Comparative Study of Performance of K Nearest Neighbor and Support Vector Machine Classifiers in Sentiment Analysis

Tula Kanta Deo[1], Rajesh Keshavrao Deshmukh[2], Gajendra Sharma[3]

[1] Department of Computer Science and Engineering, Kalinga University, Naya Raipur, India
[2]Department of Computer Science and Engineering, Kalinga University, Naya Raipur, India
[3] Department of Computer Science and Engineering, Kathmandu University, Kavre, Nepal

---***---

Abstract – *Sentiment analysis is the most important branch of natural language processing. It deals with the classification of text. The class can be positive, negative or other. This study evaluate and compare the performance of K Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers. The datasets used in this study are all_tweets dataset and financial phrase bank dataset. These datasets are preprocessed. The preprocessed datasets are split into 80% training and 20% testing subsets. The training dataset are used for feature extraction and training of the classifiers. The testing datasets are used for feature extraction and evaluation of the classifiers. The results and discussions of this study shows the performance of KNN and SVM is consistent with most of the studies. In this study, SVM outperform KNN.*

Keywords: Sentiment Analysis, K Nearest Neighbor, Support Vector Machine, Precision, Recall, Accuracy, F1 score.

1.INTRODUCTION

Sentiment analysis(SA), also known as opinion mining, is a subfield of natural language processing (NLP) that deals with automatically determining the emotional tone of a piece of text. It aims to understand whether the text expresses positive, negative, or neutral sentiment towards a topic, entity, or event[1].

A comparative study between the KNN and SVM classifiers for sentiment analysis involves evaluating the performance of these classifiers on sentiment classification tasks. The objective of this study is to evaluate and compare the performance of KNN and SVM classifiers on sentiment classification tasks. This includes assessing their accuracy, precision, recall and F1-score metrics to understand how well each classifiers performs in sentiment analysis.

KNN classifier is renowned for its simplicity and effectiveness in classification tasks. It belongs to the family of instance-based algorithms, where predictions are made based on the similarity of new instances to known instances in the training data. KNN is a non-parametric algorithm; that is, it makes no assumptions about the underlying data distribution. This makes it versatile and applicable to a wide range of datasets. KNN is a lazy learning algorithm because it postpones the learning process until the prediction phase. It stores the entire training dataset in memory and performs computation only when a prediction is required. The choice of the hyper-parameter K (number of neighbors) is crucial in KNN. A smaller value of K leads to a more flexible decision boundary, potentially resulting in overfitting, while a larger value of K may lead to under-fitting. KNN is easy to implement and understand, making it an ideal choice for beginners and as a baseline model for comparison with more complex algorithms[2].

KNN is intuitive and easy to understand, requiring minimal assumptions about the data. Since KNN does not build an explicit model during training, the training phase is fast and computationally inexpensive. KNN can handle both binary and multi-class classification problems and is robust to noisy data[3].

KNN requires computing distances between the new instance and all instances in the training dataset, making it computationally expensive for large datasets. KNN is sensitive to outliers and noise in the data, which can affect the accuracy of predictions. KNN's performance deteriorates in high-dimensional feature spaces due to the curse of dimensionality[3].

SVM is a widely used supervised learning algorithm known for its effectiveness in classification and regression tasks. SVM has gained popularity for its ability to handle linear and non-linear classification problems efficiently. SVM aims to find the hyperplane with the maximum margin, which represents the distance between the support vectors of different classes. This property makes SVM less sensitive to outliers and improves its generalization ability. SVM utilizes kernel functions such as linear, polynomial, radial basis function, and sigmoid to handle non-linear decision boundaries by implicitly mapping the input space into a higher-dimensional feature space. SVM introduces slack variables to handle misclassification errors and soft-margin classifiers, allowing for some instances to be misclassified to achieve better overall performance. SVM often yields sparse solutions, meaning the decision boundary depends only on a subset of the training data, making it memory-efficient and suitable for large-scale datasets[4].

SVM performs well even in high-dimensional feature spaces, making it suitable for complex classification tasks such as image recognition and text classification. SVM's maximum margin property and ability to handle slack variables make it robust to overfitting, especially when using regularization techniques. SVM can be applied to both linear and non-linear classification problems by choosing appropriate kernel functions, providing flexibility in model selection[5].

