Connection Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ccos20

Comparative relation mining of customer reviews based on a hybrid CSR method

Song Gao, Hongwei Wang, Yuanjun Zhu, Jiaqi Liu & Ou Tang

To cite this article: Song Gao, Hongwei Wang, Yuanjun Zhu, Jiaqi Liu & Ou Tang (2023) Comparative relation mining of customer reviews based on a hybrid CSR method, Connection Science, 35:1, 2251717, DOI: 10.1080/09540091.2023.2251717

To link to this article: https://doi.org/10.1080/09540091.2023.2251717

Published online: 06 Oct 2023.

CONNECTION SCIENCE
2023, VOL. 35, NO. 1, 2251717
https://doi.org/10.1080/09540091.2023.2251717

Comparative relation mining of customer reviews based on a hybrid CSR method

Song Gao[a,b], Hongwei Wang[c], Yuanjun Zhu[c], Jiaqi Liu[c] and Ou Tang[d]

[a]China Information Technology Security Evaluation Center, Beijing, People's Republic of China; [b]Chongqing Key Laboratory of Social Economic and Applied Statistics, Chongqing, People's Republic of China; [c]School of Economics & Management, Tongji University, Shanghai, People's Republic of China; [d]Division of Production Economics, Department of Management and Engineering, Linköping University, Linköpin, Sweden

ABSTRACT

Online reviews contain comparative opinions that reveal the competitive relationships of related products, help identify the competitiveness of products in the marketplace, and influence consumers' purchasing choices. The Class Sequence Rule (CSR) method, which is previously commonly used to identify the comparative relations of reviews, suffers from low recognition efficiency and inaccurate generation of rules. In this paper, we improve on the CSR method by proposing a hybrid CSR method, which utilises dependency relations and the part-of-speech to identify frequent sequence patterns in customer reviews, which can reduce manual intervention and reinforce sequence rules in the relation mining process. Such a method outperforms CSR and other CSR-based models with an F-value of 84.67%. In different experiments, we find that the method is characterised by less time-consuming and efficient in generating sequence patterns, as the dependency direction helps to reduce the sequence length. In addition, this method also performs well in implicit relation mining for extracting comparative information that lacks obvious rules. In this study, the optimal CSR method is applied to automatically capture the deeper features of comparative relations, thus improving the process of recognising explicit and implicit comparative relations.

ARTICLE HISTORY
Received 13 November 2022
Accepted 19 August 2023

KEYWORDS
Comparative relation mining; class sequence rule; dependency parsing; implicit comparative relation

1. Introduction

Comparison is usually used to understand the relation between two or multiple things. Generally, the comparison study is conducted with an aim of identifying aspects that the things hold in common, and in the meantime citing areas where the two things differ. For example, when a consumer searches for a product by entering keywords, an online store may recommend dozens of similar products correspondingly, then the consumer repeatedly compare the features of the products before they can make a final choice.

The customer review is a review of a product or service made by a customer who has purchased and used, or had experience with, the product or service. When shopping online, consumers often make shopping decisions by reading online reviews (Subhashini

CONTACT Song Gao hwwang@tongji.edu.cn; Hongwei Wang gaosongyouxiang@126.com Building A, Tongji Mansion, No. 1 Changwu Road, Yangpu District, Shanghai 200092, People's Republic of China

et al., 2021). Comparative reviews, on the other hand, are customer reviews that contain opinions indicating the strengths and weaknesses among products as experienced in one or more shopping cases. Such reviews become more compelling because the reviews are usually written by reviewers with extensive shopping experience (Gao et al., 2018). By mining and analyzing such comparison-rich relations in reviews, it can help consumers identify and compare products with the same attributes and provide them with evaluation benchmarks for judging whether a product is good or bad, thus helping them make better decisions and ultimately influencing their online shopping behaviour (Messaoudi et al., 2022; Wei et al., 2023).

Recent studies related to comparative reviews involves the recognition, extraction, and application of comparative relations in customer reviews: (1) comparative relation recognition mainly detects whether comparative opinions appear in reviews (Wang et al., 2017b); (2) comparative relation extraction is based on identifying comparative viewpoints, and further proves that product A is superior to product B in a certain attribute or vice versa; (3) the application of comparative relations mainly focuses on the extracted relations to analyze the ranking and competitiveness of products (Guo et al., 2022). Various methods such as text mining, pattern recognition, and machine learning are widely used with this research area (Bondarenko et al., 2022; Jain et al., 2022).

However, the identification and extraction of comparative relations remains challenging in some aspects, such as some customer reviews are difficult to process efficiently due to the presence of fragmented rules, or the lack of complete sentence semantics, or the presence of non-obvious comparative features (Gao et al., 2023). These problems are more common in unstructured texts such as customer reviews. For example, in the review "*ANDROID gt ios and in some aspects os gt Android*", it is difficult to understand the comparative relations in this review due to the incomplete sentence structure and the use of diminutive phrases (*gt* for greater). And those reviews that do not contain obvious comparative features are difficult to be identified as comparative reviews due to the lack of keyword localisation. Many scholars have tried to solve the above issues by extracting rules from comparative reviews through pattern recognition, sentence component analysis, deep learning, etc., but this leads to the problems of extensive rule generation, inefficient generation, and the need for more manual intervention (Vedula et al., 2023), which is not conducive to the further analysis and application of comparative relations. Thus, it becomes a challenging topic to propose an efficient and accurate method to recognise comparative relations in customer reviews.

In summary, this study aims to improve the efficiency of comparative rule generation, reduce manual intervention in the rule extraction process, and thus improve the effectiveness of comparative relation mining. To achieve this goal, we propose a framework based on hybrid Class Sequence Rule (CSR) models that aims to integrate rules by using CSR and dependency parsing. In this framework, comparative relations in customer reviews are categorised into two parts, namely explicit comparative relations (customer reviews with obvious comparison indicators) and implicit comparative relations (customer reviews lacking obvious comparison indicators). For the mining of explicit comparative relations, a new hybrid CSR model is used to improve the accuracy and efficiency in generating rules. As for the mining of implicit comparative relations, we adopt an "algorithm + strategy" solution to improve the whole recognition process by identifying whether there are two or more entity names in a customer review, and achieve satisfactory results. In this paper, we chose the more challenging Chinese text for experiments and model training, and our approach is also applicable to English text. The expected contributions of this research are as follows:

(1) This paper proposes a hybrid CSR approach that utilises dependency relation and part-of-speech to identify comparative relations in customer reviews, which can reduce manual intervention and reinforce sequence rules in the relation mining process.
(2) This new method is characterised by shorter computation time and improved efficiency of sequence pattern generation, as the dependency relation helps to shorten the sequence length.
(3) Regarding the implicit comparative relation, we propose a double recognition method based on hybrid CSR model and product named entity recognition. Based on the non-comparative reviews identified by the hybrid CSR method, we use the entity extraction technique to recognise them again and consider customer reviews containing more than two product naming entities as comparative reviews.

The paper is organised as follows. Section 2 reviews the related literature. Section 3 presents our comparative relation mining framework based on three hybrid CSR models. Section 4 conducts experiments using data from consumer reviews on jd.com. Section 5 summarises our study and discusses some future research directions.

2. Related studies

Comparative relation mining belongs to text categorisation tasks and is related to the fields of relation mining, dependency parsing, and competitiveness analysis (Liu et al., 2021b; Serrano-Guerrero et al., 2021; Wei et al., 2023). This paper is concerned with the related techniques of comparative relation mining, which mainly include three types of methods, namely pattern matching (PM), machine learning (ML) and natural language processing (NLP), which are discussed in this section.

2.1. Pattern matching

Pattern recognition methods are traditional techniques for comparative relation mining, which can be used for frequent sequence rule extraction when the syntactic structure in comparative reviews has fixed rules. Specific methods include CSR, feature extraction, semantic rules and so on. Pattern recognition generally involves learning from a large amount of annotated corpus to extract the fixed relationships between entities, such as keyword strategies, pattern libraries, etc. Related studies are shown in Table 1.

PM-based approaches have achieved excellent results in comparative relation mining. Among these approaches mentioned above, CSR model is the most used method for extracting sequence rules from the training corpus. For example, Wang et al. (2017a) combine artificial pattern libraries with CSR to identify the subdivision types such as equivalent relation and un-equivalent relation, respectively. Both Ping and Chen (2018) and Liu et al. (2022) fuse CSR, feature extraction and semantic rules to improve the recognition accuracy. However, there are some obvious drawbacks of pattern matching methods. First, the comparison keywords are significantly different in different contexts. Second, many of the rules proposed by pattern matching are incomplete and require human intervention. Third, it is

Table 1. Related studies on pattern matching.

