Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Keywords Extraction of Film Reviews Based on the KeyBERT Model

Qikang Huang*

Department of Intelligent Science, Communication University Of China, Beijing, China

*Corresponding author: huangq20@uni.coventry.ac.uk

Abstract. Current film and television platforms still struggle to effectively gather information for users and companies through public film reviews. There is a lack of applications or research on keyword extraction from film reviews. This study aims to evaluate the effectiveness and feasibility of the KeyBERT model for extracting keywords from film reviews. The precision and recall rate are utilized to assess the impact of model extraction. The test results indicate that the average precision and recall rate of film review extraction are 0.600 and 0.387, respectively, which are slightly lower than those of other types of text. Specifically, the precision rate and recall rate of plot description film reviews are 0.80 and 0.50, respectively, which are higher than the rates for multidimensional subjective analysis film reviews (0.40 and 0.33). Furthermore, they surpass the precision rate of 0.20 and the recall rate of 0.20 for subjective emotional expression type reviews. It is worth noting that the precision rates decrease from 0.80 to 0.20 as the number of words reviewed increases from 100 to 500 in subjective emotional expression type reviews. While the KeyBERT model is suitable for extracting keywords from movie reviews, it is essential to consider the classification of such reviews, the structural breakdown of lengthy text, and minimizing personal bias as much as possible.

Keywords: KeyBert, Movie review, Unsupervised, Keywords.

1. Introduction

Today, major film and television platforms widely utilize sentiment analysis and scoring systems. However, these methods can only provide emotional information from film reviews and cannot reveal other crucial details such as the performance of actors, special effects, and plot scenes. The model must extract keywords from a large volume of film reviews to acquire this information. The advantage of keyword extraction for movie reviews lies in its ability to enhance the understanding of user needs by movie review platforms and movie recommendation systems. This, in turn, enables the system to recommend movies that align more closely with users' interests and preferences, ultimately leading to an improved user experience and higher satisfaction levels. In terms of market analysis, market research can be conducted to gain insights into audience preferences and trends for various genres of films. This holds significant guidance for the film industry regarding market positioning, promotion, and production decisions.

Currently, keyword extraction technology has been developed and refined. Initially, keyword extraction methods were categorized into supervised and unsupervised methods based on the requirement for annotated data [1]. Common supervised extraction methods typically fall into two categories: classifier-based and regression-based approaches [2]. Unsupervised extraction methods encompass graph-based sorting, statistical word frequency-based approaches, subject-based techniques, and language model-based methodologies [3]. Generally, supervised extraction methods are considered to be superior to unsupervised extraction methods [4]. However, the supervised extraction method requires manual extraction of keywords from the dataset, which is challenging to accomplish with a large number of training sets [5]. The most widely used method for extraction is unsupervised.

The commonly used TF-IDF [6] and TextRank [7] algorithms have demonstrated strong performance. The KeyBERT model, in particular, excels in understanding context. KeyBERT is built upon BERT, a pre-trained deep bidirectional transformation model. Compared to keyword extraction models based

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

on traditional statistical methods or shallow models, BERT demonstrates a superior ability to comprehend the semantic and contextual relationships within text. It considers the bidirectional information of the entire text and assists in accurately capturing the context of keywords [8].

Most importantly, KeyBERT combines BERT's contextual understanding capabilities with CountVectorizer's methods for weight calculation. This hybrid approach leads to a better balance and enhances the model's efficiency in extracting keywords by employing CountVectorizer for dimensionality reduction. KeyBERT has outperformed other models in various testing conditions, although processing lengthy texts may require more time [9]. Progress has been achieved in various areas of research, including search queries [10] and social media tags[11].

Therefore, this paper utilizes the KeyBERT model to extract keywords from film reviews and subsequently evaluates them. Firstly, film reviews' extraction effect is compared with that of other types of articles. Subsequently, the effects of different types of film reviews and film reviews with different wording are extracted and compared.

2. Data and Methods

2.1. Data Source

The film review data used in the experiment is obtained from Google Datasets[12]. Five movies are discussed in this study, each generating 100 comments. The length of the comments varies from tens to hundreds of words, providing a rich and comprehensive analysis of the content. In this study, three film reviews of 200-400 words are randomly selected. One review is then chosen for plot description, multidimensional subjective analysis, and subjective feelings. All the selected reviews are about the same movie. Finally, five film reviews of approximately 100, 200, 300, 400, and 500 words are randomly chosen for further analysis.

Some of the news data used in the experiment were sourced from BBC's official website. The articles cover three main topics: war, environment, and people's livelihood. To ensure that the word count falls within the 200-400 words range, it may be necessary to appropriately truncate or excerpt sections of the articles.

