUST. In turn, they also reflect the strong sensitivity of these context features to different meanings of MUST, and they only belong to and support a certain meaning of MUST.

(2) Of the three senses of MUST, the first is the strongest, followed by the second, and the third is the weakest.

(3) Among the composite attribute features unique to the class, semantic features appear the most, and the low value (odd number) point mutual information (MI, 0) features occupy the majority. Since MI, 0 means that two words (MUST and adjacent words) are not semantically related in the point mutual information, it indicates that in the semantic classification of MUST, the semantic non-correlation feature of MUST and adjacent words plays a greater role in classification, and the feature of point mutual information MI > 0 (the semantic correlation of the two words) mainly plays a generalization role.

(4) Syntactic features, the unique attribute features of the third type of semantic MUST, indicating the important role of syntactic features in the third type of semantic classification.

(5) Semantic features and syntactic features do not appear in classification rules, indicating that these features are redundant.

(6) Semantic features, syntactic features, pragmatic features, and topic features work together to form the meaning of modal verb MUST and its classification rules. Semantic features mainly appear in the rules of unique compound attributes of a class and play an important role. The unique attribute features consist of syntactic feature, pragmatic feature, and topic feature.

5. Example Verification

The study uses data from a public dataset based on a 1.5 million word multi-genre corpus. Corpus sources are shown in Table 2.

Taking the English modal verb MUST as the target word, this paper constructs the formal background of the relationship between the different meaning objects of MUST and the co-occurrence context features. From based on this background, the use of unique properties, characteristics calculation method to obtain MUST, present different meanings of simple unique attributes, such unique characteristics, and unique compound attribute characteristics, these characteristics as the meaning of classification rules, through the comparative analysis of the meaning of classification rules, found that modal verb MUST word meaning and the interaction between different contextual characteristics. The scatterplot method is adopted. The scatterplot of Ml (S + MUST) is shown in Figure 4.

As can be seen from Figure 4, if MI = 0 is chosen as the interval segmentation point, it can be well distinguished from the objects of MUST (obligation, responsibility/obligation/command) and MUST (inference, speculation). In addition, considering that Ml < 0 indicates that the meanings of two words are not correlated, and Ml > 0 indicates that the meanings of two words are correlated and whether the meanings of two words are correlated or not is of great significance in semantic analysis, the point mutual information value is divided into MI, 0 and MI > 0 intervals, and each interval is an attribute feature.

The influence of the proportion of labeled data on the method in two datasets is further studied. You can see that the average of $F1$ values at all levels is a good proxy for trends in other measures, which are still used here. Figure 5 plots the performance of the four methods and the two versions of

Table 2: Corpus sources.

Genre	Website
Law	lUtps://www.copyiglU.gov/
Literary fiction	lUtps://www.eastoftheweb.com
News reports	https://www.bbc.com/
Academic papers	hops://link.springer.com/
Popular science books	lUtps://www.nature.com/
Science fiction	lUtps://novel.tingroom.com/
Company introduction	lUtps://www.petrobras.com/
University said	https://www.upenn.edu/
Interview	https://transcripts.cnn.com/
Film subtitles	https://subscene.com/subtitles

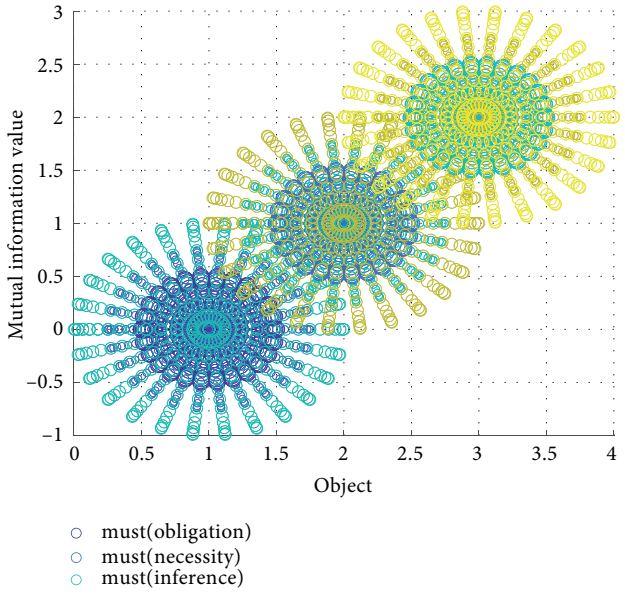

Figure 4: Scatter diagram of Ml (S + MUST).