Reference	Tasks	Methods	Data sources	Results
Wang et al. (2017a)	Identify the subdivided types of comparative reviews	CSR combined with an artificial pattern library	Dianping.com	Improve classification task, superior to baseline methods in terms of precision
Vo and Bagheri (2017)	Identify relational phrases in sentences and their arguments	Syntactic rule library	Three benchmark datasets: ReVerb, Wikipedia, and New York Times	Achieve higher recognition accuracy and more flexible relation generation
Ping and Chen (2018)	Propose an interactive system for visually exploring the semantic structure of a scientific article	Pattern matching, named entity recognition and supervised classifiers	Scientific papers	Assist users to understand the full and interesting story a scientific paper, with a concise outline and important details.
Kim and Kang (2018)	Identify the differentiating attributes of competing products	Keyword frequencies and ratios	Makeupalley.com	Extract comparison information efficiently by using ratio analysis of words and latent semantic analysis
Bondarenko et al. (2020)	Identification and classification of comparative questions	Lexical and syntactic rule library	Yandex and Otvety	Improve identification of comparison questions
Yang et al. (2020)	Improve the performance of document classification	Semantic similarity and correlation analysis	Chinese document datasets	Improve the accuracy of semantic document classification for texts in Chinese
Iso et al. (2021)	Generate two contrastive summaries and one common summary from two different candidate sets of reviews	Characterise the target summary distribution by leveraging two base summarisation models	TripAdvisor reviews	Generate higher quality product comparisons and common summaries from reviews
Liu et al. (2022)	Compare the functions from the description for App	Feature extraction and cluster	App description information on Google Play	Obtain the features that can accurately describe App functions in an automated way

difficult to recognise valid rules for comments that do not contain obvious comparison indicators. As a result, some studies are devoted to mining frequent sequence rules by deeply analyzing semantic information, proposing pattern matching methods based on relational lexicon, semantic similarity or machine learning (Iso et al., 2021; Yang et al., 2020). All of the above studies try to achieve very high recognition accuracy, but there is still room for improvement in generating efficiency and solving the human intervention.

2.2. *Machine learning*

In recent research, many scholars have combined comparative relation mining with machine learning, using Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes (NB), Conditional Random Fields (CRF), Logistic Regression (LR), and Random

Forests (RF) to identify comparative relations in text (Messaoudi et al., 2022; Sagnika et al., 2021) (Table 2).

ML-based approaches can automatically analyze the textual features of customer reviews and identify potential comparative relations using computer technology. Liu et al. (2021a), Liu et al. (2021c), and Gao et al. (2023) have proposed deep learning models for comparative relation mining based on the ANN architecture in order to improve the recognition accuracy and stability. Wei et al. (2022) also use a deep learning approach to alleviate the dependence on comparison keywords and enhance the generalisation ability of comparative relation mining. Above studies achieve better results at the sentence-level, but there is scope for improving the analysis and processing of the whole customer reviews. Moreover, the effectiveness of machine learning-based comparative relation mining methods depends largely on the combination and optimisation of comparison features, which is an unexplained process.

2.3. Natural language processing

Natural language processing is a key technology for studying customer reviews, and some classical NLP models are often used to identify comparison rules, extract comparison elements and analyze product competitiveness (Gao et al., 2018; Liu et al., 2019). The combination of comparative relation mining and NLP techniques has led to the expansion of research into a variety of domains such as online Q&A, forums, social media, etc., where a large amount of comparative information exists (Alhamzeh et al., 2021; Bondarenko et al., 2022). For example, Wang et al. (2021) perform sentence clustering by calculating the similarity of technical description statements to analyze the typical characteristics of the technology that users focus on and to demonstrate the advantages of the technology. Guo et al. (2022) propose a brand joint latent Dirichlet allocation (LDA) model for analyzing general aspects of multi-brand customer reviews and specific aspects of user opinions within a single brand. The main drawback of NLP-based approaches is that they cannot recognise comparative relations independently and thus need to be combined with methods such as PM and ML (Table 3).

2.4. Research gaps

In conclusion, prior studies have advanced the research progress of comparative relation mining, but still face some urgent problems. First, the PM-based or NLP-based comparative relation mining methods alone lead to poor results due to the lack of the ability to identify important rules from numerous cluttered sequences, which requires fusion with more feature elements to achieve better recognition results. Second, the comparison rules extracted from comparative relations have industry attributes, so the generalisation effect on large-scale datasets containing multiple domains often fails to meet the requirements of practical applications. Third, the methods proposed in the above studies are less efficient in generating rules despite good recognition accuracy, which limits the application of comparative relation mining in business domains.

In contrast, customer reviews are unstructured-free-text in natural language form with misspellings and Internet terms, which can be expressed in Chinese, English or other languages. The complex form increases the difficulty of comparative relation mining. As shown

Table 2. Related studies on machine learning.

Reference	Tasks	Methods	Data sources	Results
Tkachenko and Lauw (2017)	Identify comparative relations	SVM and NB	Reviews from the Digital Camera and Cell Phone categories of Amazon.	Achieve good performance in sentence-level comparative relation mining task
Fei et al. (2020)	Extract overlapping entity relations	ANN	ACE05, CoNLL04, and NYT	Improve accuracy of entity identification
Vo et al. (2020)	Identify temporal and causal relations from the textual passage	SVM	TempEval-3 and Causal-TimeBank corpora	Achieve good performance at recognising temporal and causal relations
Khan et al. (2020)	Investigate different algorithms for classification of comparative reviews	LR, SVM, NB, RF, etc.	Reviews from YouTube	Random forest outperforms other classifiers in this study
Liu et al. (2021a)	Identify comparative relations	Document Representation Fusion with Attention (ANN)	Product reviews on E-commerce platforms	Achieve good results in terms of performance and robustness
Liu et al. (2021c)	Identify and extract comparative relations	BERT-based multi-stage approach (ANN)	Product reviews from three datasets	ANN method outperforms baseline systems extended from traditional methods
Wei et al. (2022)	Identify comparative relations	deep learning method and Data Augmentation technology	Truck data from 360che.com and mobile phone data from JD.com	Present superior performance compared to traditional textual data augmentation methods
Gao et al. (2023)	Identify comparative relations	a hierarchical multi-attention network model (ANN)	Mobile phones reviews on JD.com	This model outperforms both traditional classification models and text classification models, and has a good performance on extracting comparative relations from long texts

Table 3. Related studies on natural language processing.

Reference	Tasks	Methods	Data sources	Results
Gao et al. (2018)	Identify competitors through comparative relation mining	NLP-based framework	Consumer reviews in Dianping.com	Competitors can be identified and market conditions can be assessed by mining reviews for comparative relations
Liu et al. (2019)	Evaluate product competitive advantage from user-generated content on social media	NLP-based framework	User-generated content from autohome.com.cn and bitauto.com	User-generated content provides an assessment of a product's competitive advantage
Alhamzeh et al. (2021)	Answer comparative questions	BERT-based model	ClueWeb12 document dataset	Present an advanced web search engine to answer comparative questions
Wang et al. (2021)	Identify and cluster descriptive texts of similar technologies	NLP-based framework	Text data from Stack Overflow	This framework covers a large set of comparable technologies and their corresponding comparative sentences with high accuracy
Beloucif et al. (2022)	Answer comparative questions	Neural-based model using ALBERT embeddings	Text data from Yahoo! Answers, Reddit, and Quora.	Supervised learning can reliably find comparative sentences and, less reliably, objects and aspects comparisons
Bondarenko et al. (2022)	Identifying the comparative questions and answers	NLP-based solution	MS MARCO, Quora Yahoo and Google Natural Questions datasets	Achieve better results in recognising comparative questions and answers
Guo et al. (2022)	Identify the strengths and weaknesses of several competitive brands	a brand joint latent Dirichlet allocation (LDA) model	Customer reviews from e-commerce platforms	Comparative relations in customer reviews can be used to identify product strengths and weaknesses

Table 4. Examples of the loose structure and oral style in customer reviews.

Product named entity	Review examples
HUAWEI nova	但是最坑的是数据线接口和其他安卓系统手机不一样。 (*The most deceptive thing, however, is that the cable interface is different from other Android phones.*)
360n5	照相感觉还没有我老 g3 好。 (*Photography doesn't feel as good as my old g3.*)

in Table 4, loose structure and colloquialisms are more common in customer reviews. In the sentence "*The most deceptive thing, however, is that the cable interface is different from other Android phones*" (但是最坑的是数据线接口和其他安卓系统手机不一样), "*the most deceptive thing*" (最坑的是) is used to introduce the subordinate clause, i.e. "*the data cable interface is different from other Android phones*" (数据线接口和其他安卓系统手机不一样), and the comparative relation is implied in the subordinate clause, thus making identification more difficult. Another example of colloquialism is that "*photography does not feel as good as my old g3*" (照相感觉还没有我老 g3 好), the normal written language could be "*It seems like that photography is not as good as my previous g3*" (感觉照相还没有我以前的 g3 好).