In addition, the data for some academic articles in the experiment are sourced from Google Scholar, specifically in the fields of medicine[13] and finance[14]. To obtain 200-400 words of data, certain articles are selected and sections such as the abstract or conclusion are extracted. Keywords will be extracted from all the data, and 6-10 keywords will be extracted from each text (depending on the length of the text). For news and academic articles, maximum reference will be made to the keywords provided by the original author.

2.2. KeyBERT Model for Keyword Extraction

2.2.1. Experimental Method of KeyBERT Model

The KeyBERT model first utilizes the Bert model to acquire the document or candidate word embedding. It then employs the word embedding model to extract n-gram words or keywords as candidates, with CountVectorizer in sklearn being the most commonly used method at present. Finally, it calculates the cosine similarity between the document and the candidate word to identify the keyword that best represents the document[15]. The process is depicted in Fig. 1.

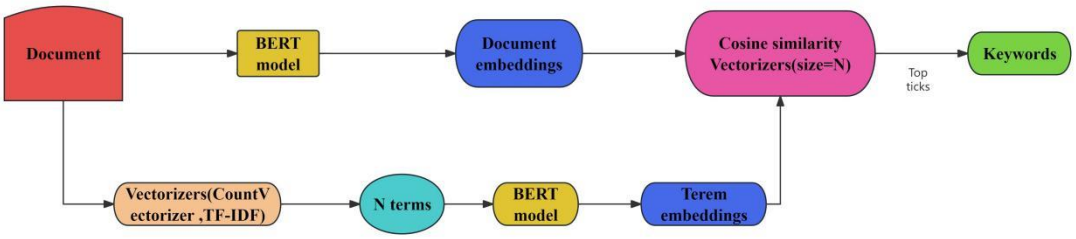

Fig.1 KeyBERT Model Keyword Extraction Process

The pre-training model selected for this experiment is the all-MiniLM-L6-v2 model. This model has shown better results in keyword extraction from English text and is currently the most commonly used model. However, processing long text content takes longer than other pre-training models.

2.2.2. Experimental Process

In terms of software and hardware testing, the test is conducted on a laptop using the Google Colab platform in the cloud. The runtime type utilized is Python3, with a T4GPU hardware accelerator, and the KeyBERT model has been installed.

In order to facilitate the test, the model has been adjusted to only output a single keyword in its actual form, without any phrases or other variations. The model is limited to extracting five keywords per text, which are the top five keywords based on their weight. Any keywords with a weight less than 0.1 will be removed.

In order to enhance result diversity and improve extraction effectiveness, the experiment utilized the model's maximum margin relevant (MMR) of the model. The diversity is adjusted to 0.4, aiming to effectively prevent repetition of different word forms or extraction of synonyms while also reducing negative effects caused by excessive diversity. Fig. 2 displays the input text and extraction results. Blue is the true keywords, yellow is the model prediction keywords, and green is the model prediction correct keywords.

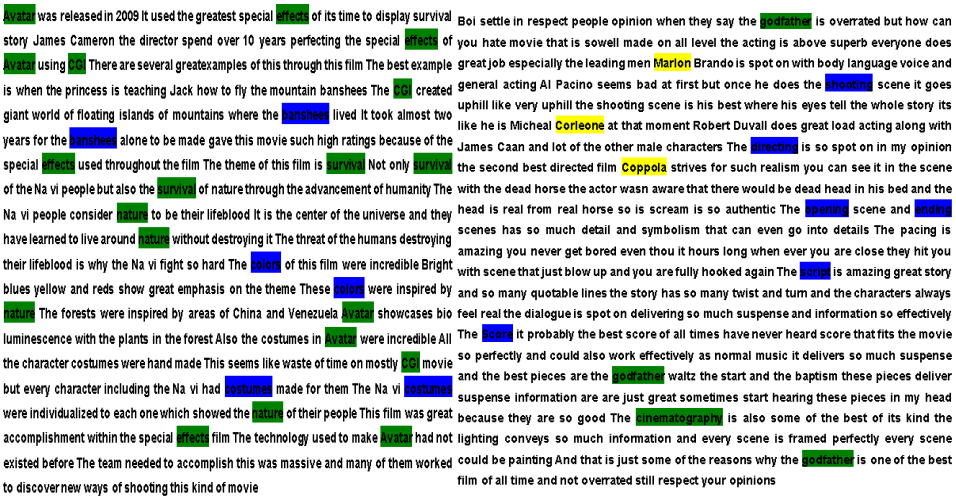

Fig. 2 The input text and extraction results

2.3. Extraction Effect Evaluation

In this experiment, the precision rate and recall rate Are utilized to assess the model's extraction effect on each text. The formula is as follows:

(1) Precision Rate:

$$\text{Precision} = \frac{\text{True case number}}{\text{True case number} + \text{False positive case number}} \quad (1)$$

Among them, the number of true cases refers to the number of keywords predicted by the model as positive cases that are positive, while the number of false cases indicates the number of keywords predicted by the model as positive cases but are negative.

