Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 11, No. 1, 2023

Research on Emotional Analysis of Online Book Reviews Based on Word2Vec Method

Jingxiu Xu[1, *], Albert A. Vinluan[2]

[1] Graduate School, ANGELES University Foundation, Angeles 2009, Philippines
[2] Graduate School, ANGELES University Foundation, Isabela State University-Echague Campus, Angeles 2009, Philippines
[*] Corresponding author: Jingxiu Xu (Email: xu.jingxiu@auf.edu.ph)

Abstract: At present, emotional analysis in the field of natural language processing is gradually becoming a popular research direction on social media. With the rise of artificial intelligence technology and deep learning algorithms, word vector representation methods based on neural networks have emerged and emerged in the field of sentiment analysis. According to the current research results of experts, the Word2Vec model has become the mainstream word vector representation method. The algorithm model is based on continuous optimization and iteration of random initialization word vectors, ultimately obtaining a stable performance word vector representation. If the research goal is to obtain sentence vectors, weighting processing is also required. Therefore, after systematically summarizing and analyzing the representation techniques of initialized word vectors and solving sentence vectors through word vector weighting, this article proposes corresponding improvement methods. This article improves the initialization algorithm of its word vector based on the Word2Vec model. On the basis of the original randomly generated word vector, the correlation information of words is added to change the distribution characteristics of the initial word vector, thereby enhancing the performance of the word vector. In the comparative experiment of the user comment dataset, the improved model showed a 1% to 3% improvement in sentiment polarity classification accuracy compared to the other model.

Keywords: Enter key words or phrases in alphabetical order, Separated by commas.

1. Introduction

With the widespread use of information technology in recent years, a large number of comments and exchange information involving users have emerged on e-commerce platforms. By browsing this information, ordinary users can browse the opinions, emotional attitudes, and product quality information expressed by purchased users at any time. With the increase of comment users, the number of comment information at various levels has also rapidly increased. According to calculations by International Data Corporation, the global digital economy will continue to maintain rapid growth in the coming years, and it is expected that by 2025, over 60% of global GDP will be contributed by the digital economy. This requires computers to help users quickly organize and organize relevant information. Comment text mining has emerged in multiple research directions, including information filtering, named entity annotation, sentiment analysis, text classification, and text clustering. Among them, sentiment analysis technology has served most fields and is committed to discovering valuable information with practical significance from text information, bringing convenience to people's lives.

2. Research Methods

2.1. Vectorized Representation of Text

Natural language has certain grammatical and structural information, composed of words, and is the basic language for human communication and understanding. In NLP, computers cannot directly recognize characters such as English, so it is necessary to convert text characters into word vector forms that computers can recognize, that is, digitization. Secondly, reflecting the semantic correlation between words through word vectors is also a difficulty in text processing. This chapter introduces several methods for generating word vectors, and explains the principles and advantages of these techniques.

2.2. One Hot Representation

The single hot representation is the simplest representation method for word vectors. Each word in the standard vocabulary is numbered sequentially, and the number of words in the vocabulary is the dimension of the word vector. The vector representation method is: the numerical value at the corresponding position is taken as 1, and the other positions are taken as 0. Assuming the length of the text is N, that is, the dimension of the word vector is N, and the position index of the words is stored as encoding. The vector of the nth word is represented as [0,0,0,..., 1], indicating that the vector space distance of each two words is equal. The word vector represented by One hot is independent between words, and the grammatical and semantic connections between words cannot be seen solely from the two vectors. This independence is not suitable for the semantic expression of sentences. Secondly, the dimension of the word vector will increase with the increase of the vocabulary size, causing a dimension explosion, which not only results in sparse numerical dispersion, but also greatly consumes computational resources. One hot indicates that it clearly does not meet the needs of large-scale corpus.

2.3. CBOW Model

Use $C_1, C_2, \dots, C_k$ to represent k observation attributes, the sample to be tested is represented by $y = (m_1, m_2, \dots, m_k)$, and the category set is represented by $B = \{b_1, b_2, \dots, b_n\}$.