CSR model has proven to be an effective technique in extracting frequent sequence rules in comparative relations (Wei et al., 2023). To overcome the shortcomings of the above work, we aim to explore an improved CSR model to achieve a balance between efficiency and effectiveness in comparative relation mining. In prior studies, most of the sequence rules generated using CSR models are based on POS tagging and manual tagging to identify key phrases, and fewer scholars used syntactic or semantic analysis to generate sequence rules. And dependency parsing is a widely used semantic analysis method to understand the interdependencies between sentence constituents (Jain et al., 2023), which provides technical support based on the semantic level for understanding comparative relations in customer reviews. Therefore, we attempt to merge dependency parsing with CSR modelling and propose a hybrid CSR approach for identifying comparative relations.

In addition, the identification of implicit comparative relations, a task that requires in-depth analysis of the semantics of customer reviews, has been neglected in the current research and has become a challenging task due to the lack of frequent sequence rules. Table 5 gives examples of typical explicit and implicit comparative relations. Although missing an obvious comparison indicator, implicit comparison relations tend to appear in two consecutive clauses which express opinions on the same attribute of different products, and such relation mining can be achieved through semantics and sentence structure. To sum up, the aim of this paper is to propose a new method that integrates the efficiency and effectiveness of comparative relation mining. Specifically, it is to optimise the strategy of sequence rule generation based on CSR model and dependency parsing technique, and to enhance the whole recognition process by using an "algorithm + strategy" solution, which is used to identify whether there are two or more entity names in a customer review.

3. Research design

The research framework is summarised in Figure 1. First, we preprocess the annotated corpus, including removing special symbols, correcting misspellings, word splitting and

Table 5. Explicit and implicit comparative relations.

Categories	Customer reviews
Explicit comparative relation	比起华为、oppo、魅族，在相同运存、内存及处理器上这是是很划算的一款手机哦。 *(Compared with HUAWEI, OPPO, and Meizu, this is a cost-effective mobile phone with the same storage, memory, and processor.)*
Implicit comparative relation	之前 360n4s 5000 毫安能用两天，现在 4000 毫安一天都用不了。 *(Regarding 360n4s, previously the fully charged battery of 5000 mAh had the power for two days' use, while now the battery of 4000 mAh only provides the power for less one days' use.)*

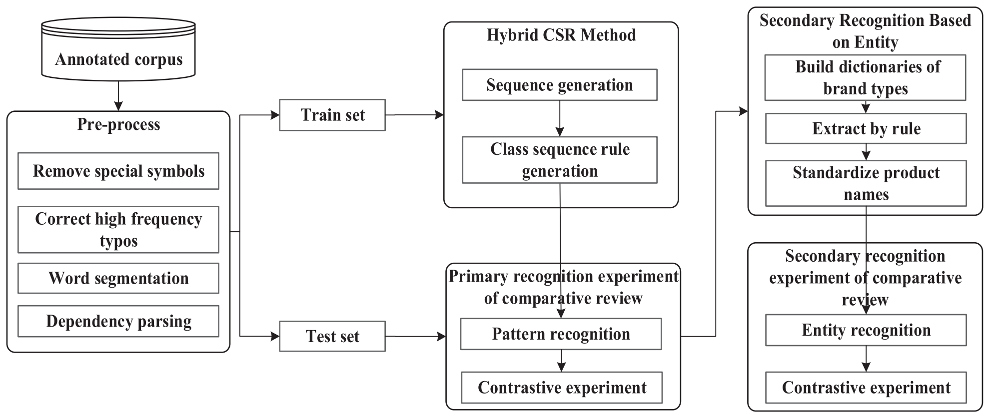

Figure 1. Research framework.

dependency parsing. Second, the annotated corpus is partitioned into a training set and a test set. Third, a hybrid CSR approach is used to identify comparative relations: (a) for sequence generation, we propose three improved CSR models, details of which will be presented in Section 3.2; (b) based on the generated sequences, we extract frequent CSRs using the prefixspan algorithm; (c) the generated rules are applied to the test set. If any frequent CSR occurs in a customer review, the corresponding feature position is marked as 1, otherwise it is marked as 0. Thus, if the sum of the elements of the feature vector is not equal to 0, it means that this review has a comparative relation; (d) for non-comparative relations classified by the hybrid CSR approach, we use information extraction techniques to recognise product named entities. This process includes building a dictionary of brand types, extracting entities based on the rules of the product structure tree, and standardising product brands.

3.1. *The hybrid CSR method*

CSR generally consists of comparison words, part-of-speech of keywords, and adjacent words. We generate sequences centred on the comparison words with a radius of three participles, and then use the PrefixSpan algorithm to generate CSRs, in which the steps for generating frequent sequences can be referred to the research results of Anwar and Uma (2022) and Noorian Avval and Harounabadi (2023). Table 6 illustrates an example where sentence 1 generates sequence 1 and sentence 2 generates sequence 2.

Table 6. Sequences generated by traditional CSR model.

Sentence one	但是 wifi 信号不如以前的华为信号好。 *(But the wifi signal is inferior to previous Huawei.)*
Sequence two	c, ws, n, 不如 (inferior)/v, nd, u, nz
Sentence two	360 跟乐视新出的手机一样。 *(The 360 is the same as the new mobile phone of LeTV.)*
Sequence two	n，跟 (as)/*p*，n，d，v，u，一样 (the same)/u.

(The notations are explained as follows: c – conjunction, ws – non-character symbols, n – noun, v – verb, nd – direction noun, u – auxiliary word, nz – other proper noun, *p* – preposition, d – adverb).

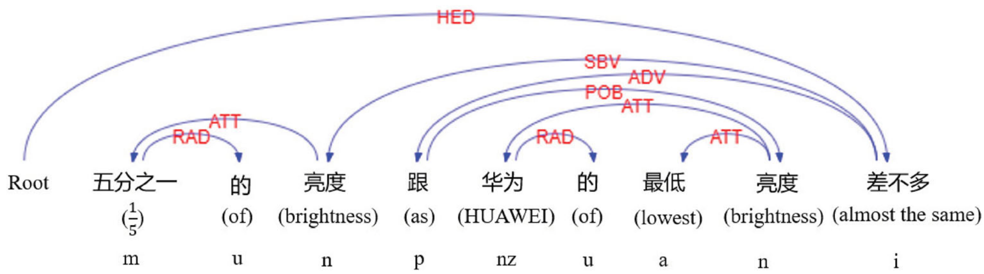

Figure 2. An example of dependency parsing.

Dependency parsing involves exploring the dependencies between words in a sentence to gain an understanding of its grammatical structure. In computer science, the dependency relation is a binary structure, including core words, dependency words, and relation between them. Figure 2 below presents an example of dependency parsing. The dependency word is dominated by a core word. The arc originates from a core word and points to a dependency word, in order to demonstrate the dependency relation between two words. So the dependency relation can be written as $\{(w_1, relation_1, head_1), (w_2, relation_2, head_2), \ldots (w_i, relation_i, head_i)\}$, where w_i represents a word, $relation_i$ represents a dependency relation component, and $head_i$ represents the position of one word pointed by dependency relation in a sentence. Thus, the sentence in Figure 2 can be expressed as {(五分之一(1/5), ATT(attribute), 3), (的(*of*), RAD(right adjunct), 1), (亮 度(*brightness*), SBV(subject-verb), 9), (跟(*as*), ADV(adverbial), 9), (华为(*HUAWEI*), ATT(attribute), 8), (的(*of*), RAD(right adjunct), 5), (最低(*lowest*), ATT(attribute), 8), (亮度(*brightness*), POB(proposition-object), 4), (差不多(*almost the same*), HED(head), 0)}. The special dependency relation component is HED, which represents the central word of the whole sentence. In the subsequent methods, the dependency component and dependency direction will be integrated into the sequence rules.

In this paper, the hybrid CSR is mainly used in the explicit comparative relation mining process. This process usually consists of three main steps: sequence generation, CSR generation, and pattern matching. As mentioned above, we use the classical PrefixSpan algorithm to generate CSRs (Anwar & Uma, 2022), and complete pattern matching by CSR matching. Regarding the sequence generation, it is not only the basis of CSR generation and pattern matching, but also a key step in determining recognition accuracy. Therefore, we explore and propose three hybrid CSR Methods to ensure the performance.

Table 7. Sequence generated by CSR_N2DP model.

