Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews

Zerui Zhan

Dublin International College, Beijing University of Technology, Beijing, 100000, China

Abstract. Sentiment analysis is an important area of natural language processing that supports applications such as market analysis, customer feedback, and social media monitoring by identifying and classifying opinions in text. Text representation is the basis of sentiment analysis, and TF-IDF and Word2Vec are two commonly used methods to carry out text vectorization by counting word frequency and capturing semantic relations respectively. This paper compares the performance of TF-IDF and Word2Vec in sentiment analysis of food reviews to provide a more effective basis for enterprises and researchers to choose text representation techniques. Based on 560,000 food review data, this paper focuses on comparing the accuracy and generalization ability of the two methods under different dataset sizes. The results showed that TF-IDF showed high accuracy in training data (99.16%), but showed obvious overfitting problems in test data (73.9%). In contrast, Word2Vec was more balanced on training and testing data (68.4% vs. 68.65%), showing better generalization. This finding has guiding implications for choosing text representation methods, especially in sentiment analysis tasks on large data sets.

1 Introduction

Sentiment analysis, also known as opinion mining, is a key area in natural language processing that involves identifying and classifying opinions in text. This area has important applications in various fields such as market analysis, customer feedback and social media monitoring. Opinion extraction technology helps businesses, governments, and organizations better understand user and public opinion by automating the extraction of valuable information from text to make smarter, more effective decisions. Text representation is a fundamental concept in natural language processing, which refers to the process of converting text data, such as words, sentences, paragraphs, or entire documents, into a numerical form that machine learning algorithms can process. This numerical representation can be a vector, matrix, or other mathematical structure that captures the features, content, and semantic information of the text Among the text representation techniques of sentiment analysis, word frequency-inverse Document frequency (TF-IDF) and Word2Vec are two prominent methods.

Corresponding author: zerui.zhan@ucdconnect.ie

TF-IDF is a statistical measure used to assess the importance of a word in a document relative to a collection of documents. The TF-IDF method reflects the importance of words only by frequency, so it is relatively simple, does not require too many resources, and can perform well on small data sets. At the same time, TF-IDF cannot understand the semantic relationship between words, so it has a weak grasp of meaning and context. Word2Vec, on the other hand, is a neural network-based approach that generates word embeddings that capture semantic relationships between words by mapping them to a continuous vector space. Word2Vec trains word vectors through neural networks, can effectively capture the semantic relationship between words and words, and can achieve some simple vector operations, which is impossible for TF-IDF. However, the Word2Vec model requires a lot of data and computational resources to generate high-quality word vectors, and it does not perform well on small corpora. These two methods still play an important role. For example, Liu et al. [1] published a study in 2022 that explored the application of weight allocation methods combining sentiment dictionaries with TF-IDF in text sentiment analysis, showing that TF-IDF still shows strong applicability in modern sentiment analysis tasks despite the emergence of many new text representations in recent years. Word2Vec, introduced by Mikolov et al. [2], has also gained significant attention. Their paper demonstrates Word2Vec's ability to capture semantic meaning and relationships, leading to its widespread use in a variety of NLP tasks, including sentiment analysis. TF-IDF and Word2Vec were compared in different contexts. For example, Kowsari et al. [3] compared several text representations for document classification, noting that although Word2Vec can capture semantic nuances, TF-IDF is competitive in some classification tasks due to its simplicity of implementation and interpretability. In addition, many studies have demonstrated that the size of the data volume directly impacts the performance of both methods. Tang et al. 's paper [4] explores scalable methods for sentiment analysis on large-scale datasets, including comparison of TF-IDF and Word2Vec, highlighting the effect of data volume on model performance. Zhao et al. 's paper [5] directly compared the performance of TF-IDF and Word2Vec under different data amounts, and concluded that TF-IDF had better performance when the data amount was small, while Word2Vec showed better semantic capturing ability when the data amount was increased. These studies highlight the need for focused comparisons between TF-IDF and Word2Vec in specific areas to understand their relative strengths and weaknesses fully.

