Research Article

Yan Wang*

Research on the TF–IDF algorithm combined with semantics for automatic extraction of keywords from network news texts

https://doi.org/10.1515/jisys-2023-0300
received December 08, 2023; accepted May 14, 2024

Abstract: As the number of online news texts continues to increase, the algorithm of automatic keyword extraction becomes a key content in facilitating users' fast access to the desired content. This article first introduced two common algorithms: term frequency–inverse document frequency (TF–IDF) and TextRank. Then, the calculation of news title weight was added to the TF–IDF algorithm according to the characteristics of network news text. Moreover, a new automatic extraction algorithm was designed by applying Word2vec to extract semantics. The experimental results demonstrated that on the ACE2005 dataset, as the quantity of automatically extracted keywords increased, the accuracy of the TF–IDF, TextRank, and the semantics-combined TF–IDF algorithms gradually decreased, and the recall rates gradually increased. When five keywords were extracted, the gap of the semantics-combined TF–IDF algorithm with the other two algorithms was the largest, and its accuracy, recall rate, and *F*-measure were 72.77, 78.64, and 75.59%, respectively. Finally, the *F*-measure of the semantics-combined TF–IDF algorithm reached 81% for network news texts. The experimental results prove the performance of the semantics-combined TF–IDF algorithm in automatically extracting keywords from network news texts, and it will have promising applications in practice.

Keywords: news text, semantics, automatic keyword extraction, term frequency–inverse document frequency, precision

AMS Mathematics Subject Classification number: 68W40

1 Introduction

The term "online news text" refers to news texts that are disseminated through the Internet, which have a faster dissemination speed and wider coverage compared to traditional news texts. With the Internet's ongoing evolution, there has been a significant increase in the number of news texts in the network, which facilitates people to get news information more quickly and directly but also makes it more difficult for people to find the news they want. The use of keywords can help readers quickly comprehend the main content of a text and thus improve search efficiency, which has a significant role in various fields, such as text categorization and information retrieval [1]. With the massive growth of information, the traditional way of manual annotation has become increasingly difficult to meet current needs; thus, algorithms for automatic keyword extraction have been widely studied [2]. Compared with ordinary texts, network news texts are different in terms of text structure and writing techniques, so the present algorithms for automatic keyword extraction are not fully applicable. In this article, the traditional term frequency–inverse document frequency (TF–IDF) method was combined with the Word2vec word vector model to improve semantic extraction, and the

* **Corresponding author: Yan Wang,** School of Literature, Cangzhou Normal University, Cangzhou, Hebei, 061000, China, e-mail: wangyan@caztc.edu.cn

performance of this combined approach was proved through experiments on a dataset. The research in this article provides a new reliable method for automatically extracting keywords from online news texts, which can serve as the foundation for classifying and retrieving online news texts, thereby further enhancing the efficiency of processing such texts.

2 Related works

Thiyagarajan et al. [3] studied three popular keyword extraction techniques: the rapid automatic keyword extraction, TF–IDF, and semantic fingerprinting algorithms, and found through experiments that the TF–IDF algorithm had the strongest correlation with the human assessment. Li et al. [4] designed a new unsupervised method for Weibo texts by combining two hashtag enhancement algorithms and found through experiments that the method was accurate. Yang et al. [5] introduced a word network based on the relationship between sentences. A new word-sentence approach proposed by them was found to be superior to the classical TF–IDF and TextRank algorithms in aspects of precision and recall rate through experiments. Hassani et al. [6] conducted a study on video text mining and proposed a new key phrase extraction method that considered the local and global features of every candidate phrase and conducted experiments on five datasets in English and Persian and found that the method performed better in aspects such as precision. Okada et al. [7] extended a multi-keyword pattern matching machine, called the Aho and Corasick machine, and proposed an effective substring search method to achieve keyword extraction. The simulation results showed that the method had good performance. Azcarraga et al. [8] put forward an approach called liGHtSOM, based on analyzing how weights distribute in the weight vector of the training graph and simple operations of the random projection matrix applied for input data compression. The experiment showed that the keywords obtained by the approach were highly accurate. Tixier et al. [9] introduced an unsupervised technique using the degradation of graphs, carried out experiments on documents of different sizes, and obtained good performance. Campos et al. [10] proposed YAKE, a lightweight unsupervised keyword extraction approach that uses statistical characteristics of the text from a single document to select the most significant keywords within it, and demonstrated the advantages of the method through experiments on 20 datasets. Yan et al. [11] integrated eye movement signals with electroencephalogram (EEG) signals and utilized neural networks to automatically extract keywords from microblogs. They verified the collaborative effect of EEG and eye movement signals through experiments. Zhang and Zhang [12] introduced a method that utilizes human reading time for keyword extraction. They extracted fixation durations from publicly available language resources and designed two neural network models for keyword extraction. The effectiveness of the proposed method was demonstrated through both quantitative and qualitative experiments. Zhang et al. [13] developed a neural framework for extracting keyphrases, which obtains indicative representations through conversation context encoders and inputs them into the keyphrase table to extract important words. The experiment found that this method had better performance than previous models.