SVM's training time complexity is quadratic with the number of training instances, making it less suitable for large-scale datasets. SVM is sensitive to noise and outliers, which can affect the placement of the decision boundary and degrade performance. The selection of the kernel function and its hyper-parameters can significantly impact the performance of SVM, requiring careful tuning[5].

2. LITERATURE REVIEW

This section summarizes related works in the domain of SA related to KNN and SVM classifiers.

Mohamed A. E. has presented the results of a comparative study of the four best-known machine learning and data mining techniques for classification, such as decision trees, artificial-neural network, K-nearest neighbors, and support vector machines. Research shows that each technique its own advantages and disadvantages, it is difficult to find one classifier can classify all the data sets with the same accuracy. Among all the learning algorithms on this particular dataset, the overall accuracy of the support vector machine is higher[6].

Naw N. were used SVM and KNN classifications to classify twitter data on education, business, crime and health. The aim is to measure the impact of social media usage behavior among ASEAN citizens. In this research, both these classifications are used to compare accuracy, precision, recall and f1 score. SVM outperforms KNN[7].

Huque Abu Sayeed A. et al. Evaluated and compared the performance of KNN, SVM and Sparse Representation Classifier (SRC) in identifying characters written in Arabic handwritten characters. The main purpose of this experiment is to recognize the isolated Arabic characters. The performance of the method is evaluated on a separate Persian/Arabic character dataset, which is a large dataset containing gray images. Experiments show that SRC and SVM consistently outperform KNN, with SVM achieving the highest recognition rate[8].

Fikri M. and Sarno R. implemented the rule with the help of SentiWordNet and SVM algorithms with the help of Term Frequency-Inverse Document Frequency (TF-IDF) as the feature extraction method. The data used to conduct the research is written in Indonesian. The oversampling method is used because the sentences in the positive, negative and neutral classes are imbalanced. Balancing dataset can improve the accuracy and F-Score of the SVM algorithm by using TF-IDF as the feature extraction method; however, data balance reduces the accuracy and F-Score of rule-based SentiWordNet. Using TF-IDF as the extraction method, the SVM algorithm achieves better results than the SentiWordNet rule[9].

İrfan M, et al. Comparison of KNN and SVM algorithms for high school recommendation selection. The method used in this study is data mining. KNN and SVM are algorithms widely used in data mining and decision support. By experimenting with many training and test data, the results show that SVM is better than KNN. The accuracy of SVM is about 97.1%, while the accuracy of KNN is about 88.5%. Moreover, the processing time of SVM is faster than KNN[10].

Fawzy H. and Mohamed A. aims at univariate time series prediction using SVM and KNN. Data are used contain 362 observations. SVM and KNN models fit 90% of the training data and their accuracy is then compared by RMSE test. The results show that SVM is better than KNN in predicting the future gold price[11].

Sudhir P. and Suresh V.D. Reviewed various applications, methods, and classification models used for sentiment analysis. Accuracy results of models based on the IMDB dataset show that machine learning such as SVM, GRU, and BERT show the most accuracy. More importantly, new models (e.g., GRU and BERT) have been shown to be more accurate than traditional classification models[12].

Nihalani R., et al. SVM and KNN algorithms were compared against the training data related to brest cancer, and the more accurate method was used to evaluate the final model using a 10-fold cross-validation practice. The SVM model achieved the highest accuracy during training and was further improved during testing to achieve an accuracy of 96.93%[13].

Utami L.D. and Masripah S. eager to address and address public views on distance learning and online education, which are sure to draw positive and negative opinions. There are many classification algorithms, including Naive Bayes algorithm (NB), KNN algorithm and SVM algorithm were used for text classification. After the calculation, the algorithm suitable for identifying reviews or opinions in this study is the SVM classification algorithm[14].

Desiani A. et al. compared the results of 2 methods and 2 training methods (e.g., cross-validation and percentile comparison), it was concluded that SVM and KNN classification models are effective in classifying breast cancer. Performance results show that SVM achieves a better accuracy when using the competitive evaluation method. The KNN classification model achieved better accuracy than SVM, which when using the percentage evaluation method[15].