The aim of this study is to conduct a comparative analysis of TF-IDF and Word2Vec techniques, specifically for sentiment analysis of food reviews. Given the different approaches of TF-IDF and Word2Vec in terms of text representation, this study sought to determine which approach would perform better in sentiment analysis of food reviews. In this study, the same logistic regression model was used for TF-IDF and Word2Vec representations, and the model parameters and sample data volume were adjusted to compare their accuracy and validity. The insights gained will highlight the strengths and weaknesses of each approach, guide the choice of text representation techniques for sentiment analysis tasks, and ultimately benefit the business through a deeper understanding of customer feedback.

2 Data and Methodology

2.1 Data

The dataset consists of 560,000 food-related reviews sourced from Kaggle. Each review includes a user rating and corresponding text comment, with some users providing multiple comments over time. Repeated comments are not considered an issue, and random sampling will be used in the experiment to minimize bias. This experiment only deals with the

relationship between reviews and ratings. Due to the dataset's large size and comprehensive coverage, it encapsulates a wide range of food evaluation scenarios, providing a robust representation of user opinions. The reviews are broad, truthful and reflect real consumer sentiment. The inclusion of a rating score for each review makes this dataset particularly suitable for machine learning classification tasks, facilitating efficient analysis. In addition, the average length of the reviews is moderate, making them representative of typical evaluative texts. The dataset's open and transparent nature also ensures the experiment's repeatability, which is why it was chosen as the sample for this study.

This study aims to compare the performance of TF-IDF and Word2Vec in sentiment analysis of food reviews. The process begins with an extensive data preprocessing phase, ensuring that the original text is properly prepared for analysis. First, the original censored text was rigorously cleaned up, where all characters were converted to lowercase letters and non-alphabetic characters (such as punctuation) were removed to maintain consistency in the dataset and reduce noise. Studies by Kumar and Chadha [6] show that these steps can reduce lexical redundancy and improve the accuracy of text conversion to vectors. All of these can improve the ability of the model to capture the main content of the text.

After that, pause words - common words like "and" or "the" that don't have significant meaning - are removed to focus on more meaningful words in the comments. In addition, morphemization is used to reduce words to their basic forms to ensure that different forms of the same word, for example, "running" and "ran" are treated uniformly. The cleaned text is then tokenized using the NLTK toolkit, breaking the text down into individual words or tokens, a key step in subsequent text representations. These pre-processing steps enable the comment text to be accurately and efficiently converted into token, making subsequent vectorization easier to achieve.

2.2 Methodology

For the vectorization stage, two different methods are used in this experiment: TF-IDF and Word2Vec. The pre-processed text is first converted to TF-IDF vectors using the tfidf format converter in the scikit-learn library. By setting the ngram_range parameter to (1,2), the transformation captures both a single word and a double graph, allowing the model to recognize important phrases beyond a single word [7]. This allows subsequent models to analyze and learn phrases, rather than just words, which improves the model's performance in text classification tasks such as sentiment analysis. In order to mitigate the effect of frequently occurring words, i.e., words whose actual semantic effect is not important, the experiment uses a sublinear term frequency scaling sublinear_tf=True, which adjusts the term frequency logarithmically to reduce the effect of terms that occur frequently in many documents.

At the same time, this experiment uses Word2Vec neural network model in Gensim library to generate word embedding. According to Rong [8], vector_size=100, window=5, min_count=2, workers=4 is a common and effective parameter configuration. This experiment takes this parameter setting, and each word in the comments is mapped to a 100-dimensional vector space to capture semantic relationships between words. To represent the entire comment, the vectors of all the words in the comment are averaged, resulting in a fixed-length feature vector that encapsulates the overall meaning of the comment.

The next stage is model training, using the feature vectors generated by TF-IDF and Word2Vec to train the logistic regression model. The tuning of hyperparameters (such as C, penalty and solver) through cross-validation is crucial for achieving the best performance of the model, and appropriate hyperparameter Settings can significantly improve the accuracy of emotion classification and F1 scores [9]. This experiment also uses cross-validation to fine-tune hyperparameters such as regularization intensity (C), penalty type and solution

algorithm (solver). The model's accuracy is improved to a certain extent, and then the comparison between TF-IDF and word2vec is carried out under this parameter.