Sentence one	{(但是(*but*), ADV, 4), (wifi， ATT， 3), (信号(*signal*)， SBV， 4), (不如(*inferior*)， HED， 0), (以前(*previous*)， ATT， 8), (的(*of*)， RAD， 5), (华为(*HUAWEI*), ATT, 8), (信号(*signal*)， SBV， 9), (好(*good*)， VOB， 4)}
Sequence one	ADV, ATT, SBV, inferior/HED, ATT, RAD, ATT
Sentence two	{ (360， ATT， 7), (跟(*as*)， ADV， 5), (乐视(*LeTV*)， POB， 2), (新(*new*), ADV, 5), (出(*outcoming*)， ATT， 7), (的(*of*), RAD, 5), (手机(*mobile phone*), HED, 0), (一样(*the same*)， RAD， 7)}
Sequence two	ATT, 跟(*as*)/ADV, POB, ADV, ATT, RAD, HED, 一样(*the same*) /RAD

3.1.1. The CSR_N2DP model

Based on the traditional CSR model, we propose the CSR_N2DP model, which does not consider the dependency direction and replaces the part-of-speech with only the dependent components. The model is implemented in the following steps. First, the model generates sequences from a sentence through a fixed window. Subsequently, the comparison keyword is used as the central word and the dependency relationship of the central word in the sentence is analyzed. Then, the comparison keyword and its dependency components are extracted. And finally, the dependency components of the N-words neighbouring the comparison keyword are refined into a new sequence. In the Table 7, we take $N = 3$ as an example, in which *Sentence one* takes the comparison keyword “inferior” and its dependency component “HED” as the centre component, then the dependent components of the three words before and after the comparison keyword are extracted in an orderly manner, and finally *Sequence one* is generated. The sequence generation process of *Sentence two* is consistent with *Sentence one*. It is worth noting that the sentence contains two comparison keywords.

3.1.2. The CSR_DP model

In the subsequent experiments, we find that the fixed-window approach used by CSR_N2DP truncates the entire sentence, thus disrupting the coherence of sentence components. To ameliorate this problem, we propose the CSR_DP model, which combines the dependency components and dependency directions. The main advantage of this model is that instead of using the fixed-window approach to generate a finite number of sequences, it extracts sequences using word-to-word dependency directions in dependency parsing. The implementation steps are as follows. Firstly, this model obtains each component and its dependency direction based on the syntactic parsing of a sentence. Then the comparison keywords and their dependency components are extracted. Finally, sequence generation is realised by the following rules: if the comparison keyword is a core component labelled as HED, it is necessary to extract all the dependency constituents that are closed to and dependent on the keyword; if the comparison keyword belongs to a non-core component, it obtains the dependency constituents of its dependency words to determine whether the component is a core constituent or not; if the component is not yet a core constituent, it turns to obtain the dependency constituents of another dependency word until the core component is determined.

As shown in Table 8, the comparison keyword “不如(*inferior*)” in *Sentence one* is identified as the core component (“HED”), so we need to acquire the words that are closest and depend on it. Since this keyword is the fourth component in this sentence, we extract the words whose dependency are labelled as “four”. Thus, the CSR_DP model generates

Table 8. Sequence generated by CSR_DP model.

Sentence one	{(但是(*but*)， ADV， 4), (wifi, ATT, 3), (信号(*signal*)， SBV， 4), (不如(*inferior*)， HED， 0), (以前(*previous*), ATT, 8), (的(*of*), RAD, 5), (华为(*HUAWEI*), ATT, 8), (信号(*signal*), SBV, 9), (好(*good*), VOB, 4)}
Sequence one	ADV, SBV, 不如(*inferior*)/HED, VOB
Sentence two	{ (360, ATT, 7), (跟(*as*), ADV, 5), (乐视(*LeTV*), POB, 2), (新(*new*), ADV, 5), (出(*outcoming*), ATT, 7), (的(*of*), RAD, 5), (手机(*mobile phone*), HED, 0), (一样(*the same*), RAD, 7)}
Sequence two	跟(*as*)/ADV, ATT, HED, 一样(*the same*)/RAD

Table 9. Sequence generated by CSR_HH model.

Sentence one	{(但是(*but*), ADV, 4), (wifi, ATT, 3), (信号(*signal*), SBV, 4), (不如(*inferior*), HED, 0), (以前(*previous*), ATT, 8), (的(*of*), RAD, 5), (华为(*HUAWEI*), ATT, 8), (信号(*signal*), SBV, 9), (好(*good*), VOB, 4)}
Sequence one	ADV, SBV, 不如(*inferior*)/c, VOB
Sentence two	{ (360, ATT, 7), (跟(*as*), ADV, 5), (乐视(*LeTV*), POB, 2), (新(*new*), ADV, 5), (出(*outcoming*), ATT, 7), (的(*of*), RAD, 5), (手机(*mobile phone*), HED, 0), (一样(*the same*), RAD, 7)}
Sequence two	跟(*as*)/p, ATT, HED, 一样(*the same*) /u

Sequence one for *Sentence one*, which includes the core component "不如(*inferior*)/HED" and dependency components "但是(*but*)/ADV", "信号(*signal*)/SBV" and "好(*good*)/VOB". *Sentence two* is generated in the same order as *Sentence one* and applies the non-core constituent rules described above.

3.1.3. The CSR_HH model

Although the CSR_DP model utilises dependency components and dependency directions in syntactic analysis, it is relatively difficult to mine comparative relations in customer reviews. This is because the complex and variable unstructured text data reduces the accuracy of dependency parsing. For example, "跟(*as*)" corresponds to only "/p" in the part-of-speech, but its corresponding dependency components are numerous. The sequence pattern associated with this word hardly satisfies the minimum support and minimum confidence requirements, and is therefore defined as an infrequent sequence pattern. In addition, it is usually possible to generate some strong sequence rules by recognising the comparison keywords. At this point, if the comparison keywords are combined with their dependency components, the originally frequent strong sequence rules will be split into several weak rules. To overcome this problem, we propose the CSR_HH model, which enhances the generation of frequent sequences by utilising dependencies and parts-of-speech. Instead of considering the dependency component of the comparison keyword, the model emphasises the role of discourse in reinforcing sequence patterns.

The implementation steps of CSR_HH model are similar to those of CSR_DP model, with the difference that we replace the dependency components corresponding to the comparison keywords with their parts of speech during the sequence generation process. Thus, it is an approach that integrates dependencies and parts-of-speech. As shown in Table 9, we take *Sentence one* as an example. Based on the implementation of the CSR_DP model, this method replaces the dependency component of the comparison keyword "不如(*inferior*)" in *Sequence one* with the part-of-speech "c".

In an attempt to provide a clearer picture of the differences between the three models, this paper compares the characteristics, implementation processes, and shortcomings of the three models, as shown in Table 10.

3.2. The entity-based double recognition method

In the annotated corpus of this paper, customer reviews with implicit comparative relations represent about 4% of the total number of reviews. These reviews are typically characterised by the lack of obvious comparison keywords or comparison patterns in the reviews, but usually contain product named entities. Thereby, this study argues that implicit comparative relations can be identified by mining comparable entities in customer reviews. If there are two different entities in the non-review text, their semantic relationships usually include causal, inheritance, juxtaposition, progression, transitive, and comparative relation, etc. However, if two or more product named entities appear in the same review, the review text usually reflects a comparative relation (Wang et al., 2017a). The sampling statistics of the corpus has shown that 91.23% of the reviews containing two or more product named entities are labelled as comparative reviews, so this result corroborates the reasonableness of the above findings.

Since the product named entities in customer reviews are hierarchical in nature, the product name structure tree can be used to find out the naming rules. As shown in Figure 3, the entity structure tree contains four layers: product layer, series layer, model layer and attribute layer. Implicit comparison relation mining mainly focuses on the first three layers. The experimental data of this paper includes 32 well-known cell phone brands, each brand contains 4–8 major series, and each series contains multiple models. Therefore, it is necessary to collect and build some dictionaries involving product brands, series, and models. A non-manual approach is needed to solve this problem due to the short life cycle of smartphone products, rapid updates, and relatively large number and complexity of product models. On this basis, we hypothesise that product named entity recognition can be transformed into a hierarchical annotation task, i.e. recognising brand, series and model names separately and combining them to form detailed product named entities.

On the basis of the above analysis, this paper proposes a double recognition method to identify comparative relation by extracting product named entities, as shown in Figure 4. A non-comparative review after the first recognition classification will be classified as comparative reviews if two different product named entities appear in the same reviews. The process consists of three steps: constructing the dictionaries of brands, series, and models; extracting named entities according to the rules; standardising product named entities. The following is the implementation process.

First of all, different methods should be used to construct brand dictionaries, series dictionaries and model dictionaries respectively. Since there are relatively few brand and series names, the dictionaries can be constructed by manual collection methods. For the model dictionary, this paper proposes a similarity-based method for constructing the model dictionary, i.e. a part of the model names are collected as the seed dictionary and the seed dictionary is extended using the similarity calculation technique based on the edit distance algorithm. The edit distance algorithm reflects the absolute difference between two strings. However, if the lengths of the strings are different, the measurement for the absolute difference may not be precise enough. The value in the interval [0,1] reflects the similarity of

Table 10. The comparison of three hybrid CSR models.

