

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

**OPTIMIZING MOVIE REVIEW SENTIMENT ANALYSIS WITH MACHINE LEARNING
AND NATURAL PROCESSING LANGUAGE TECHNIQUES FOR IMPROVED
ACCURACY**

Akshay Rajshekar Shiraguppi
University of Illinois at Springfield
Springfield, Illinois
akshayrs1993@gmail.com

ABSTRACT

Businesses may learn a lot about how the public feels about their service, movie, product, etc. by using sentiment analysis. With Sentiment Analysis, businesses may learn how people feel about their product or movie without reading the whole review or post. This study leverages deep learning techniques, specifically the Bidirectional Gated Recurrent Unit (BiGRU), to enhance sentiment classification accuracy on the IMDb dataset. A comprehensive pre-processing pipeline, including tokenization, stop-word removal, and lemmatization, is applied to refine textual data. The Bag of Words (BoW) model is used for feature extraction, then the BiGRU model is trained and evaluated. Experimental results demonstrate that BiGRU significantly outperforms traditional models such as RNN-LM and Naïve Bayes, achieving an accuracy of 98%, with an accuracy rate is 98.33%, a recall rate is 98.91%, and an F1-score is 98.23%. The ROC curve with an AUC of 1.00 further validates its superior performance. Despite these promising results, limitations include high computational demands and dataset dependency.

Keywords: *Social media, Movie review, Sentiment analysis, Text classification, NLP, Deep learning, IMDB dataset.*

I. INTRODUCTION

In the digital age, social media platforms, online forums, and review websites have become primary sources for sharing opinions and feedback. Users express their thoughts on products, services, and experiences, generating vast amounts of unstructured textual data [1]. This abundance of user-generated content provides valuable insights for businesses, marketers, and researchers aiming to understand public sentiment and improve their offerings. Reviews have a major impact on consumer decisions among various online feedback forms [2]. Whether it is a product review on an e-commerce site, a service review on social media, or a user testimonial on a business website, these reviews help assess public perception.

In particular, movie reviews have emerged as a crucial tool in the entertainment industry,

International Journal Of Core Engineering & Management (IJCEM) **Volume 2, Issue 12, March 2016**

influencing audience choices, box office performance, and marketing strategies. Moviegoers share their opinions on films through social media posts, blogs, and dedicated review platforms, offering qualitative insights into different aspects such as storyline, acting, cinematography, and music. Analyzing such reviews manually is time-consuming and inefficient due to the sheer volume of data [3]. Sentiment analysis is a branch of NLP that deals with and addresses this challenge by automating the extraction of emotions and opinions from textual data [4]. By classifying sentiments, Sentiment analysis helps us understand audience preferences and reception better by classifying their feelings into categories that are neutral, negative, or favorable. In the context of movie reviews, it helps filmmakers, production companies, and streaming platforms assess public opinion, identify trends, and refine their content strategies [5].

Machine learning (ML) has significantly improved sentiment analysis by enhancing classification accuracy and contextual understanding, addressing challenges that traditional rule-based and lexicon-based approaches face with complex language structures, sarcasm, and contextual sentiment shifts [6]. ML and deep learning models have demonstrated superior performance in sentiment classification [7].

A. Motivation and Contribution of this Study

This work is driven by the necessity of an efficient and accurate sentiment analysis system for IMDB movie reviews, as manual classification is subjective, time-consuming, and prone to inconsistencies. By leveraging advanced natural language processing (NLP) techniques and deep learning models, this study aims to automate sentiment classification and enhance the accuracy of review-based decision-making. The following contributions of the paper are:

- Implemented an efficient text preprocessing approach, including HTML tag removal, tokenization, stop-word removal, stemming, and lemmatization, to enhance data quality for sentiment analysis.
- Utilized the Bag of Words (BoW) model to transform textual data into numerical representations, facilitating effective machine learning classification.
- Applied state-of-the-art ML models, including Bidirectional Gated Recurrent Unit (BiGRU), for accurate sentiment prediction.
- Assessed model effectiveness using F1-score, recall, accuracy, precision, and confusion matrices to examine patterns of misclassification.

B. Structure of the paper

This is how the study is set up: Section II lists the body of research on movie reviews. To gather the data for this investigation, the technique was applied in section III. Section IV provide the results and analysis of text classification. At last, Section V provide the conclusion provides the conclusion.