According to Wang et al. [10], it is feasible to compare two different text representations by setting different parameters on the same machine learning model. This experiment comprehensively evaluates the performance of the model by calculating the accuracy on the training set and the test set, and compares the performance of the same model using TF-IDF and word2vec. This experiment generates detailed classification reports, including accuracy, recall, and f1 scores, and visualizes the results to directly compare the model performance comparisons achieved by TF-IDF and Word2Vec with the same hyperparameter Settings, as well as the variation curves of model performance with different parameter Settings. And through the control variable method to compare the specific details of the two methods.

3 Result

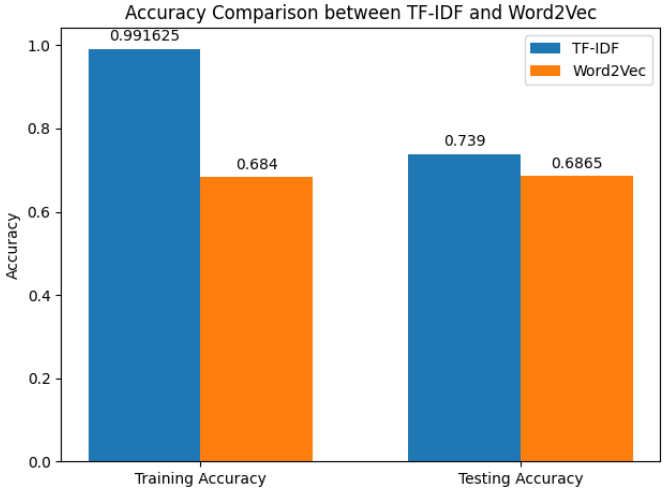

Fig. 1 Accuracy Comparison between TF-IDF and word2vec (Photo/Picture credit : Original)

The experimental results are shown in Fig. 1, which provides a comparative analysis of the training and test accuracy obtained by the logistic regression model represented by TF-IDF and Word2Vec text. As shown in Fig. 1, the model trained using TF-IDF achieved a very high training accuracy of about 99.16%, indicating strong performance on the training data. However, when applied to the test data, this accuracy dropped significantly to 73.9%, indicating a possible overfitting. In contrast, the model trained with the Word2Vec embed showed more balanced performance, with a training accuracy of 68.4% and a test accuracy of 68.65%. While lower than TF-IDF, the small difference between Word2Vec's training and test accuracy, as shown in Fig. 1 suggests a better generalization of previously unseen data. This comparison highlights the strengths and weaknesses of each approach, with TF-IDF excelling at capturing specific features in the training set, while Word2Vec provides improved generalization across different datasets.

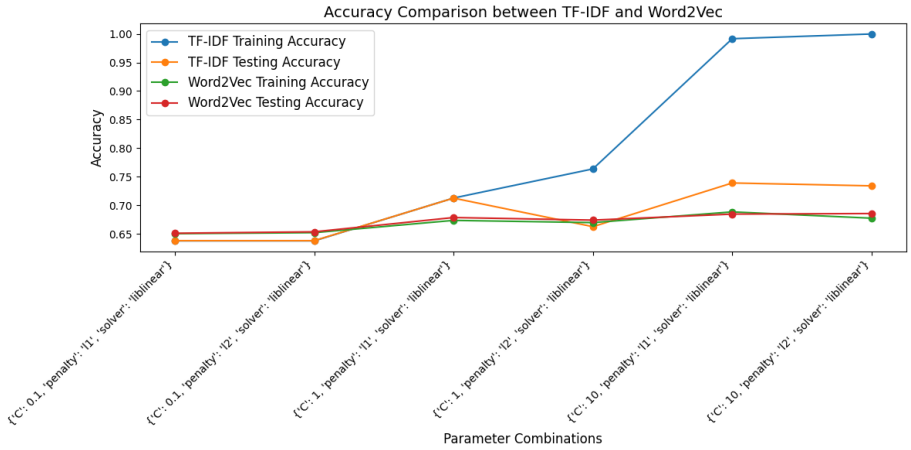

Fig. 2 Accuracy line with 10000 data (Photo/Picture credit : Original)