3 Automatic keyword extraction algorithm combined with semantics

Keywords refer to the words that have an important role in the text [14], and they are useful in helping people to quickly understand the content of the full text and find the text they need more quickly. Keywords are widely used in academic papers and also have good performance in various online texts.

The TF–IDF algorithm is an earlier and frequently used algorithm for automatic keyword extraction [15]. This algorithm considers that the significance of a word is directly proportional to how often it appears in the document but inversely proportional to its frequency in the corpus. TF refers to the term frequency. The calculation method of TF is

$$\mathrm{TF}_{i,j} = \frac{N_{i,j}}{\sum_k N_{k,j}}, \tag{1}$$

where $N_{i,j}$ denotes the occurrence frequency of word i in text d_j and k denotes the quantity of different words in text d_j.

IDF refers to inverse document frequency

$$\mathrm{IDF}_i = \log\frac{|D|}{|j : t_i \in d_j|}, \tag{2}$$

where $|D|$ is the total quantity of texts in the corpus and $|j : t_i \in d_j|$ refers to the quantity of texts containing word i in the corpus.

The TF–IDF value is obtained by

$$\mathrm{TF} - \mathrm{IDF}_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{IDF}_i. \tag{3}$$

If a word has a high TF value and a low IDF value, the word is considered to have great criticality [16], and this method is simple to operate and widely used [17].

The TextRank algorithm is a refined version of the PageRank algorithm [18]. The principle of PageRank is that if a web page is linked to many other web pages, it indicates that the web page is relatively important, which means its PageRank value is high. PageRank is calculated using the following equation:

$$S(V_i) = (1 - d) + d \times \sum_{j \in \mathrm{In}(V_i)} \frac{1}{|\mathrm{Out}(V_j)|} S(V_j), \tag{4}$$

where $S(V_i)$ is the PR value of a web page V_i, V_j is the web page linked to V_i, i.e., the inbound link, $\mathrm{In}(V_i)$ is the set of inbound links, and $\mathrm{Out}(V_j)$ is the quantity of elements in the set of links pointing to external web pages in web page j.

TextRank is an algorithm for ranking based on graphs. It treats sentences or words in a text as nodes of a graph, treats the relationships between them as edges, and determines their importance by calculating the weights between the nodes.

The calculation formula of the PageRank-based TextRank algorithm is

$$\mathrm{WS}(V_i) = (1 - d) + d \times \sum_{V_j \in \mathrm{In}(V_i)} \frac{W_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} W_{jk}} \mathrm{WS}(V_j), \tag{5}$$

where $\mathrm{WS}(V_i)$ is the weight of sentence i, W_{ji} refers to the resemblance of sentences, and d is the damping factor, 0.85 usually.

A text is segmented into sentences. Candidate keyword graph $G = (V, E)$ is built after preprocessing. V is the set of nodes, i.e., the obtained candidate keywords, and E is the edge between two points, which indicates the co-occurrence relationship. Subsequently, the node weight is calculated according to the above formula to obtain the most important T words.

Most of the ordinary texts are single texts, while online news texts are generally composed of titles and bodies. According to the characteristics of news texts, the titles are usually a high summary of the main content of the news. To further enhance the performance of automatic keyword extraction from online news texts, this article improves the TF–IDF algorithm by combining semantics.

