Text vectorization in sentiment analysis: A comparative study of TF-IDF and Word2Vec from Amazon Fine Food Reviews

Jiaxin Lu*

ECS, University of Southampton, SO16 1BJ, United Kingdom

Abstract. Sentiment analysis is a practical tool for marketing and branding teams. Companies can collect and analyze opinions or reviews from social media platforms, blog posts, and other numerous forums. It may help them acquire positive feedback to reinforce strengths or identify negative emotions to make improvements. The research is to compare two text vectorization methods in opinion mining: Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec, using Amazon Fine Food Reviews dataset. This study will use these two methods to vectorize preprocessed text data and also input the vectorized data to the emotion classification model, analyzing the performance of two methods in the emotion classification task. The consequence indicates that the former outperforms the latter in handling large datasets, particularly in distinguishing between different sentiment categories, but latter is superior in capturing the semantic relationship of words. Therefore, it is suggested that the advantages of the two methods be combined in practical applications to improve the accuracy and efficiency.

1 Introduction

With the development of e-commerce, consumer reviews and ratings have become an important basis for customers to understand product quality and purchase decisions. As a branch of natural language processing (NLP), sentiment analysis is used in machine learning to analyze and classify the emotional tone of textual data. It can extract users' emotional tendencies to provide valuable feedback for businesses in the amount of text data[1].

In recent years, the TF-IDF method and the Word2Vec model have been widely used in sentiment analysis. Singh et al. applied the TF-IDF method to perform sentiment analysis on the Twitter dataset on May 26-27, 2022[2]. They combined the TF-IDF algorithm with a variety of machine learning algorithms, including support vector machines (SVM), extreme gradient boosting (XGBoost), and Random Forest (RF) to classify the sentiment of tweets. The results of the study show that the SVM classifier achieves 83.74% which is the highest accuracy, and also has the best performance among the tested algorithms. Additionally, the SVM and F1-score also recorded the highest precision and recall, was meant that when combined with TF-IDF, SVM is extremely suitable for emotional analysis of social media

* Corresponding author:jl29u23@soton.ac.uk

data, because of high accuracy and reliability. It also emphasizes the significance of selecting an appropriate machine learning algorithm to combine with TF-IDF in social media sentiment analysis. Ram Krishn Mishra et al. discussed the application of the TF-IDF method in sentiment analysis of hotel reviews on December 11–12, 2019[3]. They used a dataset of 515,000 hotel reviews from across Europe to analyze both positive and negative feedback provided by users, discovering that the combination of TF-IDF and cosine similarity provides a reliable recommendation and significantly improves the accuracy of the system. The findings show that TF-IDF is immensely effective and useful at extracting relevant terms from reviews, particularly in the travel and hospitality industry.

Syamsul Rizal et al. conducted sentiment analysis on movie reviews from Rotten Tomatoes using the Word2Vec model[4]. They combined Word2Vec with a naive Bayes classifier to analyze a dataset of 45,739 movie reviews. Reviews labeled positive emotions as' fresh 'and negative emotions as' rotten'. Then, the researchers used a skid-Gram variant of Word2Vec to generate word embeddings which train naive Bayes classifiers[5]. They discovered that the model is able to effectively capture the semantic relationship between words, and the overall classification accuracy reaches 72.23%, which means that Word2Vec is useful for extracting meaningful features, especially in the context of film criticism. In addition, Farhan Wahyu Kurniawan et al. collected data on tweets related to the Indonesian disaster from Twitter and trained the Word2Vec model using the Indonesian-language Wikipedia as a large corpus[6]. They opted for two Word2Vec model architectures: Continuous Bag–of–Words (CBOW) and skip-gram[7], and put the generated word vectors into an SVM classifier for sentiment classification. The study demonstrated that when Word2Vec is processing Indonesian tweets, it performed particularly excellently in the emotion classification task.

