IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 6, December, 2023, pp. 1613-1618
Journal homepage: http://iieta.org/journals/isi

Enhancement of Sentiment Analysis in Hotel Reviews through Latent Semantic Indexing and Convolutional Neural Networks

Nobogh Husssein Baqer[1*], Ahmed T. Sadiq[2], Zuhair Hussein Ali[1]

[1] Department of Computer Science, College of Education, Mustansiriyah University, Baghdad 10001, Iraq
[2] Department of Computer Science, University of Technology-Iraq, Baghdad 10001, Iraq

Corresponding Author Email: nobogh.hussein@uomustansiriyah.edu.iq

https://doi.org/10.18280/isi.280618

Received: 4 May 2023
Revised: 22 August 2023
Accepted: 9 October 2023
Available online: 23 December 2023

Keywords:
trip advisor, LSI, CNN, SVD, classification, standard evaluation metrics

ABSTRACT

Sentiment Analysis (SA) is a prominent field of study concerned with the classification of sentences within a document as either positive or negative. The extraction of features from a document plays a crucial role in SA for achieving precise text classification. This study employs the Latent Semantic Indexing (LSI) algorithm for feature extraction, designed to address the limitations inherent in the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The features extracted are then utilized in the Convolutional Neural Network (CNN) classification algorithm, which encompasses two convolutional layers, a single polling layer, a fully-connected layer, and two output nodes. This is done to evaluate the efficacy of the proposed model. Experimental results indicate that the combination of LSI and CNN significantly improves text classification. Customer reviews exert considerable influence on individuals' travel plans, with a preference typically shown towards hotels with a preponderance of positive reviews. Consequently, these reviews serve as crucial resources for managers seeking to enhance their services. In this study, a dataset of hotel reviews is employed, and the resulting data is evaluated using standard metrics such as precision, recall, f-score, and accuracy, yielding results of 89%, 77%, 80.5%, and 87% respectively.

1. INTRODUCTION

Sentiment Analysis (SA) is a text mining technique employed to extract the emotional sentiment of a given text or sentence, identifying its polarity as being either "positive" or "negative". Particularly in the context of hotel reviews, SA has gained considerable significance as each review expresses specific sentiments, either positive or negative, which ultimately shape the hotel's reputation [1]. The impetus for this study on SA for hotel reviews is derived from the need to comprehend customer opinions and sentiments thoroughly, given that hotel reviews notably sway people's travel plans, and businesses can harness these reviews to refine their services.

In light of technological advancements and the surge in interactions on social media platforms, where individuals voice their opinions regarding specific services or products, it has become incumbent on any business to take user reviews into account. These reviews play a crucial role in facilitating better service provision to customers. Conversely, businesses can leverage user reviews to discern customer preferences, identify areas of weakness, and amplify their robust offerings. With the proliferation of content on the Internet, businesses confront the challenge of analyzing ample volumes of reviews to effectively comprehend opinions. The manual analysis of such data proves to be impractical and time-consuming, necessitating automated methods for sentiment classification. This underscores the importance of SA techniques, which offer automated methods to analyze and categorize sentiments in large datasets [2].

This paper is centered on the implementation of SA using a Convolutional Neural Network (CNN) classifier on a dataset comprising over 20,000 hotel reviews. The objective is to bifurcate the reviews into two clusters, "negative" and "positive". Given that many of the reviews are devoid of meaningful content and that Trip Advisor's star rating does not accurately reflect the customers' experiences, the preprocessing of the dataset is essential to eliminate irrelevant data, symbols, blanks, and to apply word-stemming techniques [3]. For encoding words and constructing internal representations of documents, Latent Semantic Indexing (LSI) is utilized as a feature extraction method. LSI facilitates the comprehension of the structure and meaning of the text by transforming the original statements into a smaller semantic space. This transformation is realized through Singular Value Decomposition (SVD), whereby terms used in similar contexts are grouped together. Consequently, documents that employ diverse terminologies to express the same meaning can be positioned close to each other in the new semantic space [4, 5]. The question this study aims to answer is: How can SA techniques be applied to a large dataset of hotel reviews for classification, and achieve high accuracy?