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

II. LITERATURE REVIEW

This section discusses the Literature review on Movie review sentiment analysis with machine learning model Also, Table I provides the summary of these literature reviews discussed below:

Ouyang et al. (2015) Obtain a vector representation of a word and reflect its distance using word2vec, a Google-proposed method for computing word vector representations. This design applies PReLU, normalization, negative-positive, and word2vec and CNN with three pairs of convolutional and pooling layers for sentiment analysis that is negative-positive. When compared to other neural network models, such as RNN and MV-RNN, our network outperforms them despite its poor accuracy [8].

Nagamma et al. (2015) Predicting box office sales using online review data may be achieved with high accuracy using a sentiment-aware autoregressive model. This resulted in a less complex model that may be easier to train and employ. Based on the tone of reviews, one may predict future box office revenues by creating a classification model with the use of an SVM classifier [9].

Parkhe and Biswas (2014) The mood expressed by the user regarding the film may be understood in large part through sentiment analysis of a review. It uses sentiment analysis based on aspects of movie reviews to identify the elements that drive each aspect. According to the experiment, the maximum accuracy in the analysis of movie reviews was achieved by assigning high-driving elements to the film's acting, plot, and movie [10].

Mouthami, Devi and Bhaskaran (2013) To convey a good or negative feeling, sentiment categorization is utilized. a novel method called to increase classification accuracy on the benchmark dataset for the proposed study, the Sentiment Fuzzy Classification technique using parts of speech tags is utilized. A dataset of movie reviews [11].

Basari et al. (2013) focus on binary categorization, which divides people into two groups. Both positive and negative classifications are included. To tackle the dual optimization issue, the best parameter is chosen more effectively by using the hybrid PSO. The outcome demonstrates that the accuracy level increased from 71.87% to 77% [12].

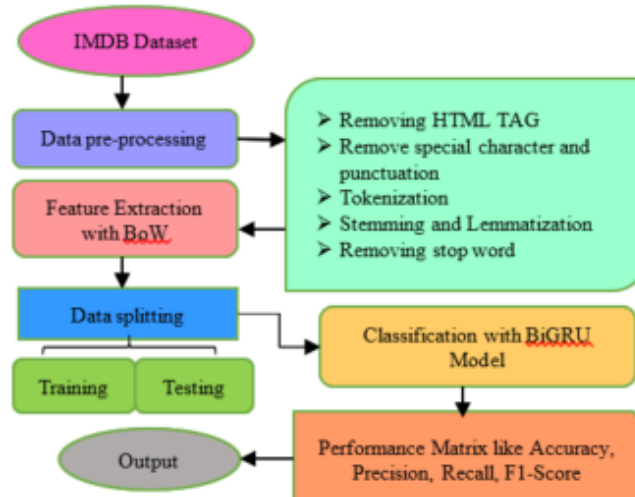
P.Kalaivani and K.L.Shunmuganathan (2013) Sentiment analysis techniques were used to analyze movie reviews, and sentiment classification, feature-based classification, and managing negotiations are the most common community online movie reviews. According to empirical results, the SVM strategy achieved accuracies of at least 80% and beat the kNN and Naive Bayes methods. There were many reviews in the training dataset [13].

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

An overview of various sentiment analysis approaches applied to movie reviews, highlighting datasets, methodologies, performance, and limitations shown in Table I.

**TABLE I. OVERVIEW OF SENTIMENT ANALYSIS APPROACHES IN MOVIE REVIEWS
USING MACHINE LEARNING**

Author	Dataset	Methodology	Performance	Limitations
Ouyang et al. (2015)	Word2Vec-generated dataset	Word2Vec + CNN with PReLU and normalization	Higher accuracy than RNN and MV-RNN	Accuracy still low; potential improvements with hybrid deep learning models
Nagamma et al. (2015)	Online review data	Sentiment-aware autoregressive model with fuzzy clustering using SVM and TF-IDF	Effective in predicting box office sales	Could be extended with deep learning for enhanced accuracy
Parkhe & Biswas (2014)	Movie reviews dataset	Aspect-based sentiment analysis (Movie, Acting, and Plot aspects)	Highest accuracy with aspect-driven factors	Needs evaluation with other aspects like cinematography and direction
Mouthami, Devi & Bhaskaran (2013)	Movie reviews dataset	Sentiment Fuzzy Classification with POS tags and SVM	Improved classification accuracy	Multi-theme document analysis remains challenging
Basari et al. (2013)	Movie reviews dataset	SVM for parameter optimization using PSO and 10-fold cross-validation	Accuracy improved from 71.87% to 77%	Potential for further enhancement using deep learning techniques
P. Kalaivani & K.L. Shunmuganathan (2013)	Online movie reviews	Supervised learning (kNN, SVM, and Naïve Bayes)	SVM outperformed Naïve Bayes and kNN with at least 80% accuracy	Increasing dataset size may further improve accuracy