Sutriawan, et al. compared the performance of NB, SVM, KNN and DT classification method for classification of positive and negative polarity behavior in Indonesian video analysis. Test results show that SVM algorithm type performs best accuracy. Therefore, SVM performs better than other classification methods[16].

3. METHODOLOGY

In this section, the general process of sentiment classification is explained. This research is an experimental study of sentiment analysis using KNN and SVM Classifier.

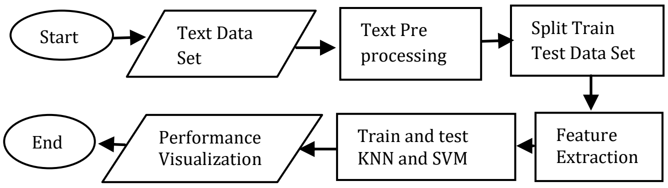
Start
Text Data Set
Text Pre processing
Split Train Test Data Set
Feature Extraction
Train and test KNN and SVM
Performance Visualization
End

Figure 1: Schema for performance evaluation of KNN and SVM

This research begins with data collection through kaggle public repository and then preprocessing the data through removal of retweets, removal of non-English words, removal of stop-words, removal of special characters, removal of digits, removal of unwanted words and removal of punctuation marks. Tokenizing, POS tagging, lemmatization is done . Divide the training and testing data with the proportion of 80% training data and 20% testing data after feature extraction carried out using the TF-IDF method. The training and testing for classification process begins with preparing models such as KNN and SVM ready for analysis, and then the models are run to make the models for sentiment prediction. The analysis phase of the model used is to evaluate the f1-score or f-measure, precision, recall, and accuracy. At this stage, performance visualizations are also made to display graphs and diagrams of data .

3.1 Working Principle of KNN Classifier

KNN classifies a new instance by examining the classes of its nearest neighbors in the feature space. During the training phase, the algorithm memorizes the entire training dataset without building an explicit model. In the prediction phase, when a new instance needs to be classified, the algorithm calculates the similarity (often using distance metrics like Euclidean or Manhattan distance) between the new instance and all instances in the training dataset. It then selects the KNN based on these similarities. The class label of the new instance is determined by a majority vote among its KNN[17].

3.2 Working Principle SYM Classifier

SVM is a discriminative classifier that constructs a hyperplane in a high-dimensional feature space to separate instances of different classes. The goal of SVM is to find the optimal hyperplane that maximizes the margin between the closest data points of different classes, known as support vectors. In linearly separable cases, SVM finds a hyperplane that perfectly separates the classes. In non-linearly separable cases, SVM uses kernel functions to map the input space into a higher-dimensional feature space where the classes become separable[18].

4. EXPERIMENTAL SETUP

This section contains the details about the experimental setup for the study. This section contains details about the dataset and matrices associated with the performance. The experiments have performed on python language with Scikit Learn, NLTK, and other library packages for implementation.

4.1 Datasets

Two datasets have used in this study. "all_tweets" dataset can be used for text analysis and analysis of sentiments and emotions through classification using machine learning or deep learning and natural language processing. "Financial Phrase Bank" dataset contains marketers' views on financial news. This file has two columns: "Sentiment" and "News Headlines". Sentiments are negative, neutral or positive.

Table 1: Descriptions of the datasets

Dataset	Data Set Description	Positive	Negative	Neutral
all_tweets [19]	Tweet Sentiment and Emotion Analysis	2974	796	2262
Financial Phrase Bank[20]	Sentiment Analysis for Financial News	1363	604	2879

4.2 Algorithm

Algorithm: Performance Evaluation of KNN and SVM classifiers
Input: Sentiment labeled text data
Output: Metric values and graphs
Process: Step1: Importing sentiment labeled text dataset Step2: Preprocessing the data Step3: Splitting the preprocessed dataset into training and testing data set Step4: Feature extraction using TF-IDF victimizer Step5: Train the KNN and SVM classifiers with the

```
training dataset
Step6: Calculating predicted labeled dataset using the trained KNN and SVM classier and feature test dataset
Step7: Calculating Matrices
Step 8: Plotting the graphs
```

4.3 Performance Measurements

Performance measurements in sentiment analysis are crucial for evaluating the effectiveness of the models. Some common performance metrics used in sentiment analysis include:

Precision: The ratio of the true positive instance to all instances predicted as positive. It measures the accuracy of positive predictions.