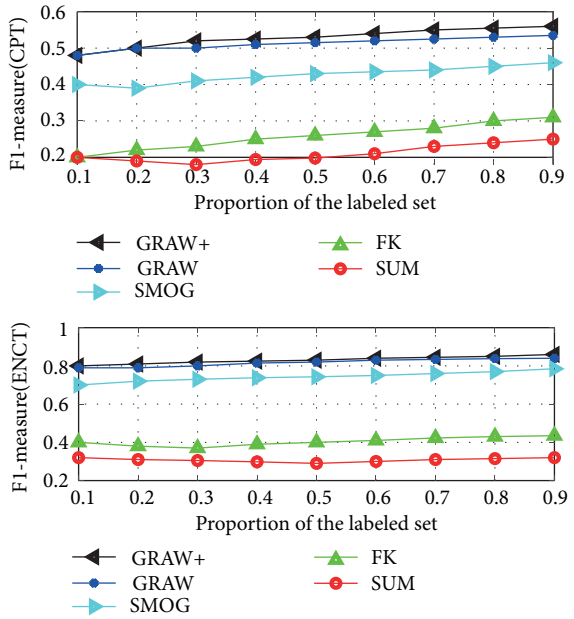

Figure 5: Average Fl values with a marker ratio varying between 0.1 and 0.9.

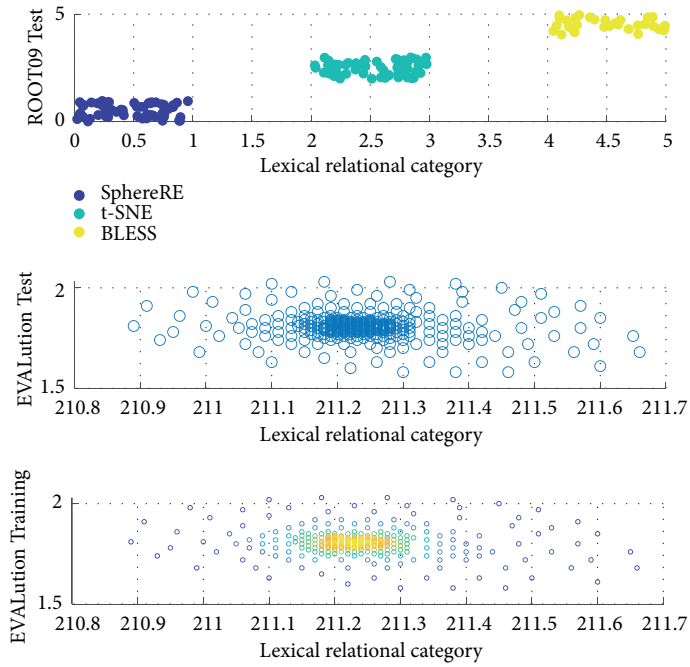

Figure 6: Visualized results of T-SNE algorithm for English language complexity text classification.

