Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/35/20230361

Deep learning for sentiment analysis on IMDB movie reviews using N-gram features

Sicheng Ouyang

Wuhan Britain China School, Wuhan, China

yangyiwu@whut.edu.cn

Abstract. In the rapidly evolving digital landscape, the synergy of deep learning techniques and abundant datasets has opened new frontiers in various domains. This research delves into the film industry, specifically harnessing the potential of the International Movie DataBase (IMDB) dataset for sentiment analysis. Through a deep learning paradigm, we embark on sentiment classification of movie reviews, discerning between positive and negative sentiments. By navigating data preprocessing and N-gram feature extraction, we engineer a deep learning model comprising embedding, global average pooling, and multi-layer dense architectures. The experimental results underscore the model's prowess in sentiment analysis, emphasizing its capacity to empower informed decision-making within the film industry.

Keywords: IMDB movie dataset, deep learning, sentiment analysis, N-gram feature extraction, model architecture.

1. Introduction

In the era of burgeoning digital information, the availability of extensive datasets and the advancements in deep learning methodologies have ushered in new dimensions of research across various domains. One such domain is the film industry, where the analysis of movie datasets offers valuable insights into understanding audience preferences, predicting movie success, and enhancing recommendation systems. Among these datasets, the International Movie DataBase (IMDB) movie dataset emerges as a pivotal repository of comprehensive movie-related information and audience ratings, serving as a crucial resource for film-related investigations.

While the potential of the IMDB movie dataset for empirical analysis and predictive modelling is well recognized, there remains an unexplored avenue in the realm of applying specific deep learning approaches to extract latent patterns and nuances from this dataset. The dynamic landscape of deep learning technologies presents an opportunity to unearth untapped dimensions in movie rating predictions and success factors. Furthermore, delving into the application of deep learning models to the IMDB movie dataset can extend the understanding of deep learning's applicability in various other contexts.

This study assumes significance as it delves into uncharted territories, uncovering the performance of the IMDB movie dataset when subjected to a specific deep learning methodology. The aim is to enhance the accuracy of movie rating predictions and the formulation of personalized recommendation strategies, thereby offering valuable decision support to movie industry stakeholders, including filmmakers, distributors, and audiences. This research endeavour contributes to the cinematic landscape

DOI: 10.54254/2755-2721/35/20230361

by bridging the gap between data-driven insights and informed decision-making, fostering the growth of the film industry and elevating user satisfaction. By exploring the performance of the IMDB movie dataset through the lens of deep learning, this study aspires to make a distinctive contribution to the realm of film research and industry practices.

2. Related work

Research on the IMDB movie dataset has brought substantial advancements across diverse domains. Notably, breakthroughs have been made in network security through Intrusion Detection Systems (IDSs) that identify intrusions and enhance accuracy using context-aware feature extraction [1]. Concurrently, predictive models for film success, employing RoBERTa embeddings and neural networks, seek to optimize decisions and investments within the film industry [2].

In the domain of malware detection, heightened accuracy has been achieved through N-Gram methodologies and feature selection guided by genetic algorithms [3]. For sentiment analysis, the adoption of SVM classifiers and diverse N-Gram sizes has deepened emotional comprehension on social media platforms [4]. Additionally, the cybersecurity landscape has gained from the Instruc2vec framework, harnessing deep neural networks to bolster protection [5].

Advancements in file fragment classification have yielded refined categorization by implementing sparse coding techniques [6]. Similarly, automated topic extraction from articles via N-Gram analysis has enriched information retrieval capabilities [7]. Privacy safeguards have been fortified through differential privacy-based N-Gram extraction approaches [8]. Lastly, network representation learning, particularly in node embedding for homogeneous networks, has offered crucial insights for network analysis [9].