The contributions of this study are two-fold:

(1) The implementation of LSI as a feature extraction

method in the proposed model, which overcomes the limitations of traditional approaches such as TF-IDF.

(2) The identification of synonyms within the dataset, leading to a deeper understanding of the underlying semantic relationships and subsequently improving the accuracy of classification.

The remainder of the paper is organized as follows: Section 2 presents related work, Section 3 describes the preliminary concepts of techniques, Section 4 outlines the methodology employed for hotel reviews, and Section 5 presents the experimental results. Finally, Section 6 concludes the paper.

2. RELATED WORK

Numerous studies have explored Sentiment Analysis (SA) in the context of hotel reviews, employing various classification techniques. This section presents a selection of these studies.

In 2011, Kasper and Vela introduced an opinion mining system that gathers comments on hotel reviews from the web, aiding hotel management in monitoring online publications about their establishments [6]. The system provides structured and classified overviews of comments, facilitating access to this information. Despite certain issues, it demonstrated satisfactory performance in analysis and classification tasks. Accuracy levels achieved were 82% using Statistical Polarity Classification, 68% with Information Extraction (IE) Polarities, and approximately 83% when both techniques were utilized in tandem [6].

Narayanan, in 2011, applied SA to a dataset of Trip Advisor hotel reviews, classifying them as either negative or positive, thus analyzing customer sentiment [2]. The Term Frequency-Inverse Document Frequency (TF-IDF) technique was used to extract frequent words from the sentences. Various classification methods were employed to classify sentences and ascertain accuracy: Naïve Bayes Multinomial achieved 79.12%, Support Vector Machine (SVM) attained 75.29%, and Naïve Bayes Bernoulli reached 78.86%. These accuracy rates increased with the enlargement of the training data. The study concluded that machine learning methods can surpass human-produced SA baselines, based on the experimental results [2].

In 2019, Tran et al. analyzed sentiment for hospitality data from Trip Advisor with a precision of 88.55% [3]. The researchers proposed a framework to summarize reviews by merging the Aspect Term Extraction-Polarity Classification (ATE-PC) task with the Latent Dirichlet Allocation (LDA) model. This combination was used to analyze large datasets to identify the features and their respective polarities that customers focus on, with the aim of improving service operations and strategies. The analysis led to the conclusion and collection of 11 topics from the LDA model [3].

Most recently in 2021, Anis et al. conducted SA on user reviews using three classifier models: Random Forest, Naive Bayes, and Support Vector Machine [7]. After computing the confusion matrix for each model, it was found that the SVM performed better than the other classification techniques, yielding an accuracy of 81.6% and an F1-score of 66.5%. The sentences were categorized into positive, negative, or neutral labels, and the performance of the algorithms was assessed on these two operands [7].

The models presented by Narayanan [2] and Anis et al. [7], which are compared to the model proposed in this paper, rely on TF-IDF for feature extraction and implement various classification algorithms. Among these, the SVM algorithm yielded the best results. However, a significant drawback of classification algorithms is the large volume of training data required. Furthermore, TF-IDF cannot convey semantic meaning; it merely assigns weights to words to determine their relevance but cannot construe the context of a phrase or ascertain its significance. Consequently, the model proposed in this paper employs LSI to classify hotel reviews. LSI can somewhat handle the problem of synonymy by decomposing the term-document matrix, rendering it faster compared to other models.

3. PRELIMINARY CONCEPTS

This section presents an overview of the preliminary concepts used in this research which are LSI and CNN, to understand the methodology used in the proposed model. Where this model contributes to improving the performance of the SA and giving good results with huge databases.

3.1 Latent semantic indexing

LSI is a method of analyzing documents set to discover statistical duplicates of words that give insights into the topics of those words and documents. The words that tend to happen together or happen with similar words are regarded as to be semantically similar. To build the LSI model, first, create a matrix of the document, the columns corresponding to documents and rows corresponding to words. Every entry to the matrix is frequency weighted of the word in the document. This weighting is to minimize the effect of frequently occurring words. By using SVD this large matrix will be reduced into a compressed matrix [8].