Nevertheless, their performance is still insufficient on large-scale review data sets, especially in the comparison of the advantages and disadvantages of different sentiment categories. Consequently, the significance of this study is to fill this research gap. This paper offer a valuable suggestion about the relative strengths and weaknesses of the two methods through a detailed comparison of the large-scale datasets. Understanding these differences is immensely important to improving techniques of sentiment analysis, particularly for dealing with large volumes of textual data efficiently and accurately, such as in social media platforms (Twitter, Facebook). The objective of this study is to evaluate the performance of these two methods in different categories of emotion, as well as computational efficiency when they are handling a large amount of data. This study used the Amazon food review data set to perform a visual comparison by principal component analysis (PCA). Compared with existing research, the findings are expected to offer valuable insights to practitioners and researchers in the field of sentiment analysis, helping them select reasonable methods in different application scenarios.

2 Research design

2.1 Dataset description

The research utilizes the Amazon Fine Food Reviews dataset, which consists of 568,454 reviews from October 1999 to October 2012. The dataset includes product and user information, ratings, and detailed text reviews. A total of 256,059 users commented and covered 74,258 products. Among these users, 260 individuals have posted more than 50 reviews in the dataset. It also includes other Amazon category reviews.

2.2 Distribution of User Reviews Sentiment Polarity

The reviews are scored from 1 to 5 points. To standardize the classification of emotions, reviews with scores of 5 are classified as positive, 1 to 2 as negative, and 3 to 4 as neutral. The distribution of each review across these categories is summarized as follows: there are 443,777 positive reviews, 82,037 negative reviews, and 42,640 neutral reviews, accounting for 78.07%, 7.50%, and 14.43% of the total reviews, respectively, as shown in Table 1.

Table 1. The distribution of positive, negative, and neutral reviews.

Category	Count	Proportion
Positive reviews	443777	78.07%
Negative reviews	82037	7.50%
Neutral reviews	42640	14.43%

2.3 Data Pre-processing

The section will detail the steps to pre-process the raw data before applying any machine learning model or analysis technique. The key steps involved in this process are detailed below.

2.3.1 Text processing

Text Cleaning is a crucial pre-processing step that involves multiple operations to clean and standardize textual data[8]. Initially, it removes the special characters, numbers, and punctuation, which removes all non-alphabetic elements, leaving only letters and spaces, helping us eliminate noise such as punctuation marks and numerical values. Then, it converts all letters to lowercase so that ensure consistency in the text, and avoid words like "Excellent" and "excellent" being treated as separate entries to reduce redundancy. Also, the cleaned text is marked as single words, and common words are removed, known as stopwords (e.g., "the," "and," "is"). This approach focuses on the more meaningful content relevant to sentiment analysis. Following this step, the remaining words are stem-extracted, using tools like Porter's stem algorithm to reduce the words to their root form. This approach not only simplifies the vocabulary but also the model to enhance its ability to understand and recognize patterns and trends in the text. Consequently, by unifying different forms of the same word into the root form, the dimensionality of data is significantly reduced, making subsequent text vectorization efficiently.

2.3.2 Vectorization Methods

After the text cleaning is completed, it must be ensured that the text data need to be converted into a numerical form for further analysis and model training, which is known as vectorization. As mentioned before, this study employed two widely used vectorization methods.

- TF-IDF

Term Frequency Inverse Document Frequency of records is used to evaluate the importance of words in a document based on their relative frequency across a collection of documents. It could help us highlight significant words and reduce the influence of common words by balancing frequency of a word appearance within a specific document (term

frequency, TF) with its rarity across the corpus as a whole (inverse document frequency, IDF)[9].