Nevertheless, despite these advancements, limitations persist in existing research. For instance, intrusion detection systems might inadequately account for the diversity within network traffic, possibly leading to elevated false positive rates. Predictive models for film success could inadvertently neglect dynamic market factors, thus limiting their predictive accuracy. While N-Gram techniques display promise in malware detection, they may not comprehensively cover intricate, deep-level malware variants. While sentiment analysis excels on social media, its generalizability across diverse contexts and cultures remains a challenge. Both file fragment classification and automated topic extraction could be vulnerable to noise, influencing the precision of their results. Furthermore, integrating differential privacy techniques introduces a nuanced trade-off between data safeguarding and utility.

Given these insights, this study seeks to contribute by addressing these limitations. Specifically, the study endeavours to explore the application of deep learning to sentiment analysis on IMDB movie reviews, with the potential to overcome challenges associated with sentiment analysis across varied contexts. By integrating N-gram feature extraction and a well-architected model, the study aims to elevate sentiment classification accuracy, thus empowering more informed decision-making within the film industry.

3. Work preparation

In this research, a deep learning-based approach is utilized for sentiment analysis of the IMDB movie review dataset. The dataset undergoes tokenization as a preprocessing step to prepare it for neural network training. To capture contextual information and enhance the feature space, N-gram feature extraction is applied. The deep learning model comprises an Embedding layer that converts integer-encoded words into dense vectors, followed by a GlobalAveragePooling1D layer that obtains a compact representation of the input sequences. Intermediate processing is performed using two dense layers with ReLU activation, and the final step involves a dense layer with a sigmoid activation function for binary classification.

3.1. Tokenization

Tokenization is a fundamental natural language processing (NLP) technique used to split text data into individual tokens, such as words, punctuation marks, or numbers.

DOI: 10.54254/2755-2721/35/20230361

3.2. N-gram feature extraction

N-grams are widely used in NLP to capture contiguous sequences of N words. The introduction of the N-gram features aims to better capture the contextual information in the text, thereby improving the model's understanding and expression of the text's semantics.

Based on collinear relationships, the N-gram model serves as a language model. It aims to determine the probability of a given word string, denoted as p(W). Due to the inherent complexity, direct calculation is not feasible. As a solution, the multiplication law of discrete probabilities is commonly employed to compute the conditional probability of each segment within W independently. By utilizing this associated probability formula, the calculation of p(W) unfolds as follows [10](1).

$$p(W) = p(w_1, w_2, \dots, w_n) = p(w_1)p(w_2|w_1) \dots p(w_n|w_1, w_2, \dots, w_{n-1}) = \prod_{i=1}^{n} p(w_i|w_1, w_2, \dots, w_{i-1}) \quad (1)$$

3.3. Embedding

The embedding technique transforms discrete word tokens into continuous vector representations, providing a more effective way to represent text for sentiment analysis. By employing embeddings, the model gains a better understanding and captures the semantic relationships and contextual information between words. Specifically, the embedding layer in the code maps integer sequences (word indices) into continuous vectors with fixed dimensions, assigning each word a position in the vector space. This enables the model to learn the relative positions and similarities between words, leading to a better comprehension of the meaning and sentiment tendencies in the text data. In the context of sentiment analysis, the application of embedding greatly enhances the model's capability to comprehend text data, thereby improving the accuracy and performance of sentiment prediction.

3.4. GlobalAveragePooling1D

In the context of sentiment analysis, the Embedding layer converts integer-encoded words into dense vectors, capturing the semantic relationships between words. However, the generated vector sequences have varying lengths depending on the input text length. To address this issue and ensure that the model can handle texts of different lengths, the GlobalAveragePooling1D layer is introduced.

The GlobalAveragePooling1D layer performs a global average pooling operation on the vector sequences, transforming each text input's embedding vector sequence into a fixed-length representation. This pooling operation compresses and fuses information from the entire text, enabling the model to handle variable-length texts effectively.

By employing GlobalAveragePooling1D, different-length text inputs can obtain uniform-length vector representations, which contributes to the model's consistency and stability. Additionally, this pooling operation reduces the number of parameters, mitigates overfitting, and enhances the model's generalization capability.