First, consideration of the title is added when calculating the importance of words:

$$\mathrm{HF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} + 1, \tag{6}$$

$$\mathrm{TF} - \mathrm{HF} - \mathrm{IDF}_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{HF}_{i,j} \times \mathrm{IDF}_i, \tag{7}$$

where $n_{i,j}$ means the quantity of word i in the title of news text d_j and k is the quantity of different words in the title of d_j.

The TF–HF–IDF is combined with the Word2vec model [19] to improve the extraction of semantics. Suppose there is text $d_j = (t_1, t_2, \ldots, t_m)$, the word vector obtained after every word is trained by the Word2vec model is: $w_i = (v_1, v_2, \ldots, v_M)$. Then, the word vector is weighted. The obtained vector is expressed as

$$t_i = \sum_{i=1}^{M} \mathrm{score}(t_i, d_i) w_i, \tag{8}$$

where $\mathrm{score}(i, d)$ denotes the TF–HF–IDF value of word i in the text, which is used as the initial weight of the word.

The specific process of the method is as follows. After processing the text by word segmentation and stop word elimination, the TF–HF–IDF value is computed, and then the individual words are represented by Word2vec word vectors. After that, the semantic-based similarity of the processed words is calculated. The set of semantic topic concepts is obtained by the hierarchical clustering algorithm [20], i.e., the set of words with similar semantics. Finally, the comprehensive weight value is calculated:

$$W - \mathrm{score}(t_i, d) = \mathrm{score}(t_i, d) + \sum_{j=1}^{N} \mathrm{sim}(t_i, t_j), \tag{9}$$

$$\mathrm{sim}(t_i, t_j) = \cos\theta = \frac{e_i \cdot e_j}{|e_i| \cdot |e_j|}, \tag{10}$$

where $\sum_{j=1}^{N} \mathrm{sim}(t_i, t_j)$ refers to the sum of the semantic similarity between word t_i and the other words, $\mathrm{sim}(t_i, t_j)$ is the Word2vec-based semantic similarity between words t_i and t_j, e_i is the word vector of t_i, and e_j is the word vector of t_j.

Finally, according to the comprehensive weight of words, the word with the highest weight in every semantic topic concept set is used as a keyword to obtain the keyword set of the document.

4 Experimental analysis

The experiments were conducted on a Windows 7 system with 4 GB memory. The word separation system was Institute of Computing Technology, Chinese Lexical Analysis System [21]. The algorithm was implemented through Java language programming. The experimental dataset came from the ACE2005 corpus [22], containing news reports from Xinhua News Agency and China National Radio. Table 1 presents the statistics of the corpus.

Table 1: ACE2005 corpus

	1.5 characters = 1 word
Broadcast	20,000 words
Newswire	20,000 words
Weblog	10,000 words

There were 500 texts in the dataset. The semantics-combined TF–IDF algorithm was compared with TF–IDF and TextRank algorithms. The evaluation indexes as follows.

(1) Precision: $P = A/B$, where A is the quantity of keywords extracted correctly by the algorithm and B is the quantity of all keywords extracted by the algorithm.

(2) Recall rate: $R = A/C$, where C is the actual total number of keywords.

(3) F-measure: F-measure $= 2PR/(P + R)$, indicating the overall performance of an algorithm.

First, for Word2vec, the chosen dimension of word vectors will affect the results. Under other consistent conditions, the performance of the proposed method with different dimensions (64-dimensional, 96-dimensional, 128-dimensional, and 200-dimensional) was compared. Five keywords were extracted, and the outcomes are presented in Table 2.

Table 2: The effect of word vector dimensions on the algorithm performance

	Training time (min)	Accuracy (%)
64 Dimensions	8.32	68.92
96 Dimensions	9.73	67.34
128 Dimensions	10.67	71.16
200 Dimensions	12.24	71.77

It was seen that with the gradual increase in word vector dimension, the training time of the algorithm gradually increased. When the dimension was 200, the training time of the algorithm was 12.24 min, which was increased by 14.71% compared to that when the dimension was 128, and the accuracy was 71.77%, which was increased by 0.61% compared to that when the dimension was 128. This indicated that the training time was significantly increased, but the improvement of the accuracy was limited. Therefore, the word vector dimension was set as 128 in the following experiments.