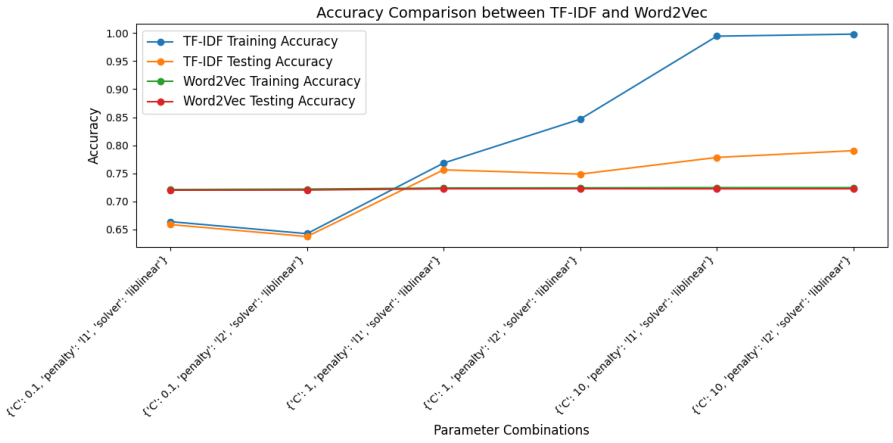

Fig. 3 Accuracy line with 50000 data (Photo/Picture credit : Original)

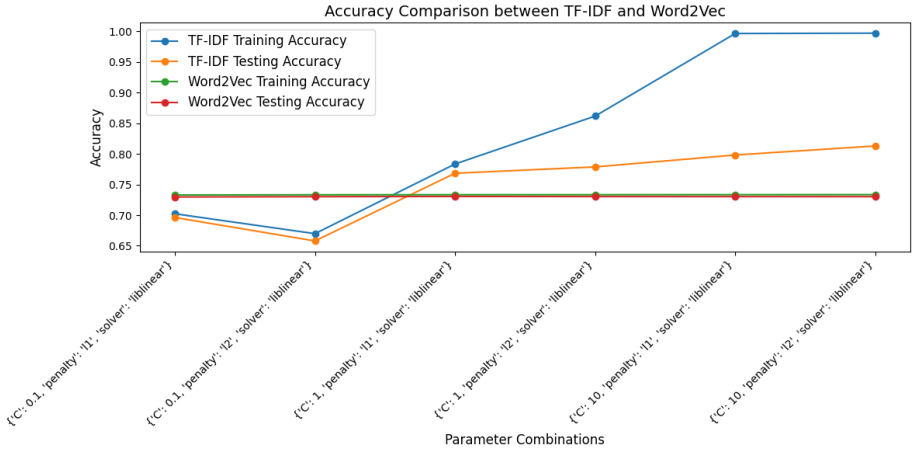

Fig.4 Accuracy line with 100000 data (Photo/Picture credit : Original)

The results shown in Fig. 2, Fig. 3, and Fig. 4 illustrate the accuracy of the comparison of TF-IDF and Word2Vec in different data set sizes (10,000, 50,000, and 100,000 samples,

respectively). These graphs provide insights into how each text representation performs when the sample size and hyperparameters of the logistic regression model change.

In Fig. 2, representing a dataset of 10,000 samples, TF-IDF shows a significant improvement in training accuracy, close to perfect, in different parameter combinations. However, the surge in training accuracy highlighted an obvious overfitting problem, as the corresponding test accuracy remained relatively flat at around 0.74. In contrast, Word2Vec performs more consistently, with training and test accuracy remaining close. While Word2Vec is less accurate overall, it generalizes better than TF-IDF when dealing with smaller data sets.