3.5. Suitable activation function

In the deep learning-based sentiment analysis model, the essential activation functions, namely ReLU (Rectified Linear Unit) and Sigmoid, play a crucial role. These functions serve as fundamental building blocks, with ReLU introducing non-linearity by transforming negative input values to zero while retaining positive input values. The following equation illustrates this behaviour. (2).

$$F(x) = \max(0, x) \quad (2)$$

By incorporating ReLU activation in the intermediate layers, the model gains the ability to learn complex features and non-linear relationships in the data. This helps in better capturing the intricate patterns and semantic information present in the text data.

On the other hand, the Sigmoid function, which maps the model's output to a probability range of 0 to 1, is employed as the activation function in the output layer for binary classification. This function signifies the likelihood of a positive sentiment. The equation below illustrates this mapping. (3).

DOI: 10.54254/2755-2721/35/20230361

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

This enables the model to generate probabilistic predictions for each input text, indicating the probability of the sentiment being positive.

By utilizing ReLU and Sigmoid activation functions, the deep learning sentiment analysis model becomes adept at handling non-linear relationships, making more informed predictions, and ultimately improving the overall performance and accuracy of the sentiment analysis task.

4. Methodology

The methodology employed in this research involves a multi-step process to perform sentiment analysis on the IMDB movie review dataset using deep learning techniques. The goal is to classify movie reviews as either positive or negative based on their sentiments.

4.1. Data collection and preprocessing

The IMDB movie review dataset is obtained in this research, which consists of movie reviews labeled with positive or negative sentiments. The dataset is then preprocessed to prepare it for training the deep learning model. Tokenization is applied to convert the raw text reviews into sequences of integers, ensuring a consistent input format for the model. Additionally, padding is performed to make all review sequences have the same length.

4.2. N-gram feature extraction

To enhance the model's understanding of the text and capture contextual information, the technique applied is the technique of N-gram feature extraction. In this step, bi-grams are extracted from the tokenized sequences. Each bi-gram is mapped to an integer-encoded representation for further processing Figure 1.is the process of bigrams feature extraction.

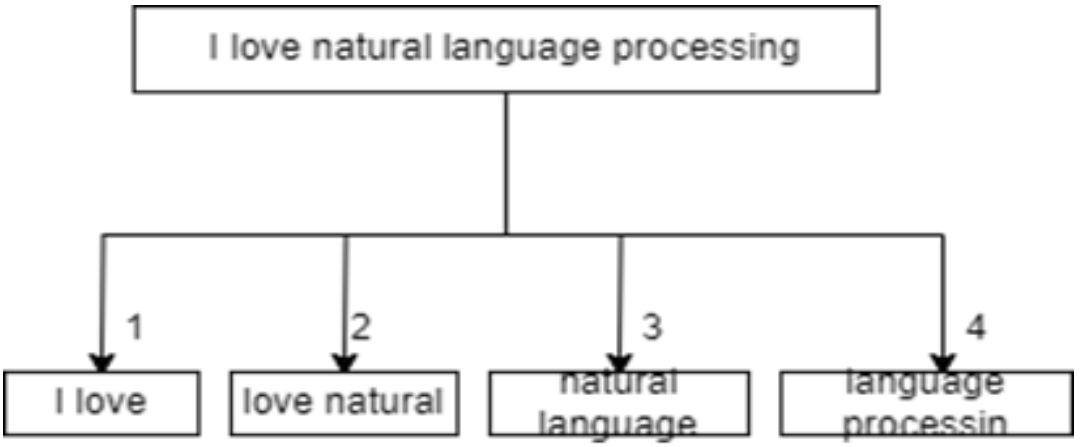

Figure 1. The process of bigrams features extraction.

4.3. Model architecture

Employing TensorFlow's Keras API, the deep learning model is designed with a sequence of layers. It begins with the addition of an embedding layer, which transforms the integer-encoded bi-grams into dense vector representations. Subsequently, a GlobalAveragePooling1D layer is employed to yield a fixed-length vector representation for each review, contributing to a reduction in the model's computational complexity. Intermediate processing involves the inclusion of two fully connected layers, both utilizing ReLU activation functions. The final touch is a dense layer that incorporates a sigmoid activation function, facilitating binary classification for the prediction of positive or negative sentiment.