III. METHODOLOGY

The methodology for movie review sentiment analysis natural language processing (NLP) techniques follows a systematic approach to enhance classification accuracy. The publicly available IMDb dataset, comprising 50,000 labeled reviews (25,000 positives and 25,000 negatives), is utilized. To normalize textual data, data preparation includes removing HTML elements, special characters, and punctuation. This is followed by lemmatization, stemming, tokenization, and stop-word elimination. Feature extraction uses the BoW model to convert text into numerical vectors for machine learning-based categorization. The dataset is split 80:20 between testing and training to ensure generalizability. F1-score, accuracy, precision, and recall are used to evaluate BiGRU, which is used to categorize sentiment. Optimization involves hyperparameter tuning, advanced word embeddings, and ensemble learning to enhance classification robustness and accuracy. The systematic methodology steps are shown in Figure 1.

The suggested flowchart's further phases are listed below:

A. Data Collection

Sentiment analysis and natural language processing applications frequently use the publicly available IMDb dataset. The reviews contain the label of the movie (whether good or bad). Twenty-five thousand reviews with positive tags and twenty-five thousand reviews with negative tags are included in the dataset. It is sourced from the popular movie database IMDb. Figure 2 displays the following word cloud of good reviews.

International Journal Of Core Engineering & Management (IJCEM)

Volume 2, Issue 12, March 2016

- **Removing HTML Tags:** IMDB evaluations frequently include HTML elements that are unnecessary for sentiment analysis. Python's BeautifulSoup library or regular expressions are used to eliminate them.
- **Removing special characters and punctuation:** The dataset's odd letters and punctuation that don't contribute to sentiment analysis are eliminated using regular expressions and Python's string module.
- **Tokenization:** Phrases can be divided into distinct portions via tokenization according to words, characters, and subwords. Additionally, tokenization may be used to identify phrases that are often used throughout the whole data collection.
- **Stemming and Lemmatization:** These techniques minimize lowering dataset dimensionality and enhancing model performance by reducing the number of variants of a word that are inflectional and sometimes derivatively tied to a single base form. While stemming is the process of eliminating a word's suffixes, lemmatization is the act of restoring a word to its dictionary or base form.
- **Removing stop word:** To improve model performance and decrease dataset dimensionality, common terms like "the," "and," and "is" are removed out.

B. Feature Extraction

In NLP, one popular and simple feature extraction technique is the BoW. It describes the appearance of every word in the text. Word order and grammar are ignored in favour of building a matrix of occurrences for a sentence or text. A corpus (or set collection) of words is used in the Bow encoding, which employs a vector of the corpus' length to represent any given text.

C. Data Splitting

The resampled dataset is separated use the train test split to separate training and testing sets. 80:20 is the splitting ratio.

D. Proposed Bidirectional GRU (BiGRU) Model

GRU is an LSTM variation as nodes go farther back, the perception of the nodes in front decreases because RNN has a significant gradient disappearing issue while processing sequences. The LSTM neural network is developed to overcome the gradient disappearance problem. a GRU was proposed to address the issues of multiple parameters and lengthy training times. In order to solve the gradient disappearing issue, it may also process sequential input and employ "gate" to memorize the information of previous nodes. Because the GRU only has two gates—an update gate and a reset gate—it has fewer parameters than the LSTM [14]. This can cut down on training time while producing the same result as LSTM. In the following way, x represents the input data, h_t the GRU's output, r_t the reset gate, and z_t the update gate. Together, z_t and r_t govern the calculation from the

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

h_{t-1} hidden state to the h_t hidden state [15]. The update gate controls the previous memory information $h_{(t-1)}$ as well as the current input data by producing a value z_t that ranges from 0 to 1. z_t determines the quantity of h_{t-1} communicated to the subsequent stage. The particular gate unit is determined using the Eq. (1 to 4):

$$z_t = \sigma(w_z \cdot [h_{t-1}, x_t]) \quad (1)$$

$$r_t = \sigma(w_r \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\bar{h}_t = \tanh(W \cdot [r_t \times h_{t-1}, x_t]) \quad (3)$$

$$h_t = (1 - z_t) \times h_{t-1} \times \bar{h}_t \quad (4)$$

This formula uses the Sigmoid function, σ , as a gate control signal to transform the data into a value between 0 and 1. Because GRU is a one-way neural network structure, even if the output state is sent from the front row to the rear row, it is still linked to the states that came before and after it. To tackle this issue, a bidirectional GRU is required. The BiGRU model is capable of extracting information in both directions.