The SVD is the product of three matrices, as in Eq. (1). It has algebraic features and transfer an important geometrical and theoretical insights about linear transformations. The SVD is also widely utilize in system identification to gain balanced reduced-order models [9]. LSI performs a kind of noise reduction and has the benefit of detecting synonyms as well as words that refer to the same subject [10, 11].

$$X_{t\times d} = W_{t\times n} S_{n\times n} \left(P_{d\times n}\right)^T \tag{1}$$

where,

S is the diagonal matrix

W represents the terms matrix of dimension t* n (t terms number, n words number)

P represents the documents matrix of dimension d* n (d number of documents, n number of words).

X is the term-document matrix of dimension t* d

The diagonal components of S are arrange by magnitude. Matrix X is the product of these three matrices [12].

3.2 Convolutional neural network

CNN consists of neurons that self-optimize via learning that is utilized for classification. It comprises an input layer, output layer, and an hidden layers every neuron will receive an input and implement an operation. The network expresses a single perceptive score function (the weight), from input vectors to the final output. The hidden layers include layers that perform convolutions, and the last layer involves loss functions

relevant to the classes [13]. The architecture of CNN is shown in Figure 1, here's a more comprehensive explanation of understanding CNN's key components and role:

(1) Convolutional Layers: these are the first blocks of CNNs, composed of filters that are convolved over the input data. The filters capture local patterns in the data by applying a dot product operation between the filter weights and the input data, which helps in capturing spatial relationships and extracting important features.

(2) Pooling Layers: play an important role in down-sampling and reducing the dimension of the attribute maps and make the model more robust to variations in the input. Pooling techniques include max pooling (selects the maximum value within each local region), and average pooling (compute the average value). It helps extract the most salient features from the feature maps while reducing the computational complexity of the network.

(3) Fully Connected Layers: these layers take the feature maps from the previous layers and perform classification tasks. It captures global relationships and learns representations based on the extracted features to make predictions about the sentiment of the input text [14].

The convolution takes the most time for the training of the neural network. The major objective of the convolution layer is to extract features from the input. The dropout algorithm is introduced for training neural networks through dropping units randomly while training to prevent their co-adaptation. Dropout setting the output of every hidden neuron to zero with 0.5 probability. This algorithm will drop out the neurons which don't contribute to the forward pass and don't involve back-propagation [15].