The idea of CBOW model is to optimize the probability of the target word. In the design of network structure, a three-

layer neural network model consisting of input layer, hidden layer, and output layer is used to infer the probability of the occurrence of intermediate words based on context, that is, $P(m_i|\ text(m))$. Assuming that the context consists of k words m and the intermediate word is m_i, the CBOW model predicts the probability of m_i using 2k words in the context. Among them, the input layer is not a concatenation of contextual words, ignoring word order information and using the average of all word vectors, ignoring the hidden layer. The input layer is directly connected to the output layer, adding the sum of the input word vectors in the projection layer, and using contextual words to predict the central word (Gao, 2021; Hou et al., 2021). In the CBOW model, the word vector L is the only neural network parameter (Xie, 2020; Li et al., 2020; Li & Xiong, 2017; Cuevas Molano et al., 2021) to maximize the logarithmic likelihood of all words as shown in the formula:

$$L^* = \underset{L}{\operatorname{argmax}} \sum_{m \in N} logP(m_i \mid text(m))$$

2.4. Skip Gram Model

The Skip gram word vector model is similar to the CBOW word vector model and an improved neural network model. However, the Skip gram model adopts the opposite network structure design method, which infers the probability of words in the context of the word based on the current central word. It allows certain words to be "skipped", unlike the CBOW word vector model, which must be done in order, making it more flexible, Compared with the CBOW word vector model, Skip Gram also has better robustness.

Assuming the current word is m_i. The word set containing 2K words in the context is composed of $m_{i-k}, m_{i-k+1}, \dots, m_{i+k}$, and the intermediate word used in the Skip gram model is m_i, which can deduce the probability of a certain word in the context in the word set $m_{i-k}, m_{i-k+1}, \dots, m_{i+k}$, which is $P\ (text(m)|\ m_i)$. The final maximum logarithmic likelihood of all contextual words is shown in the formula:

$$L^* = \underset{L}{\operatorname{argmax}} \sum_{m \in N} logP(text(m) \mid m_i)$$

2.5. Word2vec Model

Word2vec is an open source word vector modeling tool released by Google in 2013. The One hot representation of words is represented by a mapping matrix as a low dimensional dense vector, which is then used as input to the language model. Finally, the optimal network parameters are obtained through training. Word2vec and other word vector models are based on the same assumption: if adjacent words "recognize" each other, it indicates that they have similarity between words, which is called "distance similarity" in linguistics. Word2vec can simplify text words into K-dimensional real number vectors in vector space through training, and the distance between vectors in vector space can reflect the semantic relevance of the text (Yilmaz & Toklu, 2020; Yang et al., 2020). This method solves the traditional problem of sparse text vectors and reduces the training difficulty and time of the model. Therefore, Word2vec technology is often used for NLP related tasks such as clustering, finding synonyms, and part of speech analysis. Word2vec technology is a model proposed based on CBOW and Skip gram word vector models. The basic idea is to treat the document as a paragraph of a larger document, add corresponding ID identifiers, and continuously update them during the training process, for example, for book review d_i, where i is the identifier of the corresponding review document d. Figure 1. and Figure 2. are schematic diagrams of using CBOW word vector model and Skip Gram word vector model to train text vectors, respectively.