Fig. 3 represents the results of 50,000 samples, showing the trends observed in Fig. 2. The training accuracy of TF-IDF improves rapidly with the setting of some parameters, but the test accuracy does not improve correspondingly, which strengthens the overfitting problem. However, Word2Vec began to show slight improvements in training and test accuracy, suggesting that its ability to capture semantic relationships improved with larger datasets. The generalization of Word2Vec is still evident because the gap between training and test accuracy is small compared to TF-IDF.

Finally, in Fig. 4, the overfitting problem of TF-IDF becomes more obvious when the maximum dataset size is 100,000 samples. The training accuracy is close to 100%, the test accuracy is stagnant, and the overfitting is serious. At the same time, Word2Vec shows consistent (albeit modest) improvements in accuracy as the amount of data increases. The relatively flat lines in Fig. 4 show that Word2Vec is less affected by changes in parameter Settings and maintains a balanced performance between the training and testing phases.

Overall, Fig. 2, Fig. 3, and Fig. 4 show that although TF-IDF performs well in training accuracy, it is more prone to overfitting as the size of the dataset increases. This is mainly because TF-IDF relies on word frequency and inverse document frequency to represent text, and cannot effectively capture semantic relationships between words, coupled with the lack of diversity of training data sets, resulting in poor performance of the model on previously unseen data. In contrast, Word2Vec shows greater generalization across different data set sizes by better capturing semantic relationships. Therefore, the data set size and overfitting risk should be considered when selecting a method, and Word2Vec has advantages in applications that require robust generalization.

4 Conclusion

By analyzing a dataset of 560,000 food reviews from Kaggle, this paper compares the application of TF-IDF and Word2Vec text representation techniques in sentiment analysis tasks. The study highlighted that while TF-IDF achieved high accuracy on training data (about 99.16%), its accuracy on test data dropped significantly to 73.9%, indicating a significant overfitting problem. This shows that while TF-IDF is effective when capturing specific features in smaller or more controlled datasets, it is difficult to generalize when the dataset size increases. In contrast, Word2Vec showed more balanced performance, with training and test accuracy of 68.4% and 68.65%, respectively, indicating that it has stronger generalization ability across different data sets. The study highlights that Word2Vec provides better generalization by capturing semantic relationships between words, making it more suitable for sentiment analysis tasks, especially in scenarios involving large data sets. Future research could focus on developing hybrid models that combine the advantages of TF-IDF and Word2Vec, potentially improving the model's performance. In addition, exploring the applicability of these methods in other areas, such as social media analysis or interpretation of customer feedback, could further expand the impact and utility of this research.

References

1. H. Liu, X. Chen, and X. Liu. A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis. *IEEE Access*, 10, pp. 32280-32289. [Accessed 10 Aug. 2024] (2022).
2. T. Mikolov et al. Efficient estimation of word representations in vector space, *arXiv.org*. Available at: https://arxiv.org/abs/1301.3781 [Accessed: 04 August 2024] (2013).
3. K. Kowsari et al. HDLTEX: Hierarchical deep learning for text classification, *arXiv.org* (2017).
4. D. Tang, B. Qin, & T. Liu. 'Scalable Sentiment for Large Datasets', *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. [Accessed: 3 August 2024] (2014).
5. L. Zhao and K. Mao. 'Comparative Study of Text Representation: TF-IDF vs Word2Vec', *International Conference on Computer, Communications and Mechatronics Engineering (CCME 2017)* (2017).
6. A. Kumar & A. Chadha. An empirical study of the applications of text preprocessing in machine learning for sentiment analysis. *Procedia Computer Science*, 125, pp. 491-498 (2017).
7. Z. Yin & F. Zhuang. A comparative study of TF-IDF, LDA and Word2Vec for document representation. In *Proceedings of the 2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, pp. 602-607. IEEE (2017).
8. X. Rong. Word2Vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).
9. Z. Zhang & Z. Chen. A novel hybrid deep learning model with regularization techniques for sentiment analysis. *Neural Computing and Applications*, 31(7), pp. 3981-3991 (2019).
10. H. Wang, P. Li & W. Zhang. A comparative study on the performance of text classification models based on TF-IDF and Word2Vec. *Journal of Computer Applications Research*, 34(8), pp. 2452-2456 (2017).