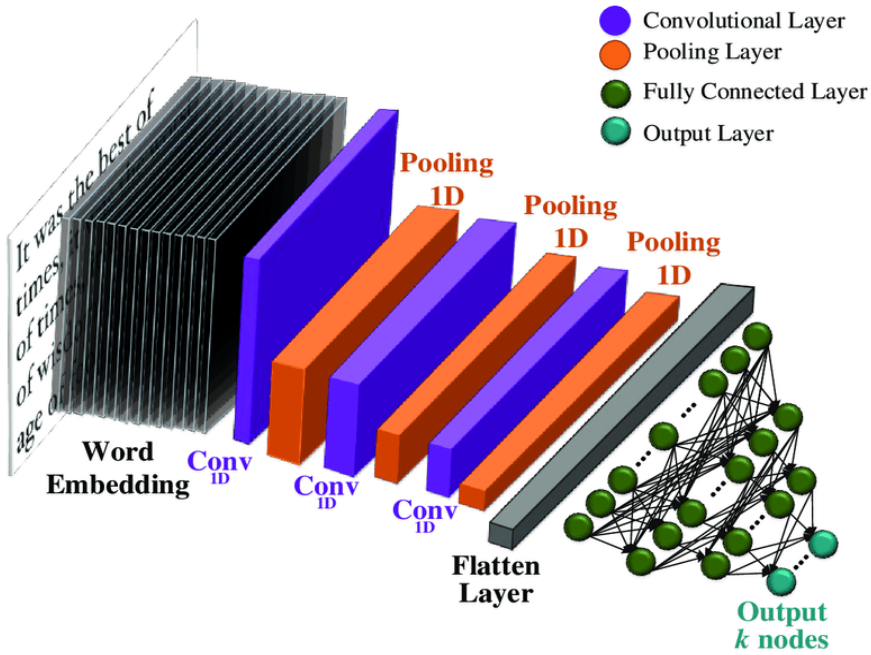

Figure 1. CNN architecture

4. METHODOLOGY

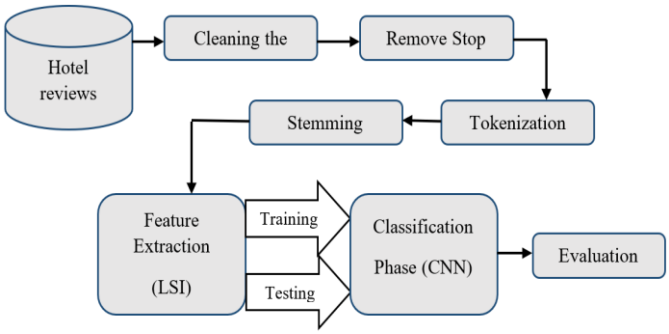

Figure 2. Steps of the proposed model

This part will discuss the phases involved in analyzing sentiment by using the proposed model to classify review as "positive", or "negative". Consider that the dataset consists of a large number of comments (or sentences) and each sentence consists of several words with some symbols or numbers etc. This proposed model pre-processed the dataset and then the features are extracted by LSI to collect similar synonyms in one class, and then classify them by CNN classifier. Finally, evaluating the result through standard evaluation metrics, as shown in Figure 2. The detail for every component will be discussed in the following sub-sections.

4.1 Dataset

Hotels play a crucial role in travel and with increased access to customer experience information, new paths have emerged to choose the best one. The hotel review trip advisor dataset used in the proposed model consists of 20491 reviews in the English language. Where it was divided as: 1421 Very negative (one star), 1793 Negative (two star), 2184 Neutral (three-star), 6039 Positive (four star), and 9054 Very positive (five star). If suppose rate from 1 to 3 is negative and 4, to 5 are positive the percentage will be 73.7% reviews for positive and 26.3% for negative. The number of positive and negative reviews is summarized in Table 1 after converting the data to only two classes.

Table 1. The number and percentage of positive and negative reviews in the hotel dataset

Review Type	No. of Review	Percentage
Positive	15093	73.7%
Negative	5398	26.3%

4.2 Data pre-processing

To perform any operation on the dataset, in the first phase, pre-processing must be carried out and by type of processing required to prepare the text for the next phase. Any data collected from platforms will contain some noise like emojis, blanks, hyperlinks, punctuation, and frequent repetition of letters as a kind of emphasis on the word [16, 17]. The pre-processing steps used to process the hotel review sentences are:

(1) Tokenization: The first step is to put each word in a separate row called a token.

(2) Text Cleaning: There are numerous text cleanings like converting upper case letters to lower case, Spell checking, and removing punctuation and special characters.

(3) Remove Stop words: these are words that are common or frequently used in sentences but are often not useful for analysis and are usually removed.

(4) Stemming: reduces words to their roots by using Porter stemmer [18].

4.3 LSI as feature extraction

The feature extraction process plays a significant role due to its effect on the efficiency of the classifier in analyzing hotel reviews. In this paper, removing conjunctions, pronouns, and common verbs, this help in isolate the terms that contain the main content of a phrase that is done by applying the LSI algorithm to (the Semantic Vector Space Model) and then analyzing the relationship between the word and document. LSI extracts and expresses the word's semantics by using the statistical method. Then, these terms are placed in a Term Document Matrix (TDM). Where TDM is a 2D grid containing the frequency of every particular word that happens in the document within a dataset. Then the SVD algorithm is utilized to minimize the number of rows in the matrix in the status of column information and represent the similarity of every two words of its row vector, where the similarity value is between 1 and 0 and the higher the value, the greater the similarity between two words. It works as in the following steps:

(1) Calculate the frequency for words, as shown in Figure 3.

(2) Convert text to the M matrix as in equation 1, where the line constitutes terms, and the column constitutes the documents.

(3) Calculate SVD to reduce the dimension from the high to low dimensional space.

(4) Gain words similarity to express drop dimensional.

The reason LSI is preferred over TF-IDF is because LSI can measure the semantics similarity between words and this doesn't exist in TF-IDF. The other reason the storage space of the text can be reduced and improve classification efficiency of by utilizing the SVD algorithm.