E. Performance Metrics

To evaluate the performance models, used a collection of assessment criteria. A confusion matrix is a table that evaluates the model's performance by contrasting the expected and actual results. The resultant models were evaluated using five assessment metrics: F1-score, recall, accuracy, and precision. Initially, confusion matrix classes: False Negatives (FN) show how many examples from the real class were mistakenly projected to be another class, whereas True Positives (TP) show how many instances were properly predicted to be the actual class. The number of records successfully identified as normal is known as True Negative (TN), whereas the number of cases from other classes that were mistakenly projected to belong to the actual class is known as False Positive (FP).

1. Accuracy

This statistic determines the percentage of cases in a dataset that are properly classified relative to all occurrences, indicating how well instances are classified within the dataset. It is calculated as Eq. (5):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (5)$$

2. Precision

Precision is the proportion of comparable cases that occur between the retrieved examples; it is often referred to as instant from doubt predictive value. It is calculated as Eq. (6):

$$Precision = \frac{TP}{TP+FR} \times 100 \quad (6)$$

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

3. Recall

In contrast, the proportion of complementary cases that have been recovered overall, or the total number of relevant examples, is called extract (also called sensitivity). It is formulated as Eq. (7):

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (7)$$

4. F1 Score

A metric that combines accuracy and recollection is called the harmonic means of accuracy and recall, often referred to as the conventional F-measure, which balances recall and accuracy. It is computed as Eq. (8):

$$F1 - score = 2 * \frac{recall \times precision}{recall + precision} \quad (8)$$

5. ROC Curve

The model's capacity to differentiate between the classes is assessed by plotting the receiver operating characteristic (ROC) curve and AUC score.

IV. RESULT ANALYSIS AND DISCUSSION

The experimental results for movie review utilizing machines and This section display deep learning on the IMDb dataset model. The performance indicators include recall, accuracy, precision, and the F1-score for text classification. Python programming language, Jupyter Notebook, Google colab, and Python Keras, pandas, NumPy, seaborn, tensor-flow and matplotlib etc. To satisfy the computational demands of the suggested models, library toolboxes were loaded onto the NVIDIA GTX 1660i GPU, which has 8 GB of VRAM and 16 GB of RAM. The following sections provide the results of proposed modes for Movie Review using the BiGRU model.

TABLE II. FINDINGS OF THE BIGRU MODEL ON MOVIE REVIEW SENTIMENT ANALYSIS

Performance Matrix	Bidirectional GRU (BiGRU)
Accuracy	98.2
Precision	98.33
Recall	98.91
F1-score	98.23

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

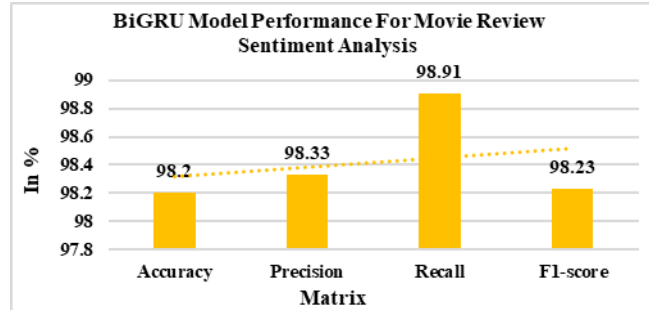


Fig. 4. Bar Graph for BiGRU Model Performance

Table II and Figure 4 present the performance of the BiGRU model on the IMDb dataset. The BiGRU model performed exceptionally well, with a 98.2% accuracy rate. The model exhibited strong precision at 98.33%, indicating its effectiveness in minimizing FP. With a recall of 98.91%, it efficiently captured relevant instances, showcasing its ability to detect positive cases accurately. The model's balanced accuracy and recall performance is demonstrated by its F1-score of 98.23%, which makes it extremely dependable for classification tasks. These metrics collectively validate BiGRU's robustness and effectiveness in handling sequential data with high predictive accuracy.