Term Frequency (TF): Term Frequency measures the frequency of a term within a single document. Generally, it reflects the importance of a term in a particular document, indicating its relevance in that context. It can be defined as -

$$\mathrm{TF(t,d)} = \frac{\text{Number of occurrences of t in document d}}{\text{Total number of terms in document d}} \quad (1)$$

Inverse Document Frequency (IDF): Inverse Document Frequency is a measure of the importance of words. A higher value of IDF represents the word is less common and more important than when it appears in a particular document.

$$\mathrm{IDF(t)} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term t}}\right) \quad (2)$$

TF-IDF Value: The importance of words in a specific document is determined by combining these two factors, and the calculation method is:

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \quad (3)$$

- **Word2Vec**

Word2Vec is a neural network-based technology widely used in the field of natural language processing (NLP) that converts words into numerical vectors, enabling machines to understand the contextual meaning of words[10]. Different from TF-IDF, Word2Vec mainly concentrates to capture the meaning, semantic similarity, and relationships with surrounding text.

Skip-gram: In this project, Skip-gram model was used to train Word2Vec. Skip-gram is a technique for creating word embeddings that focuses on predicting surrounding words based on a specific word, which is called the "target word". It works especially effectively for large data sets and handles uncommon words well. The formula and diagram illustrate the working mechanism[11].

$$J_\theta = \frac{1}{T}\sum_{n=1}^{T}\sum_{-n<j\leq n, j\ll 0} logp(\omega_{j+1}\,|\omega_t) \quad (4)$$

where T represents the total word count in the corpus, n denotes the context window size, ω_t is the target word, and ω_{j+1} are context words that surround the target word..

The goal of skip-gram function is to aggregate the logarithmic probabilities of contextual words located on either side of the target word to generate the following objective function, as shown in Fig. 1.

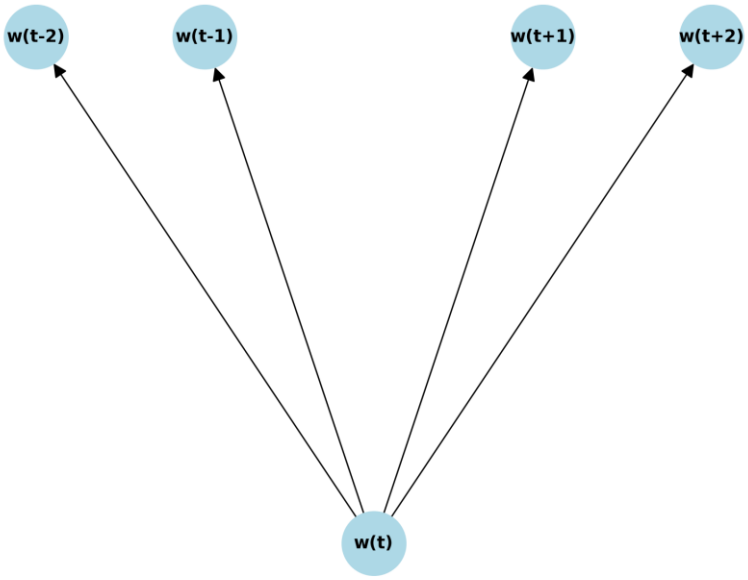

Fig. 1. The Skip-gram architecture augments the likelihood of determining context words derived from a particular target word.

2.3.3 Principal Component Analysis (PCA)

Principal Component Analysis is a machine learning method primarily used for dimensionality reduction and data visualization [12]. It simplifies complex data into fewer dimensional forms, making visualization and analysis easier while preserving important information as much as possible.

The study applied PCA to reduce the dimensionality of these vectors after converting the text data into vectors using Word2Vec and TF-IDF. This dimensionality reduction allowed us to visualize the data in two dimensions, making it easy to analyze reviews with different ratings. By leveraging PCA, the complexity of the high-dimensional vector data could be effectively managed, and meaningful patterns and insights were extracted.

During PCA dimensionality reduction, this study treated the comment text vector separately for each score intervals. By PCA transformation, the position of each review in the new coordinate system was obtained. Then, the PCA results were added to the data frame and used for subsequent scatter plot visualization and analysis significantly.