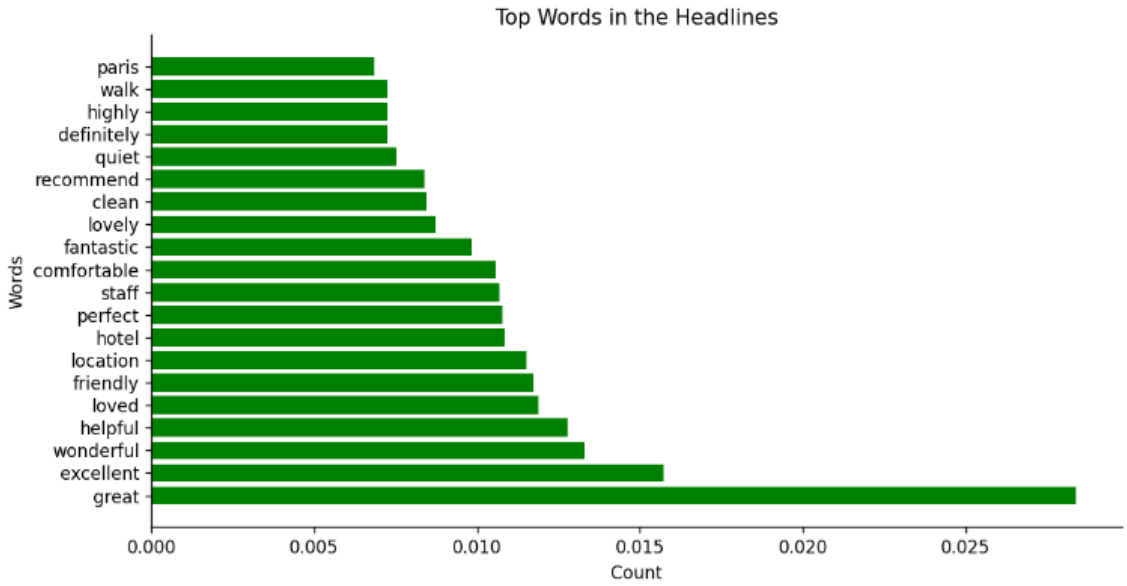

Figure 3. The more important features of positive reviews

4.4 CNN as classifier

CNN is effective in classifying the text, where the CNN takes advantage of filters of convolutional that automatically learn the characters where it can catch inherent syntactic and semantic character of sentimental expressions. The constructed CNN is composed of a word embedding, two convolutional, a pooling, a fully-connected, and 2 output nodes, as in Figure 4.

The embedding layer puts words received as input into semantic space, the words with similar meanings are placed together in the same class, and different words are far apart. The embedding layer's output relocates to first convolutional layer. The matrix of the convolutional layer keep the local information required to sentiment classification and then relocates the result to the second convolutional layer, by using 64filters in each layer, it extracts characters from the contextual information of the term based on the local information in the first layer (convolutional layer). An activation function (ReLU) is applied after each convolutional layer to introduce non-linearity and enable the model to learn complex patterns. According to the pooling layer, it chooses

the biggest value to represent values, since sentiments are expressed in several words (with different meanings) for this used max-pooling method. After the pooling layer, implement a flattening process that turns the 2 dimensional feature map into 1 dimensional format then you move the to a fully connected layer. It associates every input and output node, the vector that passes through this layer represents the last output which classifies as “positive”, or “negative”.

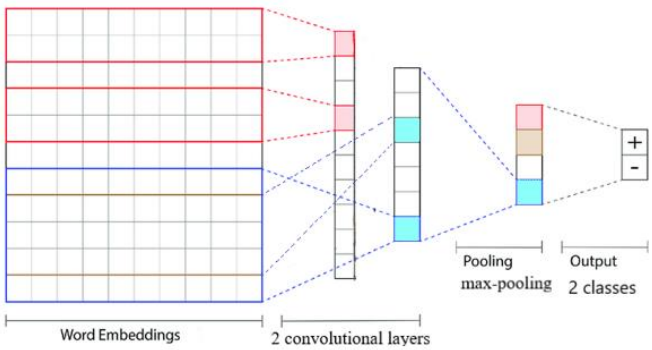

Figure 4. CNN architecture in the proposed model

5. EXPERIMENTAL RESULTS

This study practical experiments were conducted to verify the text-analyzing model that uses the LSI algorithm. Initially, preparing the text for classification operations through pre-processed by cleaning the text, converting the letters to lowercase, and returning the words to their roots using the stemming process then tokenizing them.