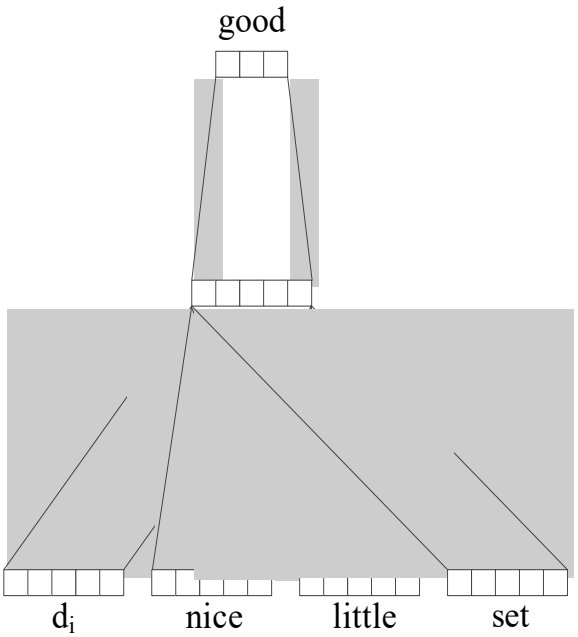

Figure 1. CBOW Word Vector Model Training Text Vector Graph

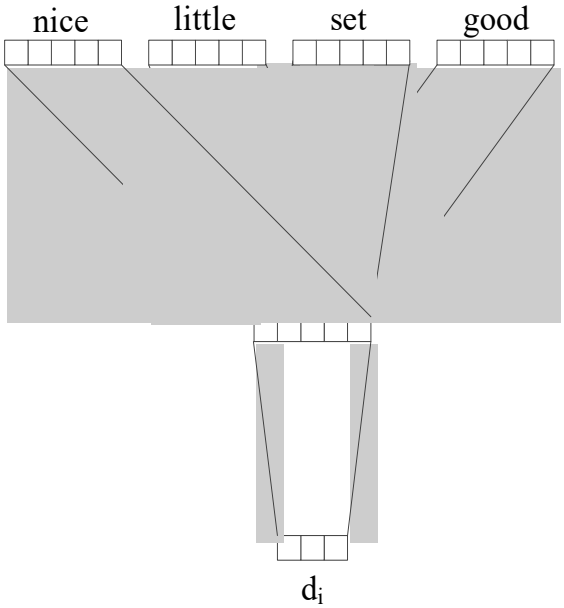

Figure 2. Skip Gram Word Vector Model Training Text Vector Graph

The Word2vec program successfully achieved direct training of text vectors. Some important parameters and corresponding functions of the Word2vec program are shown in Table 1. The research conclusion proves that the optimized Word2vec program in this article analyzes the similarity of word vector features in book review texts, that is, texts with similar meanings have similar distances in their vector spaces.

Table 1. Word2vec Program Parameters Table

Parameter Name	Function
sentences	Used to specify a training corpus list, large corpus can be constructed using BrownCorpus, Text8Corpus, or LineSentence
sg	Set the training algorithm with a default value of 0 and use the CBOW algorithm; Sg=1, using skip gram algorithm
size	The dimension of the feature vector, with a default value of 100
window	What is the maximum distance between the current word and the predicted word in a sentence
alpha	Learning Rate
seed	randomizer
min_count	Dictionary truncation value, default value is 5, word frequency is less than min_ Words with count count count will be discarded
max_vocab_size	The RAM limit during word vector construction. If the number of independent words exceeds this, the least frequent one will be eliminated. A value of none indicates no limit
sample	Configuration threshold for random downsampling of high-frequency vocabulary
workers	Control the parallel number of text training
iter	Iteration count, default to 5
batch_words	The number of words passed to the thread in each batch, default to 10000

The Word2vec program sets a corresponding text ID tag for each text in the text set. Some scholars have found that the accuracy of text features is related to the amount of text corpus. The more text corpus, the higher the accuracy. In addition, by repeatedly training randomly scrambled text, the accuracy of text features can also be improved. When training the book review corpus in this article, in order to improve accuracy, all book reviews are set as the training corpus. The specific parameter names and values for other parameter settings are shown in Table 2., and the 10 most relevant words to the word 'good' are output, sorted by relevance as shown in Table 3.

Table 2. Word2vec parameter settings for book reviews

Parameter	Parameter value
sentences	text
sg	1
size	200
window	5
alpha	0.025
min_count	2
workers	3

Table 3. The 10 most relevant words and their correlation with 'good'

Word Name	Relevance
decent	0.801
great	0.799
cool	0.799
goodthis	0.795
starsthis	0.784
shortthis	0.780
originality	0.780
k	0.779
readit	0.779
storythe	0.776

3. Literature References

Emotional analysis is an emerging research topic that has been studied by many scholars at present. In 2022, Dharma et al. adopted a neural network embedded in the Word2Vec algorithm to capture semantics, and found that it had the best accuracy in the process of text classification; Based on 4884 articles published since the establishment of RCR. In 2023, Zhu et al. utilized natural language processing and Word2Vec technology to identify potential patterns, connections, and interactions between research topics and topic vectors. In 2022, Styawati et al. conducted sentiment analysis on comments on online transportation service apps using word2vec text embedding algorithm and support vector machine (SVM) algorithm. Word2vec is used to extract text features and represent words in vector form. Construct the Word2vec model using the skip gram model. The Support Vector Machine (SVM) algorithm is used for text data classification to emotionally assess the accuracy of the text data used. In 2023, Ma et al. cleverly combined latent