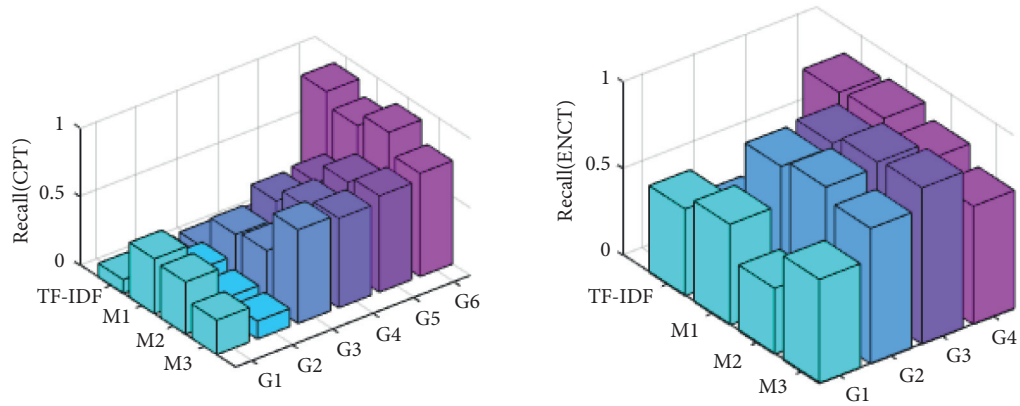

Figure 7: Comparison of recall rates between the coupled TF-IDF matrix and the basic TF-IDF matrix at various levels.

this working method (GRAW and GRAW+) when the marking ratio varies between 0.1 and 0.9.

As can be seen from Figure 5, neither SMOG nor FK benefits from the increasing marker data set. This indicates that the readability formula is difficult to improve by adding training sets. The other methods can achieve better results when the labeled data set is larger, and are still superior to the two formulas when the labeled ratio is small. Both CPT and V&M performed better than SUM on ENCT. GRAW outperformed all benchmark methods at any fraction of markers in both datasets, even at a low fraction of markers. As mentioned before, GRAW+ is improved in all scale ranges by adding new features and a new tag propagation algorithm.

We randomly sampled 300 prediction error cases for manual analysis. Most of the error cases can be attributed to the random relationship in the data set. For example, in K&H + N, BLESS, ROOT09, and CogALex, a large proportion of them are random relationships. There is no clear semantic relationship between these pairs of terms and the model is difficult to model, so, the classifier is likely to predict random relationships for tuples that have other lexical relationships. In addition, the imbalance of different types of data in the training set also makes it difficult to train the model. For example, the number of partial relations in the dataset evaluation and synonym relations in dataset CogALex is very small, and the learning effect of representation of corresponding relation tuples is also poor. For a more intuitive understanding of the SphereRE vector, the visualization of the SphereRE vector under the T-SNE algorithm is shown in Figure 6. For the training set, we can observe that the word vectors of different lexical relation categories have obvious separation in the two-dimensional plane. For the test set, the word vectors of the different term relation categories are slightly cluttered, indicating that the SphereRE algorithm is partially wrong in a model prediction.

As can be seen from Figure 7, on CPT, almost all unmarked documents are classified by the TF-IDF matrix to the first grade, which may be due to the low word frequency that makes it impossible to make meaningful discrimination between different levels. On ENCT, TF-IDF performs better but still tends to assign documents to lower grades.

As shown in Figure 7, the enhanced label propagation algorithm is superior to the general label propagation algorithm in both data sets, no matter which graph is used. This shows that the modification to the tag propagation algorithm works, and that the sequential relationship between reading levels does matter.

6. Conclusion

Four kinds of context features are extracted in this paper; semantic features, syntactic features, pragmatic features, and topic features work together to form the meaning and classification rules of the modal verb MUST, which all play an important role in the semantic classification of the modal verb MUST. Among the contextual features, pragmatic features and topic features directly limit the meaning of must. Topic features (topics related to natural laws) directly limit the meaning of MUST; The existing subject and perfect aspect directly limit the inferred meaning of MUST. Among the contextual features, pragmatic and topic-specific semantic and syntactic features play a stronger role in classification and have a more direct influence on the meaning of the modal verb MUST. The above research results fully reveal the interactive limiting relationship between modal verb MUST and contextual features, and provide a valuable basis for modal semantics, natural language processing, and intelligent semantic recognition feature selection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Liu, C. Xia, and H. Yan, "Hierarchical comprehensive context modeling for Chinese text classification," *IEEE Access*, vol. 7, pp. 134–157, 2019.

[2] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A review of text corpus-based tourism big data mining," *Applied Sciences*, vol. 9, no. 16, Article ID 3300, 2019.