Recall : The proportion of true positive instances that were correctly identified out of all actual positive instances. It measures the model's ability to accurately identify positive instances.

Accuracy: The proportion of correct instances is classified for all instances. It provides an overall measurement of how well the model performs across all classes.

F1 Score: The F1 score is a commonly used metric in classification tasks, including sentiment analysis. It is the harmonic mean of precision and recall and provides a single metric that balances both precision and recall[21].

5. RESULTS AND DISCUSSIONS

5.1 RESULTS

This section contains the details about the results for sentiment classification of data sets. This section contains the tables, charts obtained while performing the sentiment analysis on the data using KNN and SVM classifiers.

Table 2: Performance metrics for the datasets

Metrics	all_tweets		Financial Phrase Bank	
	KNN	SVM	KNN	SVM
Precision	0.81	0.85	0.68	0.72
Recall	0.65	0.85	0.59	0.70
F1 score	0.64	0.85	0.46	0.66
Accuracy	0.65	0.85	0.59	0.70

Table 2 presents a tabular representation of the outputs. It shows the metrics for both the methods when applied over the two datasets individually.

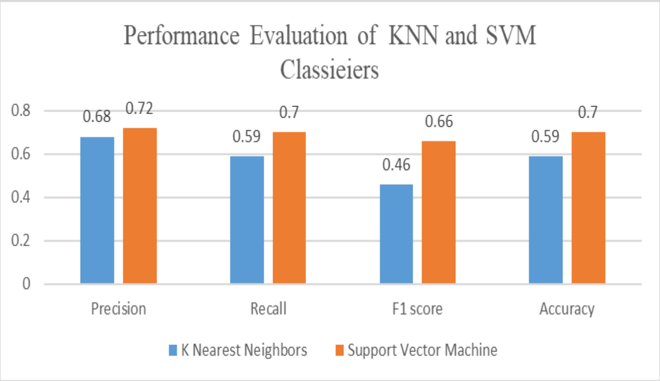

Figure 2: Performance comparison of KNN and SVM for all_tweets Dataset

Figure 2 shows all the metric results for the all_tweets dataset. This bar chart shows an overall comparison between two models of this particular data. SVM outperforms KNN in all parameters.

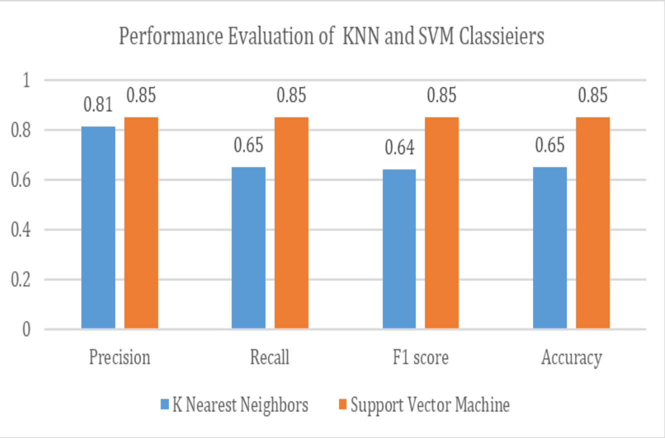

Figure 3: Performance comparison of KNN and SVM for Financial Phrase Bank Dataset

Figure 3 shows all the metric results for the Financial Phrase Bank dataset. This bar chart shows an overall comparison between two models of this particular data. SVM outperforms KNN in all parameters

Table 3: Average values of metrics for two datasets

Metrics	Average of two datasets	
	KNN	SVM
Precision	0.745	0.785
Recall	0.62	0.775
F1 score	0.55	0.755
Accuracy	0.62	0.775