DOI: 10.54254/2755-2721/35/20230361

4.4. Model compilation training & evaluation

Compiled with the Adam optimizer and utilizing the binary cross-entropy loss function, the deep learning model is trained on a dataset comprising 25,000 reviews. Throughout the training procedure, the model's performance is meticulously tracked by employing the validation dataset. To proactively address potential overfitting, an early stopping strategy is implemented.

Subsequent to the training phase, a comprehensive evaluation of the model is conducted using a distinct test dataset containing approximately 25,000 reviews. Performance metrics, encompassing accuracy and loss, are leveraged to meticulously assess the model's proficiency in effectively classifying movie reviews into distinct positive or negative sentiments.

5. Implementation

The implementation for sentiment analysis of the IMDB movie review dataset involves several key steps to construct and train the deep learning-based model. The dataset is pre-processed, and N-gram feature extraction is applied to capture contextual information. The deep learning model comprises embedding and pooling layers, followed by fully connected layers for classification fig.2. shows the flow chart of the model.

The IMDB movie review dataset is loaded and pre-processed. The reviews are tokenized and encoded as sequences of integers, ensuring a consistent input length for the subsequent model.

N-gram feature extraction is applied to enhance the model's understanding of the text. Bi-grams are extracted from the sequences and mapped to integer-encoded representations. This process helps capture the contextual information between consecutive word pairs in the reviews.

Constructed using TensorFlow's Keras API, the deep learning model encompasses multiple layers. The initial embedding layer serves to transform integer-encoded bi-grams into dense vector representations. Following this, a GlobalAveragePooling1D layer is introduced to derive a fixed-length vector representation. Further enhancing the model's capabilities, two additional dense layers, activated by ReLU functions, are included for intermediate processing.

The ultimate layer employs a sigmoid activation function, facilitating binary classification. Subsequently, the model undergoes compilation with an appropriate optimizer and loss function, with pre-emptive measures in place to counter overfitting. Through training on the designated dataset and validation on a separate dataset, convergence or adherence to early stopping criteria is ensured. To visually depict this model's structure, refer to Figure 2.