Dirichlet distribution (LDA) and Word2vec to generate a thematic evolution map from global to local corpus, discovering and revealing the multi-level information evolution of themes. This includes the distribution trend of topics throughout the entire time series, as well as the merging and splitting of semantic information between adjacent time topics. Revealed the correlation between themes and the entire lifecycle of their emergence, development, maturity, and decline.

4. Conclusion

To evaluate and test the accuracy and prediction performance of the model, Precision, Recall, and F1 were used to evaluate the classification and prediction performance of the model. A ten fold crossover method was used to evaluate the classification and prediction performance. The review book dataset Book_review is divided into 10 parts, and the training and testing sets are divided in an 8:2 ratio to obtain the Precision, Recall, and F1 values of the experimental model. The average of these values is used to evaluate the classification and prediction performance of the model.

Through experiments on the comment dataset, it was verified that the model has good predictive performance. The validated model methods include linear support vector machine (linear SVM), decision tree (DT), and word2vec classification method, which are included in the comment set book_ Conduct experiments on View to verify and compare the effectiveness of three different model methods in determining emotional tendencies. The experimental results are shown in Table 4.

Table 4. Accuracy of different classification methods

Model	Precision	Recall	F1
LinearSVM	0.81	0.81	0.81
DT	0.73	0.73	0.73
Word2Vec	0.825	0.83	0.83

From Table 4 above, it can be seen that the Word2Vec model algorithm has the best performance in sentiment classification and prediction on dataset Book_review, with the highest accuracy, recall, and F1 values. Therefore, in the study of user comment sentiment analysis, the Word2Vec algorithm is the most suitable choice.

References

[1] Dharma, Eddy Muntina, et al, "The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification," J Theor Appl Inf Technol, vol. 100, no. 2, pp. 31, Jan 2022.

[2] Zhu, Jun-Jie, Zhiyong Jason Ren, "The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining." Resources, Conservation and Recycling, vol. 190, pp. 106876, 2023.

[3] Styawati, Styawati, Andi Nurkholis, Ahmad Ari Aldino, Selamet Samsugi, Emi Suryati, and Ryan Puji Cahyono. "Sentiment analysis on online transportation reviews using Word2Vec text embedding model feature extraction and support vector machine (SVM) algorithm." In 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), pp. 163-167. IEEE, 2022.

[4] Ma, Jian, et al. "An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local," Expert Systems with Applications, vol. 212, pp. 118695, 2023.

[5] Y. Gao, "Research and Application of Emotional Analysis Based on Word2Vec Method (Master's Thesis," Xiamen University, 2019.

[6] M. Zhou, "Application of sentiment analysis based on LSTM and Word2Vec in product reviews, " Statistics and Management, vol. 12, pp.81-84, 2019.

[7] X. Yan, "A Study on Emotional Analysis of Weibo Stock Review Text Based on Word2vec and LSTM," Nanjing University, 2018.

[8] X. B. Peng, "Research on Text Emotion Analysis Method Based on Word2vec," Network Security Technology and Application, vol. 07, pp.58-59, 2016.

[9] M. B. Li, W. Wang, C. C. Wang, "Exploring the Emotional Analysis Method for Online Teaching Evaluation in Higher Vocational Education Based on Word2Vec and Bi GRU, " Journal of Guangdong Water Resources and Electric Power Vocational and Technical College, vol. 21, no. 03, pp. 58-59, 2023.

[10] Y. Z. Wang, T. Wang, "Research on Emotional Analysis Method Based on Weighted Word2Vec XGBoost, " Journal of Chongqing University of Science and Technology (Natural Science Edition), vol. 25, no. 03, pp. 62-66, 2023.

[11] M. Zhou, "Application of sentiment analysis based on LSTM and Word2Vec in product reviews," Statistics and Management, vol. 12, pp. 81-84, 2019.