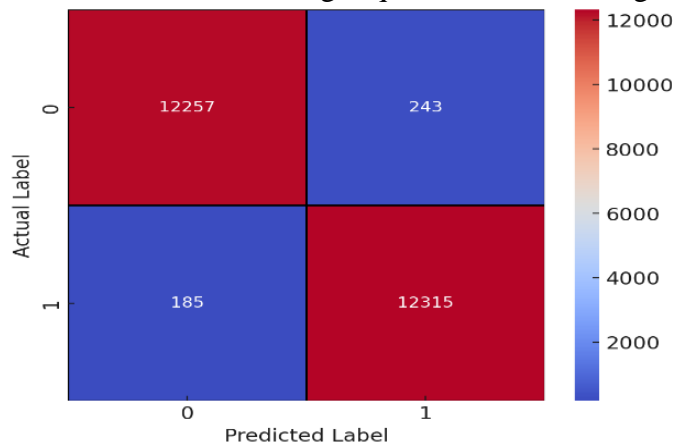


Fig. 5. Confusion Matrix of BiGRU Model

Figure 5 shows the BiGRU model's confusion matrix for the classification model. False Positives (243) are in the top-right quadrant, True Negatives (12,257) are in the top-left, The bottom-right quadrant of the matrix has True Positives (12,315), whereas the bottom-left quadrant contains False Negatives (185).

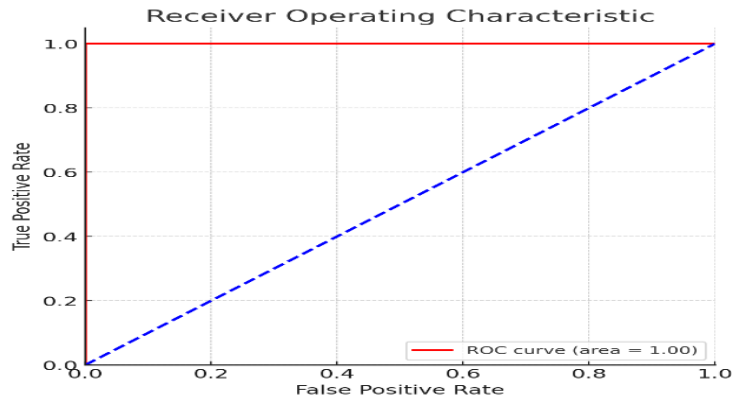


Fig. 6. Roc Curve of BiGRU Model

Figure 6 represents an illustration of the BiGRU model. The blue curve in the Receiver Operating Characteristic (ROC) curve for movie reviews displays the True Positive Rate against the False Positive Rate; an Area Under the Curve (AUC) of 1.00 indicates a high accuracy.

A. Comparative Analysis

The comparative analysis for Movie Review sentiment analysis on the IMDb dataset is provided in this section. The comparison between the BiGRU proposed and the existing RNN-LM [16] and Naïve Bayes [17] In Table III, models are shown based on matrix-like accuracy, precision, recall, and f1-score parameters.

TABLE III. COMPARISON BETWEEN BIGRU AND EXISTING MODELS FOR MOVIE REVIEW SENTIMENT ANALYSIS

Models	Accuracy
BiGRU	98
RNN-LM[16]	86.6
Naive Bayes[17]	89.4

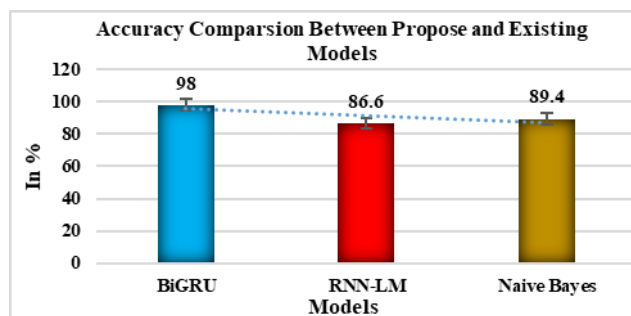


Fig. 7. Bar Graph for Comparison of Accuracy

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

Table III and Figure 7 present a comparative analysis of the BiGRU model against current sentiment analysis methods for movie reviews. BiGRU outperformed the RNN-LM model by achieving the greatest accuracy of 98%, which attained an accuracy of 86.6%, and the Naïve Bayes classifier, which reached 89.4%. The superior performance of BiGRU highlights its ability to effectively capture contextual dependencies in text, demonstrating its robustness in sentiment classification tasks compared to traditional language models and probabilistic classifiers.