3 Results And Discussion

The section presents and compares results of PCA visualizations of text vectors using Word2Vec and TF-IDF vectorization methods. These visualizations help us understand the distribution of reviews with different sentiment scores (1-5) in each methods.

3.1 PCA Results for Word2Vec Vectorization

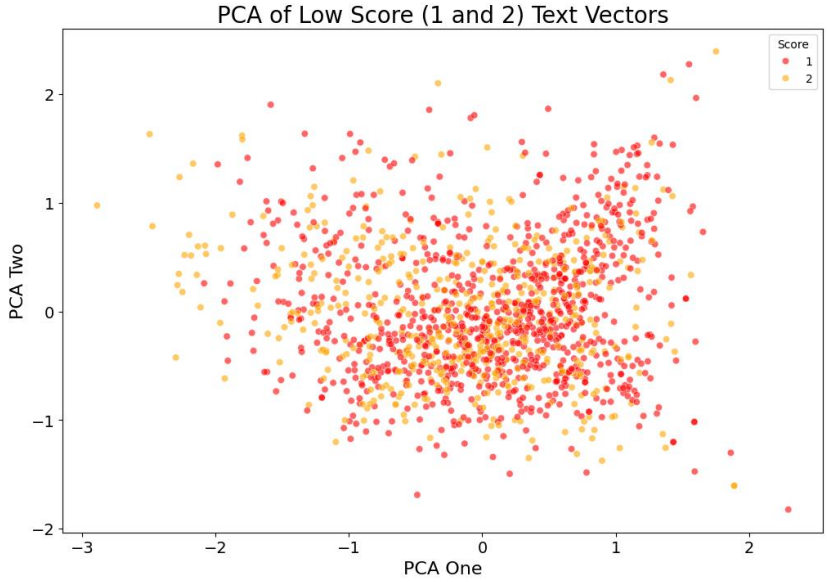

Fig. 2. PCA of Low Score (1 and 2) Text Vectors (Word2Vec).

This diagram illustrates the PCA scatter plot for reviews with scores 1 and 2, represented by red and orange points respectively, as shown in Fig. 2. The dispersion of the points indicates that the negative comments are semantically diverse, suggesting that reviewers express negative sentiments in a variety of ways.

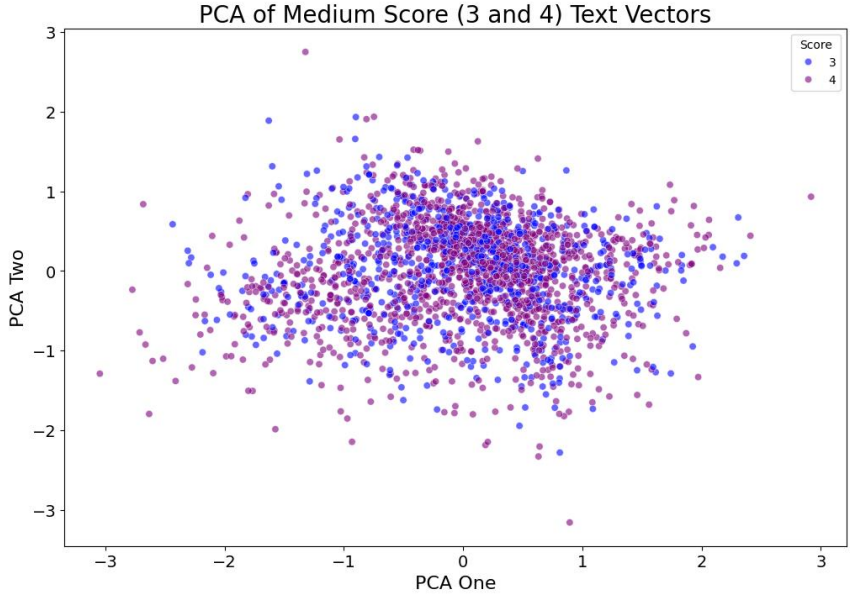

Fig. 3. PCA of Medium Score (3 and 4) Text Vectors (Word2Vec)

This diagram illustrates the PCA scatter plot for reviews with scores of 3 (purple points) and 4 (blue points), as shown in Fig. 3. The points are more concentrated, which means that neutral comments are more consistent in tone.