The impact of the quantity of keywords on the algorithm performance was compared. The number of keywords sampled was 1–10, and the precision variation is presented in Figure 1.

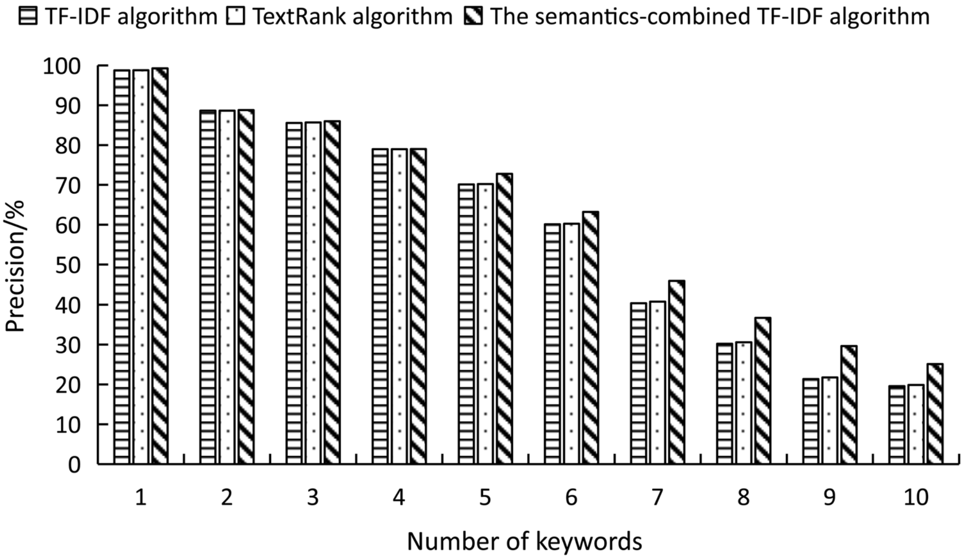


Figure 1: Comparison of precision between different algorithms.

It was seen from Figure 1 that when only one keyword was extracted, the precision of all three algorithms was close to 100%, indicating that all three algorithms performed good when only one keyword was extracted. When the number of extracted keywords reached five, the gap between the semantics-combined TF–IDF algorithm and the TF–IDF and TextRank algorithms started to increase; at this moment, the precision of TF–IDF and TextRank algorithms were 70.12 and 70.23%, respectively, while the precision of the semantics-combined algorithm reached 72.77%, which was improved by 2.65 and 2.54%, respectively. When the number of keywords reached ten, all three algorithms achieved their minimum accuracy levels, 19.56, 19.87, and 25.12%, respectively.

The variation of the recall rate of different algorithms is shown in Figure 2.