Models	CSR_N2DP	CSR_DP	CSR_HH
Features	(1) replacing the part-of-speech with dependency components of dependency Parsing, (2) excluding the dependency direction.	(1) combining the dependency component and dependency direction, (2) extracting the sequence by using the dependency direction between words in the dependency parsing.	(1) identifying frequent sequence patterns with the dependency relation and the part-of-speech, (2) **(2)** reinforcing the sequence rules based on the part-of-speech, regardless of the dependency components of comparison words.
Execution process	(1) generating sequences from one sentence through a fixed window, (2) taking the comparison keyword as the centre word, (3) extracting the comparison word and its dependency components, (4) distilling the dependency components of the N-words adjacent to the comparison keyword into a new sequence.	(1) acquiring each component and its dependency direction based on the syntactic parsing of one sentence., (2) extracting the comparison keyword and its dependency components, (3) determining that whether the comparison word is the core component and performing different extraction rules.	on the basis of the implementation of the CSR_DP model, replacing the dependency component of the comparison keyword with its part-of-speech.
Defects	the fixed-window approach makes the entire sentence truncated, thus disrupts the coherence of sentence components.	(1) the complicated and volatile unstructured text data reduce the accuracy of dependency parsing, (2) combining the dependency components and comparison words, the original frequent rules may be split into several weak rules.	

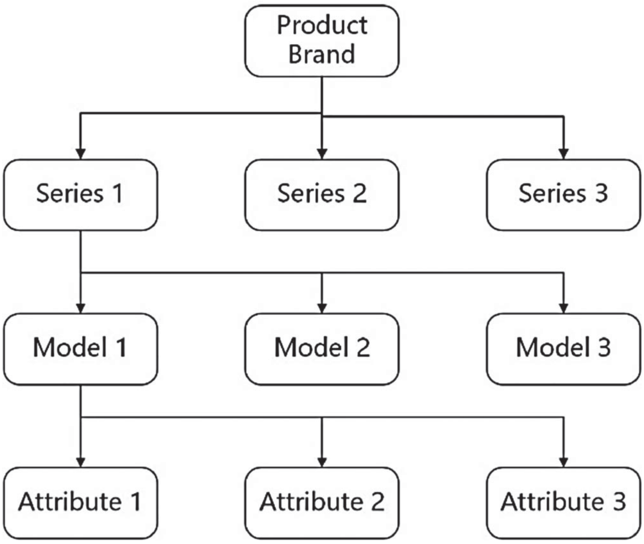

Figure 3. An illustration of product name structure tree.

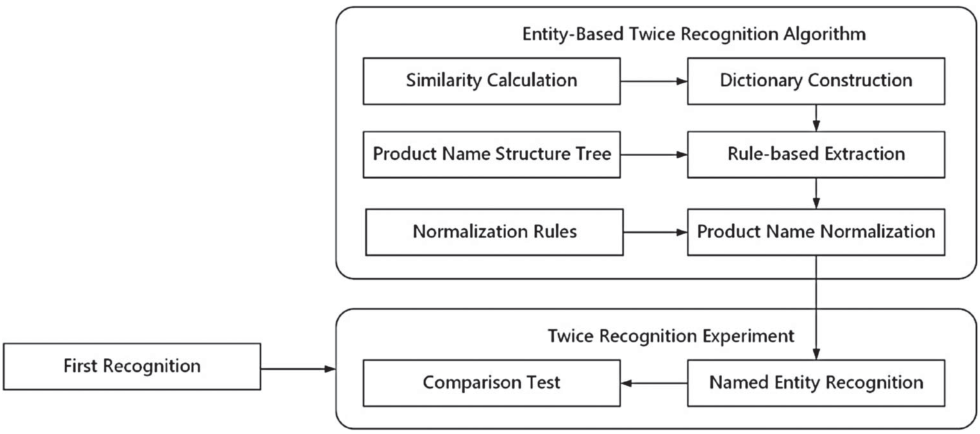

Figure 4. Entity-based double recognition method.

two strings. According to the definition of edit-distance, dist(a, b) refers to the edit-distance between strings a and b, and $sum_{a,b}$ refers to the sum of the lengths of strings a and b. Therefore, the formula for similarity is:

$$similarity_{a,b} = (sum_{a,b} - dist(a, b))/sum_{a,b} \tag{1}$$

Next, the corpus is automatically labelled with the help of the constructed dictionary of brands, series and models. We can extract entities according to the rules of entity structure tree based on the hierarchical structure of brands, series and models. Besides, in the sample corpus, sometimes some product named entities involve a mixture of brand, model and series. Thus, in this paper, we design the extraction rules (as shown in Table 11) and perform multiple rounds of matching extraction to complete the extraction of entity names.

Table 11. Extraction rules.

Elements	Extraction rules
Brand	Brand
	Brand + Series
	Brand + Series + Model
	Brand + Model
Series	Series + Model
	Series
Model	Model

Table 12. Name normalisation rules.

	Rule	Example
Rule 1	Replace the default product brand with the full product named entity for the same model in the same review	{Xiaomi 5s,5s}— > {Xiaomi 5s, Xiaomi 5s}
Rule 2	Replace the common abbreviations with full product named entity	6s plus— > iphone6s plus
Rule 3	Add the brand and series of the currently reviewed product before the incomplete entity	n4— > 360n4

Finally, due to the colloquial nature of customer reviews, product named entities are often abbreviated, e.g. 360n4 is written as n4. Also, a product may be expressed differently in different reviews. We believe that this is mainly due to the semantic omission of product named entities resulting in missing series and model information (only "brand + series" or "brand"). In this case, it is difficult to recognise specific product models. As a result, in this paper, we retain the original product named entities, and for the case of missing brands or series, we use the rules in Table 12 to realise the replacement of product named entities. The entity extraction technique is a favourable method to realise the implicit comparative relation judgment. If there exists a product named entity in a review that is different from the one being reviewed, and the distance between these two entities in the product name structure tree is large, it is determined that there is a comparative relation in the review, otherwise no such relation exists.

4. Experiments

This section aims to evaluate the performance of the proposed methods by conducting experiments from different perspectives, all data processing is completed through python programming. Details are described below.

4.1. Experimental data

The experimental data in this paper all come from Chinese customer reviews on the jd.com website. This site, well known as Jingdong, is one of the top e-commerce online shopping platforms in China, owned by Chinese company JD.com Inc. We collect a large number of Chinese customer reviews of different products, and then determine the experimental dataset according to two principles: (a) we exclude customer reviews that are automatically generated by the system or written by consumers in 10 words or less (e.g. "very good" or "very good, good service"), because they have little value to the study; (b) we select

Table 13. Overview of the experimental data.

		Quantity	Ratio
Comparative review	Explicit relation	3670	18.43%
	Implicit relation	769	3.86%
Non-comparative review		15470	77.70%

Table 14. Example of comparative words and collocations.

Comparison keywords	Comparison collocations
Than (比/n)	Make no different (没什么/i … 差别/n)
Not as good as (比不上/i)	Different from (跟/*p* … 不同/a)
Compared with (比起/v)	The same (跟/*p* … 一样/u)
Inferior to (不如/v)	Compared with (跟/*p* … .相比/v)
Be roughly the same (不相上下/i)	Make a comparison (和/*p* … .比较/n)
…	…
Beyond (超出/v)	Difference from (与/*p* … 差别/n)

mobile phones as research objects because they are relatively popular in people's lives, and customers often express their opinions in reviews by comparing the advantages and disadvantages of different brands of mobile phone products in terms of price, weight, battery usage, etc.

Thus, we firstly select 23 mobile phone products from 10 brands and randomly select 1000 reviews for each product. We next remove duplicates and delete meaningless or questionable comments, and subsequently obtain19, 909 reviews as the experimental data. After text preprocessing and manual recognition, comparative and non-comparative reviews are labelled and obtained separately, as shown in Table 13. The number of comparative reviews is 4439, accounting for 22.3% of the entire corpus, of which 18.43% are reviews with explicit comparative relations and 3.86% are reviews with implicit comparative relations.

4.2. Comparative relation mining with the hybrid CSR methods

4.2.1. Model selection

We compare the traditional CSR model with the three hybrid CSR models proposed in this paper. The abbreviations and meanings of these four models are summarised as follows: CSR for the traditional CSR model; CSR_N2DP for the model of replacing part-of-speech with dependency parsing in CSR; CSR_DP for the model of combining dependency components and directions; CSR_HH for the model of integrating dependency relation and part-of-speech.