Table 2. Classification results

	Accuracy	Precision	Recall	F1-score
positive	0.89	0.86	0.98	0.92
negative	0.85	0.92	0.56	0.69

Table 2 shows the experimental result, where the LSI technique is utilized to represent text documents in a vector space model. This technique allows the detection of hidden relationships and similarities among documents, which can help improve the accuracy of sentiment analysis. LSI is effective in reducing the dimensionality of text data and identifying the most important features for sentiment analysis. This technique solves the issue of semantic deleting of TF-IDF. Which in turn helps speed up the processing due to a decrease in the overall feature words and an increase in effective feature words. On the other hand, CNN is a multi-layer neural network that consists of stacking many hidden layers in sequence. CNN sequential design allows the network to learn hierarchical features from raw data, and it can identify data patterns that indicate positive or negative sentiment.

Overall, the combination of LSI and CNN can be a powerful tool for sentiment analysis compared to other classification methods, as it can help identify important features and patterns in the text data. Figure 5 shows the accuracy and loss over epochs for each the testing and training sets, where the dataset was split into 20% tasting set and 80% for the training set, by using a 10-fold cross-validation procedure. Cross-validation is an approach for evaluating and testing the accuracy of a model, where 9-fold is used in the training phase and the remaining fold is used in the testing phase. This process is repeated where each fold takes a chance to be nine times in the training phase and one time to be in the testing phase.

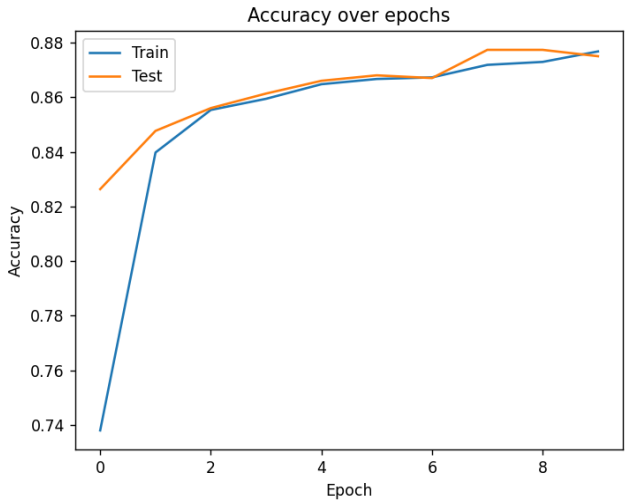

Figure 5. The accuracy and loss over epochs

5.1 Comparison with other methods

As shown in Table 3, the proposed model is presented in two methods. Firstly, attempt to implement the CNN classifier with the TF-IDF to extract features, which only overcomes the Naïve Bayes classifier in the study of Narayanan [2]. On the other hand, used the LSI with CNN which in turn outperforms the works presented in the study of Narayanan et al. [2, 7] and achieves the highest performance, especially in accuracy, which is the major factor to measure the performance of the classifier. The superiority of the proposed model is due to the use of the LSI algorithm to extract features that show it can improve classifier performance by up to 0.07. LSI can handle synonymy problems by creating a decomposing term-document matrix that in turn speeds up the processing process and classification compared to other models. While in the related works, the features are extracted based on the TF-IDF method, which does not support the semantic detection of words which increases the total number of feature terms. In addition to using the different classification methods that also has an impact on the results of the classification.

Table 3. Comparison with other methods

Model	Accuracy	Precision	Recall	F-Score	Feature Extraction
Naïve Bayes [2]	79.12%	82.65%	83.37%	83.01%	TF-IDF
SVM [7]	81.6%	70.75%	65.82%	66.54%	TF-IDF
Proposed model (CNN)	80%	75.5%	74.5%	74.5%	TF-IDF
Proposed model (CNN)	87%	89%	77%	80.5%	LSI

6. CONCLUSION

Customers write their opinions and criticisms of hotels on social media platforms where considered as an important resource of information and part of the travel plan. These reviews of hotels are very useful for travel companies, customers, and hotel managers to improve their services. To help in analyzing problems for hotel review through using SA methods. The dataset used to help in performing this study was more than 20K of customer reviews for trip advisor. In the proposed model, the combination of LSI and CNN resulted in significant performance optimizations comprised to other techniques for SA. The use of LSI can capture the semantic relationships in the dataset, where this method overcomes the limitations of TF-IDF. Additionally, the CNN architecture allowed for the effective utilization of the extracted features. The model could capture patterns within the dataset by employing multiple convolutional layers and pooling layers.