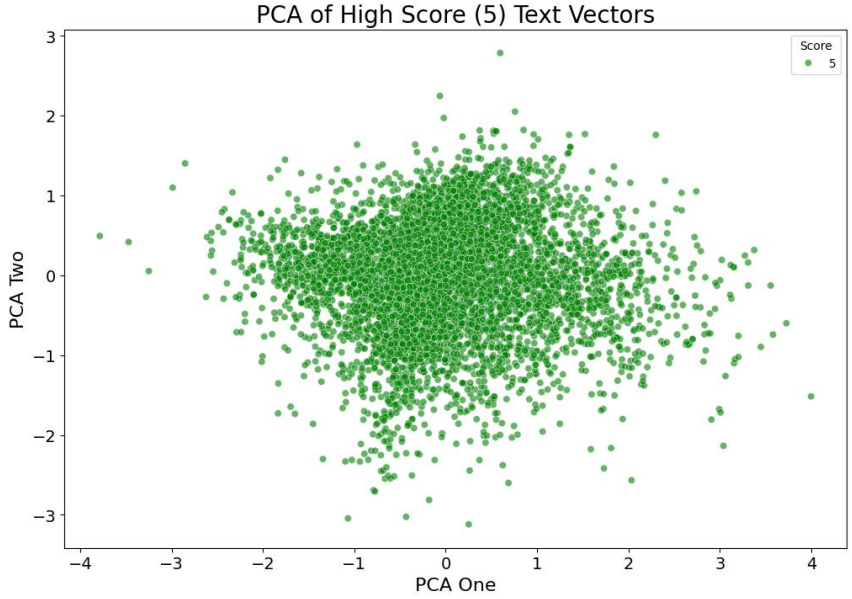

Fig. 4. PCA of High Score (5) Text Vectors (Word2Vec)

This diagram displays the PCA scatter plot for reviews with a score of 5 (green points), as shown in Fig. 4. The high score reviews are highly concentrated, giving a high degree of semantic consistency. The positive reviews tend to use similarly positive expressions, such as "excellent," "fantastic," "amazing," and "perfect."

3.2 PCA Results for TF-IDF Vectorization

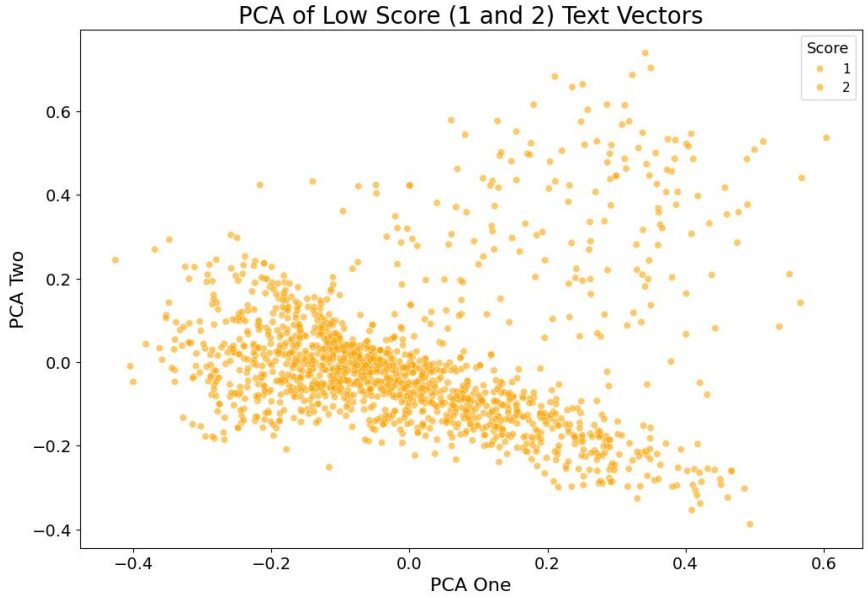

Fig. 5. PCA of Low Score (1 and 2) Text Vectors (TF-IDF).

