SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I11P121

Volume 13 Issue 11, 342-363, November 2026
© 2026 Seventh Sense Research Group®

Review Article

Enhanced Sentiment Analysis of Product Reviews using a Hybrid DBN-GRU Model

Siva Rama Prasad Kollu[1], Yugandhar Garapati[2]

[1,2]*Department of CSE, GITAM Deemed to be University, Hyderabad, India.*

[1]*Corresponding Author: kollusivaramaprasad@gmail.com*

Received: 26 May 2026 Revised: 08 September 2026 Accepted: 17 November 2026 Published: 27 November 2026

Abstract - *Product review Sentiment analysis is an important facet of interpreting customer reviews, helping businesses enhance their products and user experience. Traditional machine learning methods, such as Support Vector Machines (SVM) and Naïve Bayes, often rely on handcrafted features and fail to capture complex contextual dependencies in textual data, leading to suboptimal performance. Deep learning approaches like CNNs and LSTMs have improved sentiment classification by learning hierarchical representations, but they struggle with input noise and long-term dependencies in text sequences. To overcome these drawbacks, this study proposes a hybrid DBN-GRU model, which has recently been used for binary sentiment classification in a large-scale three-class (positive, neutral, negative) product-review classification task. The study does not propose a new architecture; rather, it presents a three-class formulation, component-level ablations, and more extensive validation of the framework. The DBN first learns unsupervised features from review embeddings, capturing higher-level representations and filtering out redundant information. These learned representations are then fed into a GRU network designed to learn sequential and contextual dependencies between review tokens. The model performs final sentiment classification using a softmax layer and trains via backpropagation with the Adam optimiser. The DBN-GRU framework was evaluated on a case study of 568,454 product reviews, and the results were competitive with the baseline method, with precision of 95%, accuracy of 95%, recall of 98%, and an F1-score of 96.47%. Based on these results, the DBN-GRU model has strong potential for three-class product review sentiment classification and can be applied in various fields such as e-commerce, recommendation systems, and market research.*

Keywords - *DBN, GRU, Hybrid deep learning model, Sentiment analysis, Product reviews.*

1. Introduction

In the digital economy, online product reviews have become one of the most potent factors in consumer decision-making and brand image development [1]. As e-commerce sites grow exponentially, millions of customers share their opinions daily through textual reviews, ratings, and feedback [2]. This large volume of user-generated content holds valuable insights that, when analysed appropriately, can help businesses understand customer sentiment, improve where needed, and tailor their marketing approach [3]. Nonetheless, manually interpreting such large volumes of data is not feasible; therefore, automated sentiment analysis has become an essential NLP application [4]. Traditional sentiment analysis techniques, including lexicon-based approaches and conventional machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), and Decision Trees, have been widely adopted for product review classification. Although these methods perform well on small, well-structured datasets, they rely heavily on handcrafted feature engineering and struggle to capture contextual semantics, long-range dependencies, sarcasm, negation, and subtle sentiment variations common in real-world customer reviews [5]. Consequently, their performance deteriorates when processing large-scale, noisy, and dynamically evolving e-commerce datasets.

Although these advancements have been made, it remains unclear how to combine hierarchical feature representation learning with lightweight contextual sequence modelling for fine-grained product-review sentiment classification. While existing CNN-based, LSTM-based, and Transformer-based approaches focus on extracting semantic or contextual representations, comparatively little work has addressed unsupervised hierarchical representation refinement before recurrent sequence modelling. This is also important for three-class sentiment analysis, where neutral/mixed sentiment is hard to distinguish from positive and negative reviews. Thus, it is essential to examine the effectiveness of DBN-based representation refinement and GRU-based contextual modelling in the sentiment classification task without compromising computational efficiency. Although deep learning models have made great progress, current models

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

have several key drawbacks. CNN-based models mainly emphasise local semantic patterns, and LSTM-based models are highly computational and lengthy in terms of training time. More importantly, the existing works have focused on either feature extraction or sequential learning and are difficult to achieve both simultaneously: reducing noisy representations and keeping the contextual information, especially in the case of ambiguous and neutral customer reviews, which motivates the study of hybrid models that combine semantic feature learning with modelling the sequential dependency.

In this work, we investigate the DBN-GRU combination for the three-class problem and benchmark its performance against selected baseline methods on a large product-review dataset-class-wise results. Leveraging the strong representational power of DBNs and the sequence-modelling strength of GRUs, the hybrid architecture is designed to improve sentiment classification. The proposed solution is evaluated on benchmark product review datasets to compare its performance with Conventional Deep Learning architectures. Results show that the hybrid model, DBN-GRU, yields high accuracy, precision, recall and F1-score. Class-wise results are analysed for the more challenging class, neutral, and ablation analysis is conducted. The framework is tested for practical use in sentiment monitoring, recommendation, and market analysis.