Specifically, the proposed model shows improvements in evaluation metrics, especially in the accuracy of classification. Businesses can leverage classification techniques to enhance their services, optimize marketing, and improve customer satisfaction.

REFERENCES

[1] Shi, H.X., Li, X.J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics, IEEE, 3: 950-954. https://doi.org/10.1109/ICMLC.2011.6016866

[2] Narayanan, V.E.G. (2011). Sentiment analysis for hotel reviews. In Proceedings of the Computational Linguistics Conference, 231527: 45-52.

[3] Tran, T., Ba, H., Huynh, V.N. (2019). Measuring hotel review sentiment: An aspect-based sentiment analysis approach. In Integrated Uncertainty in Knowledge Modelling and Decision Making: 7th International Symposium, IUKM 2019, Nara, Japan, March 27-29, 2019, Springer International Publishing. Proceedings 7: 393-405. https://doi.org/10.1007/978-3-030-14815-7_33

[4] Rosario, B. (2000). Latent semantic indexing: An overview. Technical Report INFOSYS, 240: 1-16.

[5] Chauhan, R., Ghanshala, K.K., Joshi, R.C. (2018). Convolutional neural network (CNN) for image detection and recognition. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), IEEE, Jalandhar, India, pp. 278-282. https://doi.org/10.1109/ICSCCC.2018.8703316

[6] Kasper, W., Vela, M. (2011). Sentiment analysis for hotel reviews. Proceedings of the Computational Linguistics Conference, 231527: 45-52.

[7] Anis, S., Saad, S., Aref, M. (2021). Sentiment analysis of hotel reviews using machine learning techniques. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020. Springer International Publishing, pp. 227-234. https://doi.org/10.1007/978-3-030-58669-0_21

[8] Huang, Y. (2003). Support vector machines for text categorization based on latent semantic indexing. Electrical and Computer Engineering Department, The Johns Hopkins University, Technical Report.

[9] Brunton, S.L., Kutz, J.N. (2019). Singular value decomposition (SVD). Data-Driven Science and Engineering, pp. 3-46. https://doi.org/10.1017/9781108380690.002

[10] Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50-57. https://doi.org/10.1145/312624.312649

[11] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6): 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

[12] Zelikovitz, S., Hirsh, H. (2001). Using LSI for text classification in the presence of background text. In Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 113-118. https://doi.org/10.1145/502585.502605

[13] O'Shea, K., Nash, R. (2015). An introduction to convolutional neural networks. arXiv Preprint arXiv: 1511.08458. https://doi.org/10.48550/arXiv.1511.08458

[14] Hu, H., Zheng, W., Zhang, X., Zhang, X., Liu, J., Hu, W., Duan, H., Si, J. (2021). Content-based gastric image retrieval using convolutional neural networks. International Journal of Imaging Systems and Technology, 31(1): 439-449. https://doi.org/10.1002/ima.22470

[15] Ouyang, X., Zhou, P., Li, C.H., Liu, L. (2015). Sentiment analysis using convolutional neural network. In 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, IEEE, Liverpool, UK, pp. 2359-2364. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349

[16] Pahwa, B., Taruna, S., Kasliwal, N. (2018). Sentiment analysis-strategy for text pre-processing. International Journal of Computer Applications, 180(34): 15-18. https://doi.org/10.5120/ijca2018916865

[17] Parlar, T., Ozel, A.S., Song, F. (2019). Analysis of data pre-processing methods for the sentiment analysis of reviews. Computer Science, 20(1): 123. https://doi.org/10.7494/csci.2019.20.1.3097

[18] Haddi, E., Liu, X., Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17: 26-32. https://doi.org/10.1016/j.procs.2013.05.005