Since the above four models all require the use of comparison keywords and collocations, the Chinese comparison lexicon of this paper is finally constructed based on the manual collection of comparison words and collocations, and extensive reference to the comparison words in other corpora. Table 14 lists a part of typical comparison keywords and collocations respectively.

The performance is measured by three metrics: P (precision), R (recall) and F1-measure, which are defined as:

$$\text{Precision}: \ P = \frac{T_c}{F_c + T_c} \tag{2}$$

$$\text{Recall}: \ R = \frac{T_c}{F_n + T_c} \tag{3}$$

$$\text{Overall Accuracy}: \ A = \frac{T_c + T_n}{T_c + T_{n+}F_{n+}F_c} \tag{4}$$

$$\text{F1} - \text{measure}: \ F = \frac{2P^{*}R}{P + R} \tag{5}$$

F_n is defined as the positive sample is judged as negative, whereas F_c is the case that the negative sample is judged as positive. And T_c indicates the case that the positive sample is judged as positive, T_n represents the case that the negative sample is judged as negative.

4.2.2. Parameters setting

In this study, we refer to Huang et al.'s (2008) support design to identify frequent CSRs. The formula for the minimum support is as follows, where $freq(a_{ij})$ represents the number of occurrences of the *jth* item of rule r_i, a_{ij}, in the sequence dataset D, $|D|$ indicates the size of the sequence dataset, λ denotes a threshold value between 0 and 1, ε is the support threshold and $\varepsilon > 1/|D|$.

$$\text{support}(\mathrm{r}_i) \geq \max\left(\frac{\lambda}{|D|}\min\{freq(a_{ij}\}, \varepsilon\right) \tag{6}$$

In the experiments, we follow the recommendation of Huang et al. (2008) and assign $\varepsilon = 1/|D|$, i.e. the rule occurs at least once. In both CSR model and CSR_N2DP model, the window radius used is set to 4, and the confidence threshold is assigned as 0.6. This study examines the performance of generating sequence rules when λ varies in the range of [0.1, 0.2, 0.3, 0.4, 0.5]. When $\lambda = 0.1$, each model generates the most sequence patterns, but the number of CSRs shrinks rapidly as λ becomes larger. When λ is set to 0.5, the number of rules generated for the CSR, CSR_DP, CSR_N2DP, CSR_HH models are 405, 1095, 1035, 448 respectively.

4.2.3. Model comparison

In this section, we compare the recognition effectiveness and efficiency for generating sequence rules of different hybrid CSR models in comparative relation mining. Firstly, as shown in Figure 5, the computation time for generating sequence rules is used for comparison to demonstrate the efficiency of the execution of different models. While the conventional CSR method takes 10.11 min, CSR_HH model is reduced to 0.05 min, indicating a significant improvement in efficiency after identifying frequent sequence patterns with the dependency relation and the part-of-speech. And the three models proposed above all have varying degrees of efficiency when compared to the CSR approach. In particular, the use of both CSR_DP and CSR_HH models with dependency direction to shorter sequence lengths saves time considerably and allows a remarkable improvement in the efficiency of generating sequence rules. The reason for this is that these two models generate

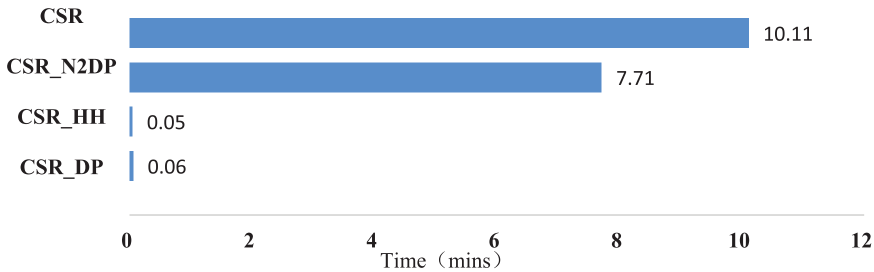

Figure 5. Computation time for different models generating sequence rules.

Table 15. Optimal parameter, CSRs quantity and computation time of four models.

	λ	CSR Quantity	Computation Time (mins)
CSR	0.5	405	10.53
CSR_N2DP	0.3	2457	7.42
CSR_DP	0.1	2243	0.06
CSR_HH	0.2	698	0.05

sequences by comparing words and their dependency components only, without breaking the coherence of the sentence components.

In the next step, each of the four models mentioned above is used to extract rules from the training set, and we collect the set of rules corresponding to each model separately.

The accuracy of these rules for identifying comparative relations is then observed with test data to compare the effectiveness of these models, the results are shown in Figure 6. The recognition process is carried out using a five-fold cross-validation approach, and the evaluation indicators for the identification results are referred to in equation (2) (3) (5). As shown in Figure 6, this study extracts and validates 20 sets of frequent sequence rules generated by four algorithms and five different λ values, and obtains precision, recall and F1-measure for each of the 20 sets of experiments. Overall, the CSR_HH model works best in assessing the indicators and the effects of the different λ values are more balanced. Furthermore, the number of sequence rules generated and the computation time of different methods are compared in Table 15 to demonstrate the advantages of the CSR_HH model in terms of efficiency and effectiveness of balancing comparative relation mining, subject to the selection of the optimal λ value for each model. Finally, we show in Table 16 the recognition results of the four different models in comparative relation mining, specifically, comparing the precision and recall of these models in identifying comparative and non-comparative relations, as well as the overall accuracy and F1-measure. Among them, CSR_N2DP, CSR_DP and CSR_HH models all have better recognition results than CSR model, with CSR_HH achieving a maximum overall accuracy of 91.83% and an F1-measure of 78.43%.

4.3. Implicit comparative relation mining with the entity-based double recognition method

Regarding the implicit comparative relation mining, we adopt a double recognition method based on the hybrid CSR model and product named entity recognition. After the

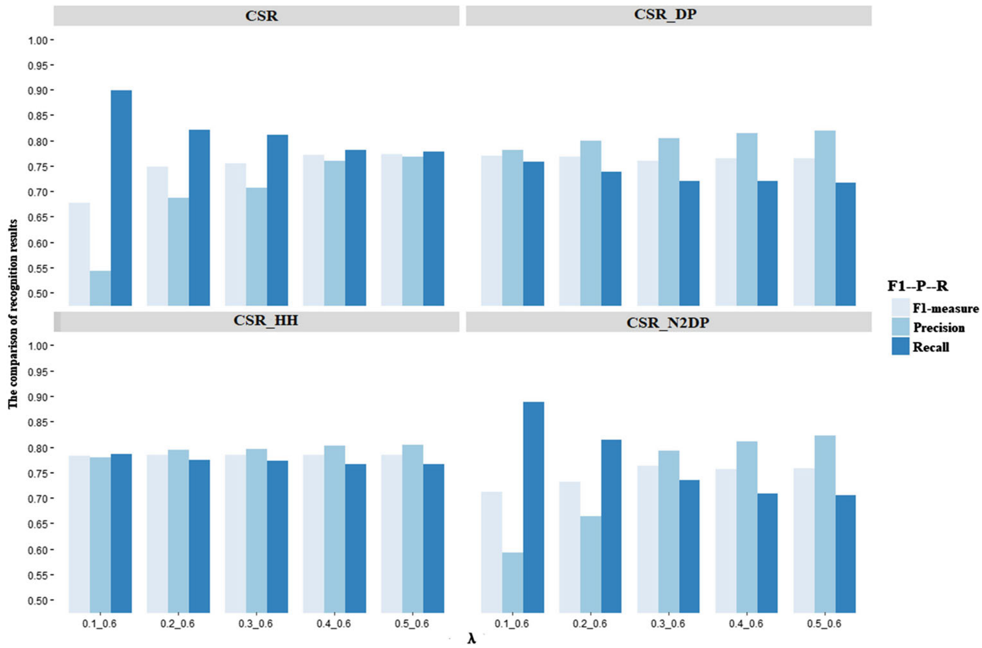

Figure 6. The precision, recall and F1-measure under different λ values.

Table 16. Recognition results.

	Precision		Recall			
Models	Comparative	Non-Comparative	Comparative	Non-Comparative	Overall Accuracy	F1-Measure
CSR	76.85%	94.72%	77.79%	94.44%	91.25%	77.32%
CSR_N2DP	79.26%	93.81%	73.43%	95.45%	91.23%	76.24%
CSR_DP	78.12%	94.32%	75.89%	94.96%	91.31%	76.99%
CSR_HH	79.36%	94.70%	77.52%	95.22%	91.83%	78.43%

comparative and non-comparative reviews are classified using the CSR_HH model, the entity extraction technology is used to identify the comparative relations in the customer reviews to better extract the comparative relations hidden in the non-comparative reviews.