(2) Recall Rate:

$$\text{recall} = \frac{\text{True case number}}{\text{True case number} + \text{False negative case number}} \quad (2)$$

Among them, the true case number refers to the number of keywords predicted by the model as positive cases and identified as positive cases. On the other hand, the false negative case number is the number of keywords predicted by the model as negative cases but positive cases.

In this experiment, precision is defined as the number of intersections between actual keywords and model-predicted keywords divided by the total number of keywords predicted by the model. Meanwhile, recall is calculated as the number of intersections between actual keywords and model-predicted keywords divided by the total number of actual keywords.

3. Results and Analysis

3.1. Horizontal Comparison

The experiment initially conducts a horizontal comparison of the KeyBERT model's effect on keyword extraction in movie reviews, news articles, and academic articles. Three groups of experiments are then conducted for each type respectively. The experimental results are presented in Table 1.

Table 1 Transverse Comparison of Experimental Results

	P1	R1	P2	R2	P3	R3	Avg P	Avg R
Film review	0.40	0.25	1.00	0.62	0.40	0.29	0.600	0.387
News	0.80	0.44	0.50	0.50	0.80	0.50	0.700	0.480
Academic article	0.60	0.38	0.60	0.50	0.80	0.50	0.670	0.460

Notes: Horizontal represents the test value, while vertical indicates the type of test text. P stands for precision and R stands for recall rate. In this experiment, a horizontal comparison is first conducted between the KeyBERT model's film review keyword extraction effect and its effect on news and academic articles. Three sets of experiments are carried out for each type, and the experimental results are presented in Table 1. Horizontal represents the test value, while vertical indicates the type of test text. P stands for precision and R stands for recall rate.

As depicted in Table 1, the KeyBERT model demonstrates lower average precision and recall rates for film review extraction compared to those for news and academic articles. This suggests that the model's keyword extraction performance in film reviews is less effective than in news and academic articles.

Based on the three experimental values of the film review, it is evident that there are significant discrepancies among the values. For instance, one group of experiments achieved an impressive precision of 1.0, while the remaining groups only reached 0.4 in terms of precision. This suggests that specific factors influence the KeyBERT model's effectiveness in extracting keywords from film

reviews. Film reviews have a consistent and structured format that is easily understandable, unlike news or academic articles. Film reviews encompass various types and components; for instance, some reviewers may focus on storytelling while others emphasize their emotional responses. These subjective elements potentially impact the outcomes of the experiment.

3.2. Comparison of Extraction Effects of Different Types of Film Reviews

Following this experiment's completion, the KeyBERT model's extraction effect on different types of movie reviews is discussed. Specifically, movie reviews are categorized into plot description type, multidimensional subjective analysis, and personal emotion expression type and evaluated accordingly. The test results can be found in Table 2.

Table 2 Comparison of Extraction Effects of Different Types of Film Reviews

	Plot description	multi-dimension analysis	Emotional perception
precision	0.80	0.40	0.20
recall	0.50	0.33	0.20

As shown in Table 2, the KeyBERT model demonstrates the highest retrieval precision rate and best recall rate for film reviews with plot descriptions. In comparison, multidimensional subjective analysis and personal emotion expression both exhibit relatively low precision and recall rates.

After analysis, the reasons may be as follows. Firstly, the difference in data distribution. Movie reviews with plot descriptions may focus more on using keywords to summarize and describe the movie's plot, which leads to better performance of the model in terms of keyword extraction. In analyzing multidimensional objective and personal emotional expression in film reviews, keyword extraction may be influenced by various factors, including subjectivity, emotional tone, and individual differences. These factors can increase the complexity of the extraction process. The second reason is related to the training content. The Bert model is trained using English Wikipedia and other articles, leading to corpus bias. As a result, the model performs poorly in personal emotion and subjective analysis. Finally, there is a difference in evaluation indicators. The multidimensional objective analysis and personal emotional expression of movie reviews may require more context and depth of understanding. In contrast, KeyBERT may be more suitable for extracting short, focused keywords, especially when it comes to describing the plot.