1.1. Research Motivation

The rapid expansion of e-commerce has generated an overwhelming number of product reviews, and businesses must extract meaningful insights to enhance customer satisfaction and decision-making. However, current sentiment analysis techniques have limitations in capturing subtle emotions, handling noisy text, and modelling contextual relationships in review text [9]. Classical machine learning models rely on handcrafted features, which limit scalability and generalisation. In contrast, DL models such as CNNs and LSTMs can struggle to capture temporal context or incur high computational costs [10]. These constraints underscore the need for a more efficient, robust framework that integrates effective feature learning with sequential modelling. To address this gap, this study investigates an existing hybrid DBN-GRU architecture that learns hierarchical semantic representations before modelling temporal patterns for three-class sentiment classification of product reviews.

1.2. Research Significance

This research is significant because it systematically evaluates the DBN-GRU framework for three-class sentiment analysis of product reviews, which is relevant to customer experience management, product development, and business decision-making [11]. The framework uses Deep Belief Networks to extract features and Gated Recurrent Units to model time dependencies. Its ability to extract high-level semantic structures and sequential dependencies aims to improve classification, particularly for neutral sentiments, which are often misclassified. Beyond contributing to sentiment analysis research, this study offers practical value to e-commerce sites, recommendation systems, and market analysts who need to derive quantifiable information about a specific market from large volumes of user-generated content.

1.3. Research Gap

While sentiment analysis has progressed significantly, significant gaps remain in capturing the full diversity of user-generated reviews. Traditional machine learning algorithms fail to generalise to different and noisy datasets because they require manual feature creation [12]. Deep learning techniques like CNNs and LSTMs have improved sentiment classification. CNNs do not capture long-term dependencies, and LSTMs perform well but are also computationally costly, particularly for large datasets [13]. Moreover, existing models struggle with padding when handling fine-grained differences in sentiment, sarcasm, and context-dependent phrases, leading to incorrect sentiment classification for non-neutral or mixed-sentiment reviews. CNN-based, LSTM-based, and Transformer-based sentiment classification models have been explored and tested. However, few studies have explored combining unsupervised hierarchical feature extraction with simplified recurrent sequence modelling in a unified model. Most recent methods either model semantic representations or model dependencies between sequence elements, and they ignore the optimisation of both and the possibility of low complexity.

To address this gap, this research aims to study the DBN-GRU, particularly for sentiment classification of product reviews with three classes. The current sentiment formulation does not use binary classification, as in very similar DBN-GRU models; instead, it classifies as positive, neutral, or negative, which makes it difficult for the model to learn more subtle sentiment boundaries. The methodological novelty of this work is leveraging DBN-based hierarchical representation refinement for the three-class scenario before GRU-based contextual sequence modelling, and conducting systematic component-level ablation and extensive validation. Thus, the DBN-GRU architecture is not new. Despite this, this study examines the performance of both representation learning and contextual modelling for fine-grained three-class sentiment classification in product reviews.

1.4. Key Contributions

This study has made the following contributions:

1. Finer-grained sentiment classification: The sentiment classification in product reviews is formulated as a three-class problem consisting of positive, neutral and negative sentiments, allowing assessment of finer-grained distinctions in sentiments than a binary classification.
2. Text classification using representation refinement with DBN and GRU sequence modelling: The study examines the ability of hierarchical DBN representation learning as a feature-refinement stage before GRU sequence

modelling for text classification of three similar sentiment classes, aiming to obtain compact and semantically meaningful representations.

3. Component-level methodological analysis: The benefit of the DBN representation learning stage is measured in the ablation experiments, and the contribution of the components of the feature is quantified.
4. Comprehensive validation: Repeated 5-fold cross-validation, bootstrap confidence intervals, calibration analysis, decision-curve analysis, and conventional classification performance metrics are employed to establish these measures beyond just test-set accuracy.

Existing sentiment analysis literature primarily focuses on stand-alone CNN, LSTM, GRU, or Transformer architectures to extract interpretative information. However, the DBN+GRU framework proposed in this study overcomes the limitations of these approaches by unifying the hierarchical semantic information provided by a DBN with the sequential context-modelling capability of a GRU. The DBN stage rescales features into a less noisy, less redundant space to help the GRU focus on contextual relations. We evaluate this design empirically by improving classification for neutral sentiment, using ablation and class-wise analysis in Section 4.

While the DBN-GRU architecture has been explored in the past, such as by Khan et al. [23], this work uses the architecture to classify sentiment from product reviews in the context of three classes and a large number of documents. This work does not introduce a new DBN-GRU architecture; instead, it systematically studies combining DBN-based hierarchical representation learning with GRU-based contextual sequence modelling on a significantly larger dataset and a three-class sentiment task. We also perform component-level ablation analysis to assess each component's contribution.