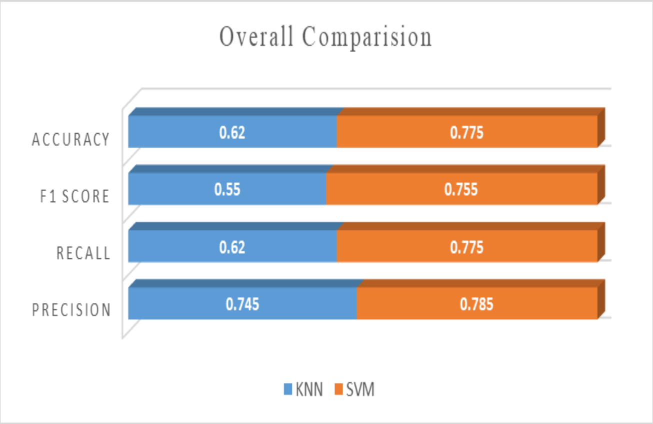

Figure 4: Versatility of KNN and SVM

To compare the versatility of KNN and SVM in sentiment analysis, the average of all parameters of SVM is higher than KNN in the two datasets. From table 3 and figure 4, SVM gives better results for all measurements compared to KNN when determining the average performance on two data sets.

5.2 Discussions

The current study found that average of precision, recall, accuracy and f1 score of KNN were 74.5%, 62%, 55%, 62.5 respectively and SMV were 78.5%, 77.5%, 75.5%, 77.5% respectively.

Mohamed A. E. used Weka software and German Credit data. He has found an accuracy of KNN as 71.3% and SMV as 76.3% which is consistent with the current study[6].

Naw N. used tweets of education, business, crime and health data. He has found precision, recall, accuracy and f1 score of KNN as 77.8%, 63.3%, 69.5%, 63.3% respectively and SVM as 71.3%, 70.8%, 74.4%, 70.8% respectively on Education data. He has found precision, recall, accuracy and f1 score of KNN as 75.8%, 59.6%, 70.5%, 68.1% respectively and SVM as 67.2%, 67.4%, 72.3%, 72.4% respectively on Business data. He has found precision, recall, accuracy and f1 score of KNN as 63%, 58.9%, 58.7%, 59.9% respectively and SVM as 69.4%, 67.6%, 72.1%, 71.7% respectively on Crime data. He has found precision, recall, accuracy and f1 score of KNN as 47%, 40.2%, 50.8%, 59.2% respectively and SVM as 58.4%, 57%, 70.9%, 64% respectively on health data[7]. on education, business, crime data, Naw N. study is consistent with the current study. On health data, Naw N. study is inconsistent with the current study. This inconsistency may be due to the health data.

Utami L.D. and Masripah S. used Rapid Miner 5.1 application with a dataset in the form of online learning reviews to carry out the analysis process. They found an accuracy of KNN as 86.33% and SVM as 87.67%[14]. This accuracy is consistent with SVM but inconsistent with KNN. This inconsistency may be due to the application and data.

Desiani A. et al. used breast cancer dataset. They found precision, recall, accuracy and f1 score of KNN as 97%, 98%, 97.85%, 97% respectively and SVM as 95%, 96%, 95.7%, 95.5% respectively [15]. This study is inconsistent with the current study. This inconsistency may be due to breast cancer dataset.

6. CONCLUSIONS

This study presents the results of a comparative study to investigate the two well-known machine learning methods for classification. Classification is important for organizing data so that it can be easily accessed. Each techniques is used in different areas and on different datasets. The performance of the learning algorithm depends on the nature of the data set. This research compares the results by analyzing sentiments of all_tweets dataset and financial phrase bank dataset using KNN classifier and SVM classifier. This study includes different metrics such as accuracy, precision, recall and f1 score, which help to clarify the comparison between the two methods. The results and discussions shows the performance of KNN and SVM on all_tweets dataset and financial phrase bank dataset is consistent with most of the other research . In this research, SVM outperform KNN..

In the future, KNN and SVM can also be tested on larger, more detailed data sets, which will help make better decisions. An ensemble of KNN and SVM could be developed that combines the results of the two methods and therefore improves performance.

REFERENCES

[1] K. P. Gunasekaran, "Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review".