V. CONCLUSION & FUTURE WORK

The study of the underlying sentiments and views represented in texts is called sentiment analysis. A potent technique for expressing the viewpoints of any person, organization, or society is data sentiment analysis. In the past, ML and NLP were employed to examine consumer data. The model's sentiment analysis abilities are assessed using ML and DL approaches on the IMDB collection of movie reviews. The experimental results demonstrate that the BiGRU model outperforms existing algorithms in movie review sentiment analysis on the IMDB dataset, with an accuracy of 98%. This is noticeably higher than RNN-LM (86.6%) and Naïve Bayes (89.4%). The model's high accuracy (98.33%), recall (98.91%), and F1-score (98.23%) indicate its ability to properly categorize attitudes while minimizing false predictions. The ROC curve's excellent predictive power is further supported by its AUC of 1.00. The examination of the confusion matrix shows a balanced classification performance. However, certain limitations exist, such as the model's reliance on extensive computational resources and sensitivity to imbalanced data. Additionally, the use of a single dataset (IMDb) restricts generalizability to broader sentiment analysis tasks. Future studies can explore hybrid deep learning architectures incorporating attention mechanisms or transformers for enhanced context understanding.

REFERENCES

1. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, 2014, doi: 10.1016/j.asej.2014.04.011.
2. V. K. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment Analysis of Movie Reviews: A New Feature-Based Heuristic for Aspect-Level Sentiment Classification," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, IEEE, Mar. 2013, pp. 712–717. doi: 10.1109/iMac4s.2013.6526500.
3. H. Ghorbel and D. Jacot, "Sentiment Analysis of French Movie Reviews," in *Studies in Computational Intelligence*, 2011, pp. 97–108. doi: 10.1007/978-3-642-21384-7_7.
4. M. Kanakaraj and R. M. R. Guddeti, "Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, IEEE, Feb. 2015, pp. 169–170. doi: 10.1109/ICOSC.2015.7050801.

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

5. Tun Thura Thet, J.-C. Na, and C. S. G. Khoo, "Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards," *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, Dec. 2010, doi: 10.1177/0165551510388123.
6. V. Kolluri, "An In-Depth Exploration of Unveiling Vulnerabilities: Exploring Risks in AI Models and Algorithms," *Int. J. Res. Anal. Rev.*, vol. 1, no. 3, 2014.
7. H. Shirani-Mehr, "Applications of Deep Learning to Sentiment Analysis of Movie Reviews," Stanford Univ., 2015.
8. X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "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, Oct. 2015, pp. 2359–2364. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.
9. P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," in International Conference on Computing, Communication & Automation, IEEE, May 2015, pp. 933–937. doi: 10.1109/CCAA.2015.7148530.
10. V. Parkhe and B. Biswas, "Aspect based sentiment analysis of movie reviews: Finding the polarity directing aspects," in Proceedings - 2014 International Conference on Soft Computing and Machine Intelligence, ISCMi 2014, 2014. doi: 10.1109/ISCMi.2014.16.
11. K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in 2013 International Conference on Information Communication and Embedded Systems (ICICES), IEEE, Feb. 2013, pp. 271–276. doi: 10.1109/ICICES.2013.6508366.
12. A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," in Procedia Engineering, 2013. doi: 10.1016/j.proeng.2013.02.059.
13. P.Kalaivani and D. K.L.Shunmuganathan, "sentiment classificationof movie review By supervised machine learning algorithm," P.Kalaivani et.al / Indian J. Comput. Sci. Eng., 2013.
14. B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
15. T. Du and V. K. Shanker, "Deep Learning for Natural Language Processing," *Electr. Comput. Eng. Comput. Inf. Sci.*, 2013.
16. G. Mesnil, T. Mikolov, M. A. Ranzato, and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," in 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings, 2015.
17. V. K. Singh, R. Piryani, A. Uddin, P. Waila, and Marisha, "Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches," in 2013 5th International Conference on Knowledge and Smart Technology (KST), IEEE, Jan. 2013, pp.



International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 12, March 2016

122–127. doi: 10.1109/KST.2013.6512800.