First of all, we collect over 300 mobile phone products from jd.com, thereby construct three dictionaries involving product brands, series, and models through product name structure tree. These dictionaries are then manually corrected in conjunction with the experimental data, resulting in 32 mobile phone brand names, 52 series names and 181 model names, as shown in Table 17.

As mentioned above, a review is considered to have a comparative relation when it contains another product name that differs from the current review subject, and is judged to be a comparative review. The CSR_HH algorithm is used to identify the entire review corpus, with the first recognition of implicit comparative relations included in this process. After that, for the non-comparative reviews classified by the CSR_HH method, we adopt the entity-based twice recognition algorithm mentioned above to identify them again, thereby regard the reviews in which other product name entities appeared as comparative reviews. As shown in Table 16, the entity-based double recognition method increases the overall

Table 17. The dictionary of brand, series, and model.

Brand	Series	Model
Apple	Iphone	7plus
Huawei(华为)	Honor(荣耀)	8
	Mate	40pro; 30; X2 …
Xiaomi(小米)	Redmi(红米)	2a
Samsung(三星)	Galaxy	s7
…	…	…
360		n4s

Table 18. Recognition results with the entity-based double recognition method.

	Precision		Recall			
Methods	Comparative	Non-Comparative	Comparative	Non-Comparative	Overall Accuracy	F1-Measure
CSR_HH	79.36%	94.70%	77.52%	95.22%	91.83%	78.43%
CSR_HH + ENTITY	85.83%	95.93%	83.55%	96.04%	93.73%	84.67%

accuracy of the comparative relation mining to 93.73% and the F1-measure to 84.67%. It indicates that the "algorithm + strategy" solution is reasonable and effective in identifying implicit comparative relations of customer reviews, thus improving the overall recognition results (Table 18).

4.4. Analysis and discussion

To achieve a higher effectiveness of sequence generation, we propose three CSR-based methods, including the CSR_DP, CSR_N2DP, and CSR_HH models. In contrast to the classic CSR model, three new models take different improvements to generate sequence rules. Among them, the CSR_N2DP model does not take into account the dependency direction, but replaces the part of speech of the original sequence with dependency components. On this basis, this model generates a large number of invalid sequence rules by the fixed window method, which is due to the fact that the model disrupts the coherence of sentence components. In order to improve the above deficiencies, CSR_DP model extracts the sequence rules by combining the dependency component and dependency direction. Although the CSR_DP model ensures the coherence of sentence components, it is limited by the accuracy of the dependency parsing, resulting in frequent rules often being split into multiple weak rules. Thus, the CSR_HH model is proposed to overcome the shortcomings of the above methods. This method extracts the frequent sequence rules based on the dependency relation and the part-of-speech, regardless of the dependency components of comparison words.

In the experiment, we illustrate the complexity and overhead analysis of four models. First, the complexity of these models is consistent since the three new models are based on the CSR approach, and the improvements are mainly in the composition of the sequences. During the execution of the models, all the input data are customer reviews and the output data are sequence rules, the difference lies in the accuracy and efficiency of these methods. In terms of computation time, as shown in Figure 5, the CSR_N2DP model takes 7.71 min, and the CSR_DP model takes 0.06 min, and the CSR_HH model takes only 0.05 min, which is a very low overhead. Comparing the experimental results with the models in Figure 6, Tables 15 and 16, the results show that the CSR_HH model generates CSRs better than the other methods and achieves a balance between efficiency and effectiveness in comparative relation mining.

As the focus of this study, the generation of frequent sequence rules is affected by the rule length and comparison words. We find that dependency relations can effectively reduce the sequence length to obtain the simplest rules. And in comparison relation mining, the use of comparison keywords and their part-of-speech helps to obtain frequent strong rules. Additionally, identifying implicit comparative relations can effectively improve the overall accuracy of comparative relation mining, which deserves attention in future research. We adopt an "algorithm + strategy" solution to improve the whole recognition process by identifying whether there are two or more entity names in a review, thus achieving satisfactory results.

5. Conclusion

5.1. Our findings

This paper aims at exploring the balance of effectiveness and efficiency in comparative relation mining. Previous research has focused on combining traditional CSR model with features such as word, part-of-speech, term dictionary, or using machine learning methods to identify comparative relations. Yet, few studies have dedicated on improving CSR model to make their application feasible and accurate in large datasets for comparative relation mining. We thus propose a new framework that utilises dependencies and part-of-speech to identify comparative relations in customer reviews. Compared with prior studies, this paper makes a great progress in the following two aspects:

(1) Traditional CSR model is based on keywords and part-of-speech. We propose improved CSR methods including CSR_DP, CSR_N2DP and CSR_HH models. These models pay attention to the dependency component and the dependency direction of each clause for syntactic parsing when extracting the sequence. Among them, the CSR_HH model achieves an excellent trade-off between efficiency and effectiveness in generating sequence rules. In short, it achieves the highest recognition accuracy with less computation time.
(2) The experiment results demonstrate that controlling the length of the sequence rules and the part-of-speech of the comparison keywords directly contribute to the acquisition of frequent sequence rules. Both CSR_DP and CSR_HH models utilise the dependency direction to shorten the length of the sequences, and enable to reduce the computation time and higher efficiency in sequence pattern generation. The CSR_HH model increases the effect of generating frequent strong sequence rules by replacing the dependency component of the comparison keyword with its part-of-speech.
(3) Regarding the implicit comparative relation, we propose an entity-based double recognition method to identify the comparative relations through analyzing product brands, series or models. This study firstly uses the CSR_HH model to identify the whole corpus. After that, for the non-comparative reviews classified by the first recognition,

we adopt the entity extraction technology to identify them again by extracting product brands, series, and models. Compared with prior studies, our "CSR_HH + ENTITY" solution realises higher precision in the whole recognition process, and achieves satisfactory results. Additionally, the better performance of "algorithm + strategy" solution also validates our conjecture that a review containing two entities may have a comparison relation, which could be applied in the study of implicit relation mining in future.

5.2. Managerial implications

Comparative relation mining is of great interest to both academic and the industry, since it would identify areas in which products may perform in a similar way, and yield insights about areas in which one product outperforms the other. Compared with the traditional text mining, the identification of comparative relations would enable our deep understanding of semantic information behind the textual data, and thereby the competitive position of the products in the market.

With reference to the traditional CSR model, we improve the generation of sequence rules well by optimising the length of the sequence and the expression of comparison keywords in the sequence, thereby achieve the balance of effectiveness and efficiency in comparative relation mining. This study is a good reference for future work on identifying comparative relations using CSR methods. Besides, we adopt an " algorithm + strategy" solution to identify the implicit comparative relations in non-comparative reviews classified by other methods, thus to improve the whole recognition process. Therefore, the combination of algorithms and recognition strategies is also a good idea to improve the accuracy of comparative relation mining. In this paper, we choose Chinese reviews to conduct experiments and train model, and our methods is also applicable to English texts.

Our work broadens the research scope of traditional text mining. It is of significance for identifying comparative relations from user generated contexts, especially for business scenario such as online shopping, product competitiveness analysis. For example, according to the comparative relations, we can construct a competitive network of products, where nodes represent products, each link between two products represents a comparative relation, and the number of links between two products indicates the strength of this relation. By understanding the structure of such a network, we can study the hierarchy of the market by identifying the position of each product. Furthermore, we can extract the competitive relations with the target product (which can be presented as the core node) and thus identify the competitors of the target product. Finally, we can obtain and illustrate the competitive advantages and disadvantages between the target product and its competitors. Therefore, the experimental results show that our proposed method achieves better results in identifying comparative relations, thus providing technical support for identifying competitors in the next step.

In summary, customer reviews are considered to be one of the resources available in social media with the potential to extract comparative relations. If a decision maker creates a monitoring system on comparative relation mining for market perception and competitive intelligence, it will be more responsive to customer feedback on target product. As a result, decision makers can adjust their market strategies faster than using traditional methods, compensating for weaknesses and highlighting strengths through comparisons with competitors.

5.3. Future work

Even though this study provides useful insights about how comparative relations are identified with customer reviews as the example, we also acknowledge some limitations and provide potential directions for future research. As online reviews are unstructured-free-texts in natural languages, yet some comparison patterns are not identified accurately by the proposed method. Such problem mainly occurs for those comparative sentences that fail to meet the grammatical rules in the commentary texts. It is difficult to extract rules because of ellipsis, substitution words, and Internet terms. Additionally, our method still needs some manual annotation for some complex rules. Finally, the recognition of implicit comparative relation by double recognition method is still in an exploration stage. It could be limited and inaccurate to evaluate the implicit comparison only by recognising different entities. To this end, future research should focus on three aspects: (1) understanding of the composition of comparative reviews through machine learning; (2) proposing an unsupervised recognition method to further reduce human intervention and facilitate large-scale commercial applications; (3) optimising the recognition of implicit comparative relations.