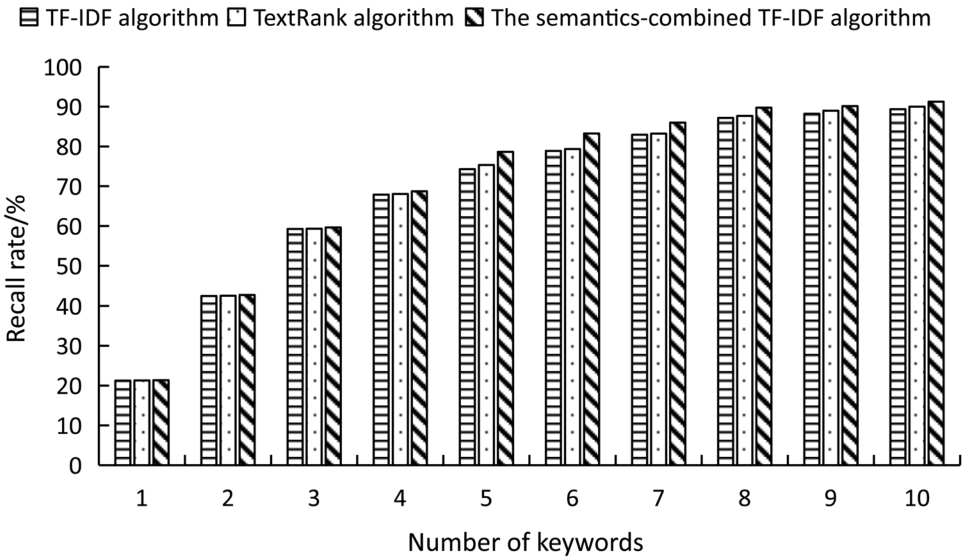
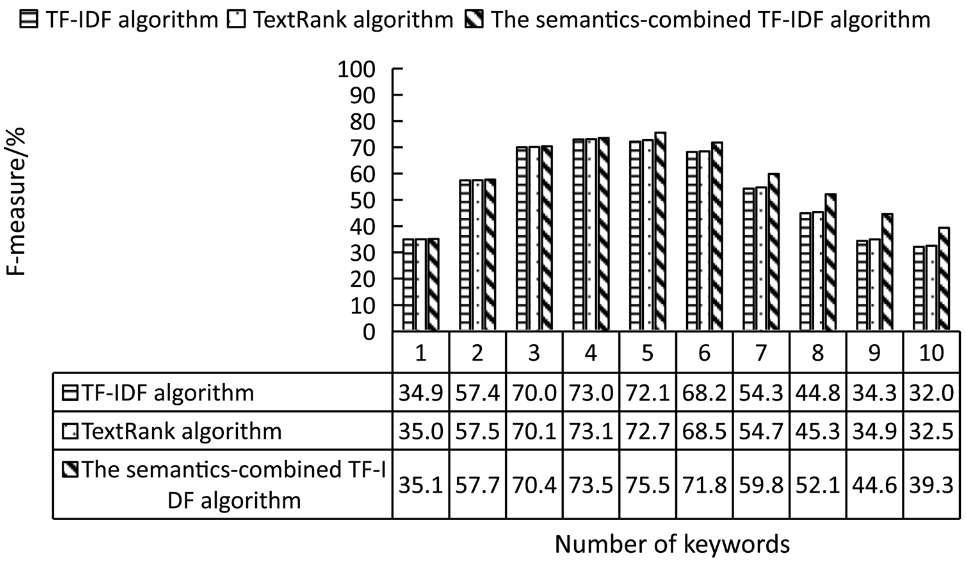


Figure 2: Comparison of recall rates of different algorithms.

It was seen from Figure 2 that contrary to the precision, the recall rates of different algorithms gradually improved as the number of keywords automatically extracted increased, but similar to the precision, the gap between algorithms started to become obvious when the number of keywords reached five; at this moment, the recall rate of TF–IDF and TextRank algorithms were 74.28 and 75.34%, respectively, while the recall rate of the semantics-combined algorithm was 78.64%, which was improved by 4.36 and 3.3%, respectively. When the number reached ten, the recall rate of all three algorithms was around 90%.

Finally, the *F*-measure of different algorithms was compared, as shown in Figure 3.