This diagram represents the PCA scatter plot for reviews with scores of 1 and 2, as shown in Fig. 5. Compared with diagrams of Word2Vec, the points show a dispersed distribution with less overlap, mainly emphasizing the level of importance of different words.

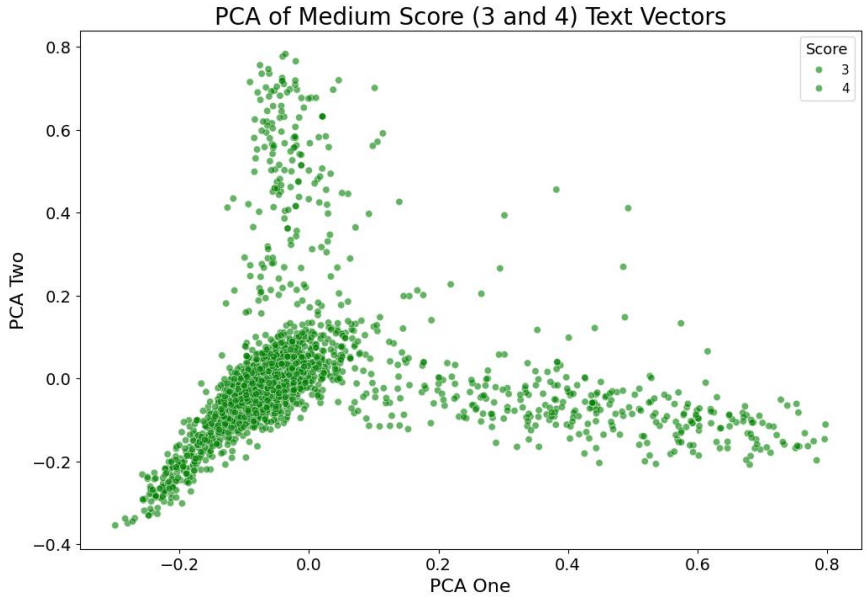

Fig. 6. PCA of Medium Score (3 and 4) Text Vectors (TF-IDF).

This diagram shows the PCA scatter plot for reviews with scores of 3 and 4, as shown in Fig. 6. The points are more concentrated with some overlap, reflecting a consistent emphasis on word importance among neutral reviews.

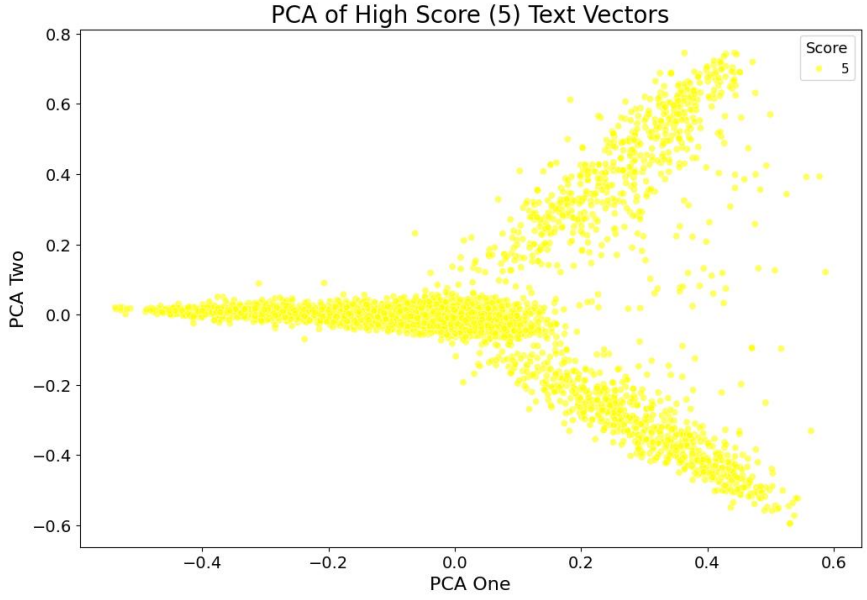

Fig. 7. PCA of High Score (5) Text Vectors (TF-IDF).