[3] J. Berger, A. Humphreys, S. Ludwig, W. M. Wendy, N. Oded, and A. S. David, "Uniting the tribes: using text for marketing insight," *Journal of Marketing*, vol. 84, no. 1, pp. 21–25, 2020.

[4] M. Pejić Bach, Ž Krstić, S. Seljan, and T. Lejla, "Text mining for big data analysis in financial sector: a literature review," *Sustainability*, vol. 11, no. 5, Article ID 1277, 2019.

[5] Y. B. . Yong-Beom Kim, "Modal categories and dynamic modality in English," *Korean Journal of English Language and Linguistics*, vol. 17, no. 4, pp. 701–727, 2017.

[6] S. García, J. Luengo, and F. Herrera, "A data mining software package including data preparation and reduction: keel," *Intelligent Systems Reference Library*, vol. 72, pp. 285–313, 2015.

[7] K. Xie, G. Di Tosto, L. Lu, and Y. S. Cho, "Detecting leadership in peer-moderated online collaborative learning through text mining and social network analysis," *The Internet and Higher Education*, vol. 38, pp. 9–17, 2018.

[8] H. Han, G Liu, and J Dang, "An interactive model of target and context for aspect-level sentiment classification," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 3831809–3831829, 2019.

[9] H. Hassani, C. Beneki, S. Unger, and T. M. Maedeh, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 14, 2020.

[10] A. V. Glazkova, "Topical classification of text fragments accounting for their nearest context," *Automation and Remote Control*, vol. 81, no. 12, pp. 2262–2276, 2020.

[11] J. Kim, S. Jang, E. Park, and S. Choi, "Text classification using capsules," *Neurocomputing*, vol. 376, pp. 214–221, 2020.

[12] C. Källestål, E. Z. Blandón, and R. Peña, "Assessing the multiple dimensions of poverty. Data mining approaches to the 2004–14 health and demographic surveillance system in cuatro santos, Nicaragua," *Frontiers in Public Health*, vol. 7, pp. 409–423, 2020.

[13] D. Montoya, J. E. H. Vargas, J. S. Giraldo, and N. C. Hincapie, "Developing a pedagogical method to design interactive learning objects for teaching data mining," *Journal of Educators Online*, vol. 17, pp. 231–243, 2020.

[14] S. Fareri, G. Fantoni, F. Chiarello, and C. Elena, "Estimating Industry 4.0 impact on job profiles and skills using text mining," *Computers in Industry*, vol. 118, p. 103222, 2020.

[15] S. Deng, H. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," *Knowledge Discovery & Data Mining*, vol. 6, pp. 1007–1016, 2019.

[16] S. Wu, "Nonlinear information data mining based on time series for fractional differential operators," *Chaos*, vol. 29, no. 1, Article ID 013114, 2019.

[17] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems*, vol. 106, pp. 92–104, 2020.

[18] J. Hou and M. Zheng, "Online spatial evaluation of residential livability based on POI data mining and LMBP algorithm," *Arabian Journal of Geosciences*, vol. 14, no. 5, pp. 410–423, 2021.

[19] S. Minaee, N. Kalchbrenner, E. Cambria, N. Narjes, C. Meysam, and G. Jianfeng, "Deep learning--based text classification: a comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 34–40, 2021.

[20] S. Liu, X. Wang, C. Collins et al., "Bridging text visualization and mining: a task-driven survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2482–2504, 2018.

[21] R. A. Sinoara, J. Antunes, and S. O. Rezende, "Text mining and semantics: a systematic mapping study," *Journal of the Brazilian Computer Society*, vol. 23, no. 1, pp. 18–20, 2017.

[22] J. Yang, Y. Li, Q. Liu et al., "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57–69, 2020.

[23] G. Morota, R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando, "Big data analytics and precision animal agriculture symposium: machine learning and data mining advance predictive big data analysis in precision animal agriculture1," *Journal of Animal Science*, vol. 96, no. 4, pp. 1540–1550, 2018.

[24] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6844–6852, 2015.