DOI: 10.54254/2755-2721/35/20230361

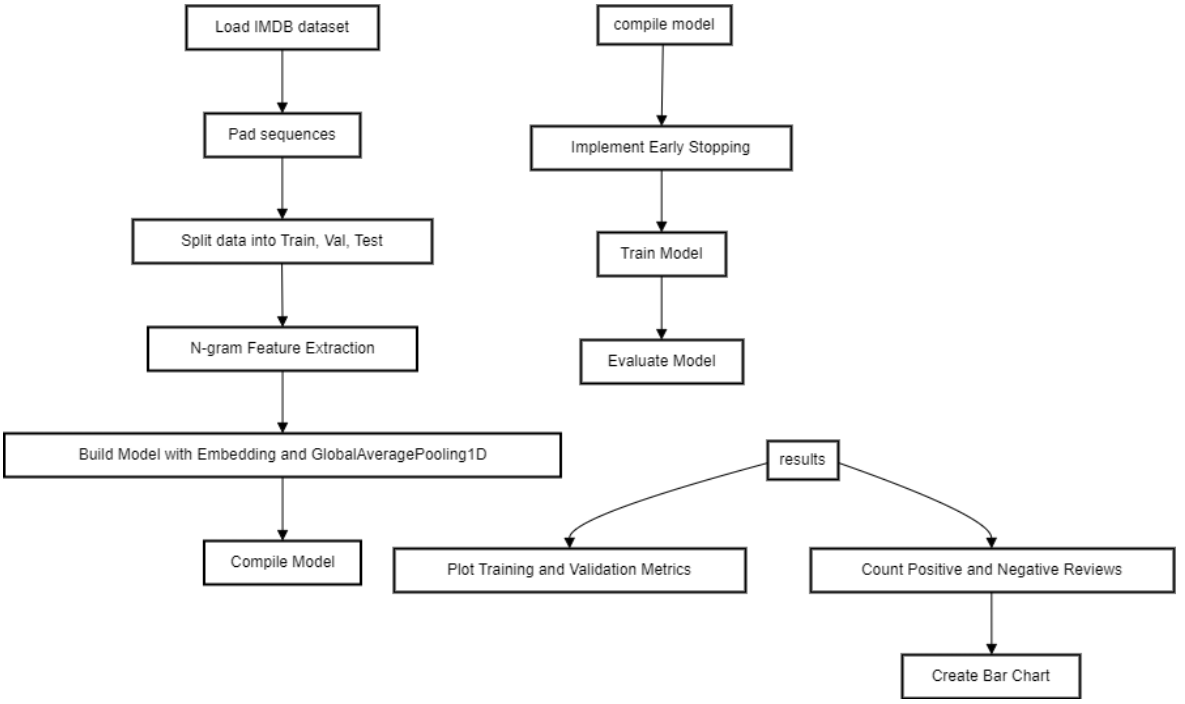

Figure 2. The basic logic of the model.

6. Results analysis

The outcomes of sentiment analysis experiment provide valuable insights into the performance of the deep learning-based approach employed on the IMDB movie review dataset. The obtained test loss of 0.298 and the corresponding test accuracy of 87.65% showcase the effectiveness of the model in accurately distinguishing between positive and negative sentiments within movie reviews. Figure 3. combining with Figure 4. shows the result of the experiment.

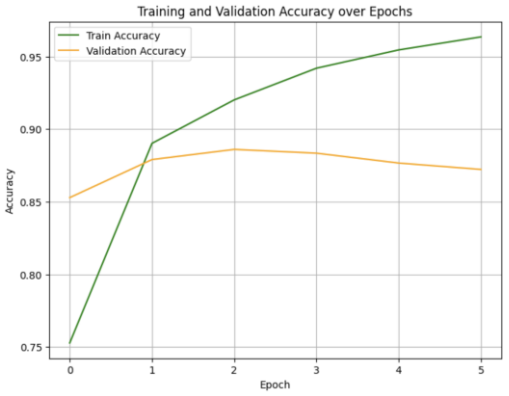

Figure 3. The curve of the loss and accuracy of train dataset.

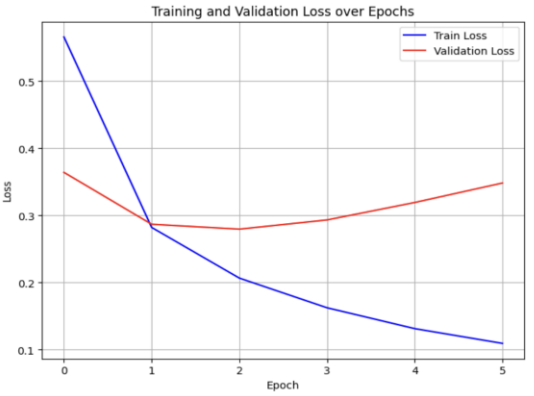

Figure 4. The curve of the loss and accuracy of test datasets.

These metrics underscore the model's ability to learn and leverage relevant features from textual data, resulting in precise sentiment classification. Table1. shows the accuracy and loss of each epoch, with 157 sets of reviews in one epoch.

DOI: 10.54254/2755-2721/35/20230361

Table1. Accuracy and loss of each epoch.

Epoch no.	Loss	Accuracy
1	0.5660	0.7527
2	0.2822	0.8902
3	0.2067	0.9201
4	0.1626	0.9419
5	0.1315	0.9545
6	0.1095	0.1095

The test loss of 0.298 indicates that the model has effectively converged during training. This value represents the difference between predicted and actual sentiment labels in the test set. A lower test loss signifies a higher level of agreement between predictions and actual sentiments, highlighting the model's robust ability to generalize to unseen data.