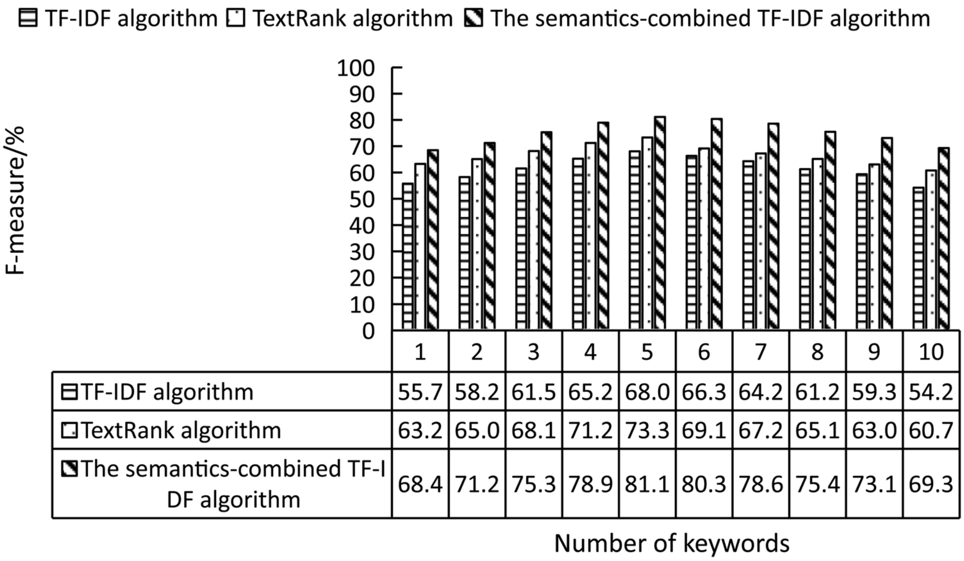


	1	2	3	4	5	6	7	8	9	10
▤TF-IDF algorithm	34.9	57.4	70.0	73.0	72.1	68.2	54.3	44.8	34.3	32.0
☐TextRank algorithm	35.0	57.5	70.1	73.1	72.7	68.5	54.7	45.3	34.9	32.5
▧The semantics-combined TF-IDF algorithm	35.1	57.7	70.4	73.5	75.5	71.8	59.8	52.1	44.6	39.3

Figure 3: Comparison of *F*-measure between different algorithms.

It was seen from Figure 3 that when the number of keywords was small, the difference in the *F*-measure was not obvious and almost the same. When five keywords were extracted, the *F*-measure of the semantics-combined algorithm was 3.45 and 2.89% higher than the other two algorithms, respectively. When the number of keywords reached ten, the *F*-measure of TF–IDF and TextRank algorithms were 32.09 and 32.55%, respectively, while the *F*-measure of the semantics-combined algorithm was 39.39%. Finally, it was concluded from Figures 1–3 that the semantics-combined algorithm had the best performance when the number of extracted keywords was five.

In order to further understand the effect of the semantics-combined TF–IDF algorithm on the automatic extraction of keywords from web news texts, 500 articles were crawled from news web pages based on the Scrapy framework through a crawler tool for experiments. Ten keywords were manually labeled. Under different numbers of extracted keywords, the comparison of the *F*-measure is presented in Figure 4.



	1	2	3	4	5	6	7	8	9	10
⊟TF-IDF algorithm	55.7	58.2	61.5	65.2	68.0	66.3	64.2	61.2	59.3	54.2
□TextRank algorithm	63.2	65.0	68.1	71.2	73.3	69.1	67.2	65.1	63.0	60.7
▨The semantics-combined TF-IDF algorithm	68.4	71.2	75.3	78.9	81.1	80.3	78.6	75.4	73.1	69.3

Figure 4: Comparison of *F*-measure between different algorithms for the automatic extraction of keywords from online news texts.

From Figure 4, it can be found that the *F*-measure of the semantics-based approach was higher than the other approaches in automatically extracting keywords from 500 crawled online news texts; when five keywords were extracted, the *F*-measure of the proposed algorithm reached the highest, 81.13%, which was 13.06% higher than the TF–IDF approach and 7.8% higher than the TextRank approach, further proving the performance of the proposed method.

Wen et al. [23] proposed an optimized weighted TextRank algorithm to extract keywords. When five keywords were extracted, the outcomes are displayed in Table 3.

Table 3: Comparison between the TextRank and weighted TextRank methods

	Result			
	Weight value	**Precision**	**Recall rate**	***F*-measure**
TextRank	—	0.25656	0.48452	0.33548
Weighted TextRank	0.3	0.28575	0.53687	0.37298
	0.5	0.28616	0.53772	0.37353
	0.7	0.28575	0.53636	0.37286

In Table 3, for the weighted TextRank approach, the precision, recall rate, and *F*-measure value were the highest in keyword extraction when the weight value was taken as 0.5. When five keywords were extracted, the comparison of the growth amplitude of different indicators of the weighted TextRank method and the semantics-based TF–IDF method compared to the TextRank method is presented in Table 4.