Disclosure statement

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

Funding

This work is supported by the Natural Science Foundation of China [72001215, 71771177], the Fund from Chongqing Key Laboratory of Social Economic and Applied Statistics [KFJJ2019099], Shanghai Municipal Education Science Research Project (Philosophical and Social Sciences General Project, No. A2023010).

References

Alhamzeh, A., Bouhaouel, M., Egyed-Zsigmond, E., & Mitrović, J. (2021). Distilbert-based argumentation retrieval for answering comparative questions. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2021)*, *2936* (pp. 209–212).

Anwar, T., & Uma, V. (2022). CD-SPM: Cross-domain book recommendation using sequential pattern mining and rule mining. *Journal of King Saud University – Computer and Information Sciences*, *34*(3), 793–800. https://doi.org/10.1016/j.jksuci.2019.01.012

Beloucif, M., Yimam, S. M., Stahlhacke, S., & Biemann, C. (2022). Elvis vs. M. Jackson: Who has More albums? Classification and identification of elements in comparative questions. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France (pp. 3771–3779).

Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., & Hagen, M. (2022). Towards understanding and answering comparative questions. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 66–74). https://doi.org/10.1145/3488560.3498534.

Bondarenko, A., Braslavski, P., Volske, M., Aly, R., Frobe, M., Panchenko, A., Biemann, C., Stein, B., & Hagen, M. (2020). *Comparative web search questions*. Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20). Association for Computing Machinery, New York, NY, USA, 52–60. https://doi.org/10.1145/3336191.3371848

Fei, H., Ren, Y., & Ji, D. (2020). Boundaries and edges rethinking: an end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, *57*(6), 102311. https://doi.org/10.1016/j.ipm.2020.102311

Gao, S., Tang, O., Wang, H., & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, *71*, 19–32. https://doi.org/10.1016/j.ijhm.2017.09.004

Gao, S., Wang, H. W., Liu, J. Q., Zhu, Y. J., & Tang, O. (2023). Comparative relation mining of online reviews: A hierarchical multi-attention network model. *International Journal of Mobile Communications*, *22*(2), 212–236. https://doi.org/10.1504/IJMC.2023.132572

Guo, Y., Wang, F., Xing, C., & Lu, X. (2022). Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *53*, 101141. https://doi.org/10.1016/j.elerap.2022.101141

Huang, X. J., Wan, X. J., & Yang, J. W. (2008). Learning to identify Chinese comparative sentences. *Journal of Chinese Information Processing*, *22*(5), 30–38. In Chinese.

Iso, H., Wang, X., Angelidis, S., & Suhara, Y. (2021). Comparative opinion summarization via collaborative decoding. In *Proceedings of the Association for Computational Linguistics*, Dublin, Ireland (pp. 3307–3324). arXiv preprint arXiv:2110.07520.

Jain, P. K., Quamer, W., Pamula, R., & Saravanan, V. (2023). SpSAN: Sparse self-attentive network-based aspect-aware model for sentiment analysis. *Journal of Ambient Intelligence and Humanized Computing*, *14*(4), 3091–3108. https://doi.org/10.1007/s12652-021-03436-x

Jain, P. K., Srivastava, G., Lin, J. C. W., & Pamula, R. (2022). Unscrambling customer recommendations: a novel LSTM ensemble approach in airline recommendation prediction using online reviews. *IEEE Transactions on Computational Social Systems*, *9*(6), 1777–1784. https://doi.org/10.1109/TCSS.2022.3200890

Khan, A., Younis, U., Kundi, A. S., Asghar, M. Z., & Ahmed, I. (2020). Sentiment classification of user reviews using supervised learning techniques with comparative opinion mining perspective. In *Proceedings of the 2019 Computer Vision Conference (CVC)*, 21 (pp. 23–29).

Kim, S. G., & Kang, J. M. (2018). Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management*, *54*(6), 938–957. https://doi.org/10.1016/j.ipm.2018.06.003

Liu, H., Yin, X., Song, S., Gao, S., & Zhang, M. (2022). Mining detailed information from the description for App functions comparison. *IET Software*, *16*(1), 94–110. https://doi.org/10.1049/sfw2.12042

Liu, J., Wang, X., & Huang, L. (2021a). Fusing various document representations for comparative text identification from product reviews. In *Proceedings of the International Conference on Web Information Systems and Applications* (pp. 531–543).

Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*, *123*, 113079. https://doi.org/10.1016/j.dss.2019.113079

Liu, Z., Qin, C. X., & Zhang, Y. J. (2021b). Mining product competitiveness by fusing multisource online information. *Decision Support Systems*, *143*, 113477. https://doi.org/10.1016/j.dss.2020.113477

Liu, Z., Xia, R., & Yu, J. (2021c). Comparative opinion quintuple extraction from product reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3955–3965). https://doi.org/10.18653/v1/2021.emnlp-main.322.

Messaoudi, C., Guessoum, Z., & Ben Romdhane, L. (2022). Opinion mining in online social media: a survey. *Social Network Analysis and Mining*, *12*(1), 25–32. https://doi.org/10.1007/s13278-021-00855-8

Noorian Avval, A.A., Harounabadi, A. (2023). A hybrid recommender system using topic modeling and prefixspan algorithm in social media. *Complex & Intelligent Systems, 9*, 4457–4482. https://doi.org/10.1007/s40747-022-00958-5

Ping, Q., & Chen, C. (2018). LitStoryTeller+: an interactive system for multi-level scientific paper visual storytelling with a supportive text mining toolbox. *Scientometrics*, *116*(3), 1887–1944. https://doi.org/10.1007/s11192-018-2803-x

Sagnika, S., Mishra, B. S. P., & Meher, S. K. (2021). An attention-based CNN-LSTM model for subjectivity detection in opinion-mining. *Neural Computing and Applications*, *33*(24), 17425–17438. https://doi.org/10.1007/s00521-021-06328-5

Serrano-Guerrero, J., Romero, F. P., & Olivas, J. A. (2021). Fuzzy logic applied to opinion mining: a review. *Knowledge-Based Systems*, *222*, 107018. https://doi.org/10.1016/j.knosys.2021.107018

Subhashini, L. D. C. S., Li, Y., Zhang, J., Atukorale, A. S., & Wu, Y. (2021). Mining and classifying customer reviews: a survey. *Artificial Intelligence Review*, *54*(8), 6343–6389. https://doi.org/10.1007/s10462-021-09955-5

Tkachenko, M., & Lauw, H. W. (2017). Comparative relation generative model. *IEEE Transactions on Knowledge and Data Engineering*, *29*(4), 771–783. https://doi.org/10.1109/TKDE.2016.2640281

Vedula, N., Collins, M., Agichtein, E., & Rokhlenko, O. (2023). Generating explainable product comparisons for online shopping. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (pp. 949–957). https://doi.org/10.1145/3539597.3570489.

Vo, D. T., Al-Obeidat, F., & Bagheri, E. (2020). Extracting temporal and causal relations based on event networks. *Information Processing & Management*, *57*(6), 102319. https://doi.org/10.1016/j.ipm.2020.102319

Vo, D. T., & Bagheri, E. (2017). Self-training on refined clause patterns for relation extraction. *Information Processing and Management*, *54*(4), 686–706.

Wang, H., Chen, C., Xing, Z., & Grundy, J. (2021). Difftech: Differencing similar technologies from crowd-scale comparison discussions. *IEEE Transactions on Software Engineering*, *48*(7), 2399–2241. https://doi.org/10.1109/TSE.2021.3059885

Wang, H. W., Gao, S., Yin, P., & Liu, J. N. (2017a). Competitiveness analysis through comparative relation mining: evidence from restaurants' online reviews. *Industrial Management & Data Systems*, *117*(4), 672–687. https://doi.org/10.1108/IMDS-07-2016-0284

Wang, W., Xin, G., & Wang, B. (2017b). Sentiment information Extraction of comparative sentences based on CRF model. *Computer Science and Information Systems*, *14*(3), 823–837. https://doi.org/10.2298/CSIS161229031W

Wei, N., Zhao, S., Liu, J., & Wang, S. (2022). A novel textual data augmentation method for identifying comparative text from user-generated content. *Electronic Commerce Research and Applications*, *53*, 101143. https://doi.org/10.1016/j.elerap.2022.101143

Wei, N., Zhao, S., Liu, J., & Wang, S. (2023). A review for comparative text mining: From data acquisition to practical application. *Journal of Information Science*, published online. https://doi.org/10.1177/01655515231165228

Yang, S., Wei, R., Guo, J. Z., & Tan, H. L. (2020). Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis. *Journal of Web Semantics*, *63*(8), 100578. doi:10.1016/j.websem.2020.100578