This diagram illustrates the PCA scatter plot for reviews with a score of 5 (yellow points), as shown in Fig. 7. The points indicate the frequent use of significant positive words because the high-score reviews are closely clustered.

3.3 Comparative Analysis of Word2Vec and TF-IDF Models

Firstly, it could be observed that the Word2Vec model is outstanding at representing the semantic connections between words from the PCA visualization. This would enable it to grasp the context and nuanced meaning of comments, especially in high-rated comments, where positive emotions can be consistently grouped. The model showed a more fragmented distribution in both low-rated and medium-rated reviews, which means that it was able to handle multiple expressions and capture nuances in negative and neutral reviews. However, the scattered distribution of low and medium ratings may also be a drawback, because it is difficult to clearly distinguish subtle emotional differences in these categories. In addition, the paper concluded that it requires more computational resources and data support to train and implement the Word2Vec model.

By contrast, the TF-IDF model demonstrates clearer and more focused clustering across all emotionally rated reviews, especially among high-rated reviews. TF-IDF can more effectively capture and reflect emotional differences in comments by emphasizing the frequency and importance of words. Furthermore, this method is very simple to implement, and barely requires fewer computational resources, making it suitable for smaller datasets. Nevertheless, TF-IDF cannot capture semantic relationships between words, so it may cause a lack of contextual information. In addition, TF-IDF extremely relies on the frequency of words, as a result, it may ignore certain subtle emotions.

To quantify these observation further, this study calculated the variance of data points across first two principal components (PCA1 and PCA2) of each sentiment score category (low, medium, and high) [13]. PCA1 represents the direction of largest variance of data, while PCA2 represents the largest portion of the remaining variance. Variance is calculated by the average of squared deviation from its mean, which reflects distribution of data point. A lower variance represents the data points are more concentrated, indicating that the text vectors are more consistently within the same emotion category. A higher variance represents the distribution is more dispersed, meaning the model can recognize more different semantic details but the boundaries between emotional categories may become blurred. Table 2 and Table 3 present the values of variance for each model.

Table 2. Variance of PCA Components for TF-IDF

Sentiment Scores	PCA1 Variance	PCA2 Variance
Low Score (1-2)	0.0364	0.0300
Medium Score (3-4)	0.0394	0.0299
High Score (5)	0.0349	0.0306

Table 3. Variance of PCA Components for Word2Vec

Sentiment Scores	PCA1 Variance	PCA2 Variance
Low Score (1-2)	0.6474	0.4116
Medium Score (3-4)	0.7005	0.4068

High Score (5)	0.8098	0.4699

According to Table 2 and Table 3, it is not difficult to see that the values of variance of PCA1 and PCA2 in the TF-IDF model were lower in all scoring intervals, which means that the text vectors generated by TF-IDF model have higher consistency and similarity within the same emotion category, so TF-IDF can easily to distinguish the different emotion categories in classifying sentiment. Compared with TF-IDF model, Word2Vec had significantly higher values of variance for PCA1 and PCA2 across all score ranges, indicating that Word2Vec model is able to deal with more complex and diverse expressions of emotion, but it may not be as accurate as TF-IDF in classifying emotional categories.