Table 4: Comparison between the weighted TextRank method and the semantics-based TF–IDF method

		Precision	**Recall rate**	***F*-measure**
The study of literature [23]	TextRank	0.25656	0.48452	0.33548
	Weighted TextRank	0.28616 (+0.0296)	0.53772 (+0.0532)	0.37353 (+0.03805)
This study	TextRank	0.75216	0.71526	0.73325
	The semantics-based TF–IDF method	0.82916 (+0.077)	0.79323 (+0.07797)	0.81125 (+0.078)

In Table 4, the object of the study in the literature [23] was 500 news crawled from Sohu news, and the object of this article was 500 randomly crawled online news. The amount of data in the two datasets was similar. The weighted TextRank method was improved by 0.0296, 0.0532, and 0.03805 in precision, recall rate, and *F*-measure compared to the TextRank method, respectively; the semantics-based TF–IDF method was improved by 0.077, 0.07797, and 0.078 in precision, recall rate, and *F*-measure compared to the TextRank method, respectively. Comparisons revealed that the growth rate of the semantics-based TF–IDF approach was higher than that of the TextRank approach, indicating that the semantics-based TF–IDF method was more beneficial to improving the keyword extraction effect than the weighted TextRank method.

Taking one of the online news texts entitled "iPhone 14 fastest price drop: record-breaking speed" as an example, the keyword extraction effect of the semantics-based TF–IDF method was analyzed. The text content is as follows.

After the iPhone 14 full series went on sale, the offline price of two models in the standard version was lower than the initial offer price. Even with the value of heavy upgrades such as the Spirit Island and 48 million pixels, the premium for the two models in the Pro version also fell rapidly after the launch, spot goods at the original price were available for some models and colors offline, and the service break can be expected soon.

Judging by the price trend in the last 2 days, iPhone 14 has seen a big drop in the e-commerce platform, and the offline spot price has also seen a new low. iPhone 14 has dropped by about 600 yuan, iPhone 14 plus has dropped by about 800 yuan, iPhone 14 Pro has also dropped slightly, and the price of the high-capacity version has dropped a little more.

After analysis, experts believe that compared to the iPhone 12 and iPhone 13 series in the previous 2 years, the lack of price reduction in the last month after the release means that the iPhone 14 is the model with the fastest price reduction in recent years.

The manually labeled keywords and the keywords extracted by the TF–IDF, TextRank, and semantics-based TF–IDF methods are shown in Table 5.

Table 5: Example of automatic keyword extraction results

Manual labeling	**The TF–IDF algorithm**	**The TextRank algorithm**	**The semantics-combined TF–IDF algorithm**
iPhone 14	iPhone 14	iPhone 14	iPhone 14
Price reduction	Price reduction	Price reduction	Price reduction
Model	Model	Model	Model
Drop	Price	Premium	Drop
E-commerce	Offline	Price	E-commerce

It is seen from Table 5 that when automatically extracting keywords from this online news text, three keywords were correctly extracted by the TF–IDF and TextRank approaches, and the other two were different from the manually labeled results, but the keywords extracted by the semantics-combined algorithm were consistent with the manually labeled ones, which further proved the reliability of the TF–IDF algorithm combined with semantics.

5 Conclusion and future works

This article designed a TF–IDF algorithm combining semantics by combining title weights and Word2vec word vector model to improve the algorithm performance for automatic extraction of keywords. It was found through experiments that this method had advantages in precision and recall rate. This method showed greater enhancement in precision, recall rate, and *F*-measure than existing methods when extracting keywords. The case study showed that the extracted keywords had a better match with the manually labeled keywords. The semantics-combined TF–IDF algorithm can be further applied in the real world. However, this study also has some limitations, such as the small number of languages studied and the small number of texts. In future work, the applicability of the proposed method will be investigated in more languages and the scale of experiments will be further expanded to determine the reliability of the method.

Funding information: Author states no funding involved.

Author contributions: The author confirms the sole responsibility for the conception of the study, presented results, and manuscript preparation.

Conflict of interest: Author states no conflict of interest.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Ahadh A, Binish GV, Srinivasan R. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. Process Saf Environ Prot Part B. 2021;155:455–65.