[2] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[3] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," vol. 3, no. 5, 2013.

[4] M. Awad and R. Khanna, "Support Vector Machines for Classification," in Efficient Learning Machines, Berkeley, CA: Apress, 2015, pp. 39–66. doi: 10.1007/978-1-4302-5990-9_3.

[5] S. Amarappa and D. S. V. Sathyanarayana, "Data classification using Support vector Machine (SVM), a simplified approach".

[6] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," vol. 7, no. 2, 2017.

[7] N. Naw, "Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers," IJSRP, vol. 8, no. 10, Oct. 2018, doi: 10.29322/IJSRP.8.10.2018.p8252.

[8] A. S. A. Huque, M. Haque, H. A. Khan, A. Al Helal, and K. I. Ahmed, "Comparative Study of KNN, SVM and SR Classifiers in Recognizing Arabic Handwritten Characters Employing Feature Fusion," Sig.Img.Proc.Lett, vol. 1, no. 2, pp. 1–10, Jul. 2019, doi: 10.31763/simple.v1i2.1.

[9] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and SentiWordNet," IJEECS, vol. 13, no. 3, p. 902, Mar. 2019, doi: 10.11591/ijeecs.v13.i3.pp902-909.

[10] M. Irfan, A. R. Nurhidayat, A. Wahana, D. S. Maylawati, and M. A. Ramdhani, "Comparison of K-Nearest Neighbour and support vector machine for choosing senior high school," J. Phys.: Conf. Ser., vol. 1280, no. 2, p. 022026, Nov. 2019, doi: 10.1088/1742-6596/1280/2/022026.

[11] H. Fawzy and A. Mohamed, "Comparison between support vector machines and K-nearest neighbor for time series forecasting," jmcs, 2020, doi: 10.28919/jmcs/4884.

[12] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," Global Transitions Proceedings, vol. 2, no. 2, pp. 205–211, Nov. 2021, doi: 10.1016/j.gltp.2021.08.004.

[13] R. Nihalaani, "Comparison of K nearest Neighbours and Support Vector Machine to Build a Breast Cancer Prediction Model," IJRASET, vol. 9, no. 5, pp. 1435–1442, May 2021, doi: 10.22214/ijraset.2021.34559.

[14] L.D. Utami .and S. .Masripah, "COMPARATION OF CLASSIFICATION ALGORITHM ON SENTIMENT ANALYSIS OF ONLINE LEARNING REVIEWS AND DISTANCE EDUCATION," 2021.

[15] A. Desiani, A. A. Lestari, M. Al-Ariq, A. Amran, and Y. Andriani, "Comparison of Support Vector Machine and K-Nearest Neighbors in Breast Cancer Classification," Pattimura Int. J. Math., vol. 1, no. 1, pp. 33–42, May 2022, doi: 10.30598/pijmathvol1iss1pp33-42.

[16] S. Sutriawan, P. N. Andono, M. Muljono, and R. A. Pramunendar, "Performance Evaluation of Classification Algorithm for Movie Review Sentiment Analysis," IJC, pp. 7–14, Mar. 2023, doi: 10.47839/ijc.22.1.2873.

[17] I. N. Chikalkar, "K -NEAREST NEIGHBORS MACHINE LEARNING ALGORITHM," vol. 8, no. 12, 2020.

[18] C. H. Miller, M. D. Sacchet, and I. H. Gotlib, "Support Vector Machines and Affective Science," Emotion Review, vol. 12, no. 4, pp. 297–308, Oct. 2020, doi: 10.1177/1754073920930784.

[19] "Tweet Sentiment and Emotion Analysis | Kaggle." Accessed: Feb. 13, 2024. Available: https://www.kaggle.com/datasets/subhajournal/tweet-sentiment-and-emotion-analysis

[20] "Sentiment Analysis for Financial News | Kaggle." Accessed: Feb. 13, 2024. Available: https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news

[21] D. P. Sengottuvelan and I. A. Regina, "Performance Evaluation of Machine Learning Models for Prediction of Sentiments in Movie Reviews," JOURNAL OF CRITICAL REVIEWS, vol. 7, no. 08, 2020.