4 Conclusion

In summary, the study discovered that TF-IDF is superior to Word2Vec in terms of both accuracy and overall classification performance when handling large-scale review data, although both models have their advantages. TF-IDF is more significant in terms of capturing textual features and differentiating comment categories with different ratings, especially emphasizing emotional differences in comments. In contrast, Word2Vec is more excellent at capturing semantic relationships, especially dealing with positive emotional comments, but it needs high computational resources and it is reasonable for larger data sets. These findings have important implications for the field of sentiment analysis, particularly in applications such as e-commerce platforms and social media, providing a practical reference for selecting appropriate text to quantify techniques. It is helpful for researchers and engineers to better select and optimize emotion analysis models in practical applications. Additionally, it is suggested that future research could focus on combining the benefits of two methods to create a more significant model, which can efficiently extract text features while capturing deep semantic relationships.

References

[1] Denis Eka Cahyani and I. Patasik, “Performance comparison of TF-IDF and Word2Vec models for emotion text classification,” Bulletin of Electrical Engineering and Informatics, vol. 10, no. 5, pp. 2780–2788, 2024, Accessed: Aug. 15, 2024. [Online]. Available: https://www.beei.org/index.php/EEI/article/view/3157/2341.

[2] S. Singh, K. Kumar, and B. Kumar, “Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques,” 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), May 2022, doi: https://doi.org/10.1109/com-it-con54601.2022.9850477.

[3] Ram Krishn Mishra, Siddhaling Urolagin, and A. Arul, “A Sentiment analysis-based hotel recommendation using TF-IDF Approach,” 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE Xplore, Dec. 2019, doi: https://doi.org/10.1109/iccike47802.2019.9004385.

[4] S. Rizal, Adiwijaya, and M. D. Purbolaksono, “Sentiment Analysis on Movie Review from Rotten Tomatoes Using Word2Vec and Naive Bayes,” 2022 1st International Conference on Software Engineering and Information Technology (ICoSEIT), IEEE Xplore, Nov. 01, 2022. https://ieeexplore.ieee.org/document/10030009.

[5] S. Manna and H. Nakai, “Effectiveness of Word Embeddings on Classifiers: A Case Study with Tweets,” 2019 IEEE 13th International Conference on Semantic Computing (ICSC), IEEE Xplore, Jan. 2019, doi:https://doi.org/10.1109/icosc.2019.8665538.

[6] Farhan Wahyu Kurniawan and Warih Maharani, “Indonesian Twitter Sentiment Analysis Using Word2Vec,” 2020 International Conference on Data Science and Its Applications (ICoDSA), IEEE Xplore, Aug. 2020, doi: https://doi.org/10.1109/icodsa50139.2020.9212906.

[7] B. Liu, “Text sentiment analysis based on CBOW model and deep learning in big data environment,” Journal of Ambient Intelligence and Humanized Computing, Oct. 2018, doi: https://doi.org/10.1007/s12652-018-1095-6.

[8] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, “The influence of preprocessing on text classification using a bag-of-words representation,” PLOS ONE, vol. 15, no. 5, p. e0232525, May 2020, doi: https://doi.org/10.1371/journal.pone.0232525.

[9] “Understanding TF-IDF (Term Frequency-Inverse Document Frequency),” GeeksforGeeks, Jan. 20, 2021. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[10] “Word Embeddings in NLP,” GeeksforGeeks, Oct. 11, 2020. https://www.geeksforgeeks.org/word-embeddings-in-nlp/

[11] “Papers with Code - Skip-gram Word2Vec Explained,” Paperswithcode.com, 2020. https://paperswithcode.com/method/skip-gram-word2vec

[12] Z. Jaadi, “A Step-by-Step Explanation of Principal Component Analysis,” Built-In, Feb. 23, 2024. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[13] L. Rydin Gorjão, G. Hassan, J. Kurths, and D. Witthaut, “MFDFA: Efficient multifractal detrended fluctuation analysis in python,” Computer Physics Communications, p. 108254, Dec. 2021, doi: https://doi.org/10.1016/j.cpc.2021.108254.