[2] Zhou Q, Shi X, Ge L. Predicting mental disorder from noisy questionnaires: an anomaly detection approach based on keyword extraction and machine learning techniques. J Intell Fuzzy Syst: Appl Eng Technol. 2021;41:7167–79.

[3] Thiyagarajan G, Prasanna S, Uma B. Automation of discussion board evaluation through keyword extraction techniques: a comparative study. IOP Conference Series: Materials Science and Engineering. vol. 1131; 2021. p. 1–7.

[4] Li L, Liu J, Sun Y, Xu G, Yuan J, Zhong L. Unsupervised keyword extraction from microblog posts via hashtags. J Web Eng. 2018;17:97–124.

[5] Yang L, Li K, Huang H. A new network model for extracting text keywords. Scientometrics: An Int J All Quant Asp Sci Sci Policy. 2018;116:339–61.

[6] Hassani H, Ershadi MJ, Mohebi A. LVTIA: A new method for keyphrase extraction from scientific video lectures. Inf Process Manage: Libr Inf Retr Syst Commun Networks: An Int J. 2022;59:1–21.

[7] Okada M, Lee SS, Hayashi Y, Aoe J, Ando K. An efficient substring search method by using delayed keyword extraction. Inf Process Manag. 2021;37:741–61.

[8] Azcarraga AP, Yap T, Chua TS. Comparing keyword extraction techniques for WEBSOM text archives. Int J Artif Intell Tools. 2008;11:219–32.

[9] Tixier A, Malliaros F, Vazirgiannis M. A graph degeneracy-based approach to keyword extraction. Conference on Empirical Methods in Natural Language Processing, (Austin, Texas), Association for Computational Linguistics; 2016, Nov 1-5. p. 1860–70.

[10] Campos R, Mangaravite V, Pasquali A, Jorge AM, Nunes C, Jatowt A. YAKE! keyword extraction from single documents using multiple local features. Inf Sci. 2020;509:257–89.

[11] Yan X, Zhang Y, Zhang C. Utilizing cognitive signals generated during human reading to enhance keyphrase extraction from microblogs. Inf Process Manag. 2024;61:103614.

[12] Zhang Y, Zhang C. Enhancing keyphrase extraction from microblogs using human reading time. J Assoc Inf Sci Technol. 2021;72:611–26.

[13] Zhang Y, Zhang C, Li J. Joint modeling of characters, words, and conversation contexts for microblog keyphrase extraction. J Assoc Inf Sci Technol. 2020;71:553–67.

[14] Chen J, Hou H, Gao J. Inside importance factors of graph-based keyword extraction on chinese short text. ACM Trans Asian Low-Resour Lang Inf Process (TALLIP). 2020;19:63.1–15.

[15] Jones KS. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28:11–21.

[16] Ramezani R. A language-independent authorship attribution approach for author identification of text documents. Expert Syst Appl. 2021;180:1–21.

[17] Li S, Ou J. Multi-label classification of research papers using multi-label k-nearest neighbour algorithm. J Phys: Conf Ser. 2021;1994:1–10.

[18] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; 2004. p. 404–11.

[19] Sumayasuhana S, Ashokkumar S. An enhancement in machine learning approaches for novel data mining serendipitous drug usage to reduce false positive rate from social media comparing word2vec Algorithm. ECS Trans. 2022;107:13329–44.

[20] Sun H. RETRACTED: business data analysis based on hierarchical clustering algorithm in the context of big data. J Phys: Conf Ser. 2021;1744:1–7.

[21] Xu W. A chinese keyword extraction algorithm based on TFIDF method. Inf Studies: Theory Appl. 2008;31:298–302.

[22] Shi X, Zeng X, Wu J, Hou M, Zhu H. Context event features and event embedding enhanced event detection. ACAI 2020: 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, (Sanya China). Association for Computing Machinery; 2020, Dec 24–26. p. 1–6.

[23] Wen Y, Yuan H, Zhang P. Research on keyword extraction based on Word2Vec weighted TextRank. 2016 2nd IEEE International Conference on Computer and Communications (ICCC). Chengdu, China: IEEE; 2016, Oct 14–17. p. 2109–13.