The test accuracy of 87.65% reaffirms the robustness of the model in sentiment classification. This metric demonstrates the model's capability to correctly identify the polarity of sentiments within movie reviews. The high accuracy implies that the model can discern complex patterns and subtle nuances within the textual data, leading to accurate sentiment predictions.

In conclusion, the results emphasize the efficacy of the proposed approach for sentiment analysis. The amalgamation of a low-test loss and a high-test accuracy reaffirms the model's prowess in extracting significant features from textual data, thereby enhancing the reliability of the sentiment analysis process.

7. Conclusion

Amid rapid digital expansion, deep learning advancements and vast datasets have opened novel avenues. This study concentrates on the film industry, specifically IMDB movie dataset analysis. While its film research potential is recognized, uncharted terrain lies in applying deep learning techniques. This promises fresh movie rating insights and cross-domain deep learning applications. Merging IMDB data and deep learning, this research targets film breakthroughs, propelling industry growth and user satisfaction.

Though this study reveals deep learning's IMDB sentiment analysis potential, improvement opportunities exist. Enriching sentiment detection through extended lexicons is one avenue. Attention mechanisms could enhance vital review capture and classification. Exploring diverse architectures, like Transformers, may amplify sentiment analysis impact. Global applications might emerge from multilingual sentiment analysis. Extensive experiments could comprehensively assess the proposed approach, enhancing sentiment analysis accuracy and utility.

References

[1] Shams EA, Rizaner A, Ulusoy AH. A novel context-aware feature extraction method for convolutional neural network-based intrusion detection systems. Neural Computing and Applications. 2021 Oct;33(20):13647-65.

[2] Jose A, Harikumar S. Predicting IMDB Movie Ratings Using RoBERTa Embeddings and Neural Networks. InResponsible Data Science: Select Proceedings of ICDSE 2021 2022 Nov 15 (pp. 181-189). Singapore: Springer Nature Singapore.

[3] Parvin H, Minaei B, Karshenas H, Beigi A. A new N-gram feature extraction-selection method for malicious code. InAdaptive and Natural Computing Algorithms: 10th International Conference, ICANNGA 2011, Ljubljana, Slovenia, April 14-16, 2011, Proceedings, Part II 10 2011 (pp. 98-107). Springer Berlin Heidelberg.

[4] de Godoi Brandão J, Calixto WP. N-Gram and TF-IDF for Feature Extraction on Opinion Mining of Tweets with SVM Classifier. In2019 International Artificial Intelligence and Data Processing Symposium (IDAP) 2019 Sep 21 (pp. 1-5). IEEE.

DOI: 10.54254/2755-2721/35/20230361

[5] Poornima S, Subramanian T. Effective Feature Extraction via N-Skip Gram Instruction Embedding Model using Deep Neural Network for designing Anti-Malware Application. In2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) 2023 Mar 17 (Vol. 1, pp. 2118-2123). IEEE.

[6] Wang F, Quach TT, Wheeler J, Aimone JB, James CD. Sparse coding for n-gram feature extraction and training for file fragment classification. IEEE Transactions on Information Forensics and Security. 2018 Apr 5;13(10):2553-62.

[7] Zhu L, Wang W, Huang M, Chen M, Wang Y, Cai Z. A N-gram based approach to auto-extracting topics from research articles1. Journal of Intelligent & Fuzzy Systems. 2022 Jan 1;43(5):6137-46.

[8] Kim K, Gopi S, Kulkarni J, Yekhanin S. Differentially private n-gram extraction. Advances in Neural Information Processing Systems. 2021 Dec 6;34:5102-11.

[9] Zhou J, Liu L, Wei W, Fan J. Network representation learning: from preprocessing, feature extraction to node embedding. ACM Computing Surveys (CSUR). 2022 Jan 18;55(2):1-35.

[10] Damashek M. Gauging similarity with n-grams: Language-independent categorization of text. Science. 1995 Feb 10;267(5199):843-8.