3.3. Comparison of Extraction Effects of Movie Reviews with Different Word Counts

At the conclusion of this experiment, the KeyBERT model is utilized to compare the extraction effects of film reviews with varying word counts. Five texts with different word counts are selected and incrementally increased from 100 words to 500 words, followed by individual evaluations. The test results are presented in Table 3. Horizontal is the number of times and vertical is the precision and recall rate

Table 3 Comparison of Effect of Film Review Extraction with Different Words

	100（words）	200	300	400	500
precision	0.80	0.75	0.40	0.60	0.20
recall	0.67	0.43	0.25	0.38	0.11

As shown in Table 3, the overall precision and recall rate of film reviews by the KeyBERT model decrease as the number of words increases. However, one set of experiments presents an exception to this trend. It has been proven that an increase in the number of words will significantly diminish the extraction effectiveness of the KeyBERT model. After analysis, the reasons may be as follows:

Firstly, there is information fragmentation. Key information in a lengthy film review may be dispersed throughout the text, and important keywords may not be located within the same section. KeyBERT is built on the representation of each token for keyword extraction. However, if key information is dispersed across different locations in the text, the model may encounter challenges in capturing the overall keyword. Secondly, the composition is intricate. The longer the film reviews, the more complex the composition becomes. In addition to proper names such as movie titles, different aspects of the story revolving around the theme also contribute to its complexity. Finally, CountVectorizer assigns weights to the impact. KeyBERT utilizes CountVectorizer internally for dimensionality reduction and weight calculations. In the case of lengthy text, the vocabulary may be more extensive, and the word frequency information in the CountVectorizer calculation is dispersed throughout the text, leading to a more intricate weight calculation and resulting in less precise extraction of keywords.

4. Conclusions

This study aims to evaluate the effectiveness of the KeyBERT model in extracting keywords from film reviews and to compare its performance with other text analysis methods. We examine the factors influencing the extraction results in different types of film reviews and varying text lengths and draw conclusions based on our findings. The KeyBERT model proves to be suitable for keyword extraction in film reviews, demonstrating an average precision and recall rate of approximately 0.1 lower than that of other article types. This suggests that its overall performance is only slightly inferior to news articles and academic papers. However, we observed significant impacts on the extraction process from both word count and content type. As the number of words in film reviews increases, we noted a decrease in precision-recall rate from 0.80 and 0.67 for 100 words down to 0.20 and 0.11 for 500 words, indicating a substantial reduction in effectiveness as text length grows. In terms of content type, we found that plot descriptions achieved a precision score of 0.80 and a recall score of 0.50, surpassing multidimensional subjective analysis (with scores of 0.40 and 0.33) as well as subjective emotional expression (scoring at 0.20 for both precision and recall). The KeyBERT model's efficacy in extracting content from film reviews appears most effective for plot descriptions but less so for other types of content.

Therefore, particular attention should be given to the following points during extraction. The initial step involves categorizing the extensive amount of film review content. It is essential to exclude emotionally expressive comments and instead focus on integrating plot descriptions and multidimensional subjective analysis into the categories. Subsequently, deconstructing the structure of the lengthy text is necessary, identifying the topics covered by each section of the comprehensive review and reclassifying them accordingly. Finally, it is important to avoid extracting excessive data all at once. This can be achieved by limiting the number of words extracted at a time, increasing the number of extractions, and integrating all relevant keywords.

In the future, to enhance the technology of keyword extraction from movie reviews, we can improve the effectiveness of this process by increasing the training of movie review datasets, preprocessing the content of movie reviews, adjusting the model structure, and incorporating sentiment analysis models.

References

[1] Turney, P.D. (2002). Learning to Extract Keyphrases from Text. ArXiv, cs.LG/0212013.

[2] Sun, Chengyu, Liang Hu, Shuai Li, Tuohang Li, Hongtu Li, and Ling Chi. 2020. "A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources" Symmetry 12, no. 11: 1864. https://doi.org/10.3390/sym12111864

[3] Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), e1339.

[4] Kim, S. N., & Kan, M. Y. (2009, August). Re-examining automatic keyphrase extraction approaches in scientific articles. In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, and Applications (MWE 2009) (pp. 9-16).

[5] Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010, October). Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 conference on empirical methods in natural language processing (pp. 366-376).

[6] Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. J. Documentation, 60(5):493–502.

[7] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

[8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[9] Nadim, M., Akopian, D., & Matamoros, A. (2023). A Comparative Assessment of Unsupervised Keyword Extraction Tools. IEEE Access.

[10] Kelebercová, L., & Munk, M. (2022). Search queries related to COVID-19 based on keyword extraction. Procedia computer science, 207, 2618-2627.

[11] Jafari, B. M., Luo, X., & Jafari, A. (2023, May). Unsupervised keyword extraction for hashtag recommendation in social media. In The International FLAIRS Conference Proceedings (Vol. 36).

[12] Benlahbib, Abdessamad (2019), "1000 Movie Reviews (Review + Attached rating + Sentiment polarity) for Reputation Generation", Mendeley Data, V1, doi: 10.17632/38j8b6s2mx.1

[13] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

[14] Dornbusch, R., & Reynoso, A. (1989). Financial factors in economic development.

[15] Grootendorst, M. (28 October 2020). Keyword Extraction with BERT. Maartengrootendorst. https://www.maartengrootendorst.com/blog/keybert/