This study applies hierarchical DBN representation refinement and subsequent GRU contextual modelling to sentiment classification for three-class product reviews. To classify into three broad classes, we add an explicit third class between positive and negative sentiment, making the model harder to train to distinguish sentiment representations that are closer together. Furthermore, the study compares performance using only the DBN representation-learning stage with that of the final hybrid model to assess the quality of the representation-learning stage. The framework also uses repeated cross-validation and bootstrap confidence intervals, along with decision curve analysis and calibration, to evaluate performance. These methodological elements present a more detailed evaluation of the DBN-GRU framework than a single hold-out performance comparison.

1.5. Novelty and Positioning

The proposed framework combines representation learning with a sequence model for sentiment analysis of product reviews. Current methods discussed in this study mostly use conventional machine Learning, CNN-based Feature Extraction, Recurrent-based (LSTM/GRU) or Transformer-based contextual representations. The proposed framework is a methodological investigation of DBN-GRU for three-class product-review sentiment classification, rather than a proposal for a new DBN-GRU architecture.

Previous studies have reported this architectural composition, combining DBN-based representation learning with a GRU-based sequence modelling mechanism. The present study is a further exploration of three sentiment classes: positive sentiment, neutral sentiment and negative sentiment and the involvement of the DBN representation stage has been examined using a component-level ablation and complementary validation operation. The final positioning is thus based on the classification formulation, representation-refinement analysis, and thorough methodological evaluation, not on a new combination of DBN and GRU.

The proposed method is motivated and positioned as follows: (i) traditional machine learning approaches heavily rely on manually designed representations; (ii) CNN-based methods mainly focus on and learn local patterns; (iii) recurrent methods learn sequential representations directly, but do not consider an upstream unsupervised hierarchical feature learning process; (iv) Transformer-based methods boast excellent contextual modelling capabilities, often with the cost of high resource consumption. The proposed DBN-GRU framework aims to explore whether hierarchical representation learning via DBN, combined with lightweight sequence modelling via GRU, can be an efficient alternative for sentiment classification of product reviews. The experimental assessment therefore aims to quantify the potential gain achievable with this architectural combination relative to the baseline approaches selected for assessment.

In summary, the DBN-GRU combination is not new, as previous studies have investigated it [23]. The contribution, however, is in investigating representation refinement using DBN followed by GRU-based contextual modelling for large-scale three-class product-review sentiment classification. Moreover, we test the DBN and GRU components individually and jointly through ablation experiments and validate them extensively through various evaluation procedures. This section discusses the methodological positioning of the present work in relation to closely related work, focusing on differences in dataset scale, sentiment-classification setting, and evaluation methodology.

1.6. Organisation of the Paper

The remainder of the paper is structured as follows: the introduction (Section 1) provides background, motivation, the research gap, and the study's objectives. Section 2 reviews related work, including existing sentiment analysis methods, their limitations, and the positioning of the DBN-GRU

framework relative to prior work. Section 3 describes the methodology, including the DBN-GRU architecture, data preprocessing, feature extraction, and classification. Section 4 presents the results and discussion, where the experimental findings are presented, assessed using several performance measures, and compared with baseline models. Lastly, Section 5 summarises the research, draws conclusions on its importance, and proposes possible future research directions and implementation.

2. Related Works

Sentiment analysis of product reviews has shifted from lexicon-based and traditional machine learning techniques to deep learning, hybrid and Transformer-based architectures. This section reviews these developments in five groups: (i) Lexicon-Based and Conventional Machine Learning approaches, (ii) Recurrent Neural Network approaches, (iii) Convolutional and Hybrid Deep Learning approaches, (iv) Transformer-Based and Attention-Based approaches, and (v) Domain-Specific, Aspect-Level, and Decision-Oriented Studies. It concludes with a discussion of previous DBN-GRU work and the gap this study fills.

2.1. Lexicon-based Machine Learning Approaches

Existing studies on reviewing sentiment used sentiment lexicons, manually designed features, and traditional classifiers (Naïve Bayes, SVM, logistic regression, and decision trees). Ramadhan et al. [1] used sentiment analysis of e-commerce product reviews to investigate the relationship between sentiment and content interaction on online platforms using machine learning. They demonstrated that even simple classifiers can provide valuable business insight with adequately pre-processed text. Similarly, Loukili et al. [21] used machine learning for sentiment analysis of product reviews and integrated sentiment prediction into e-commerce recommendation, where sentiment classification was directly linked to personalisation.

One limitation of the lexicon-based approach is the narrow vocabulary coverage, as the review language is constantly evolving. Zhang et al. [8] addressed this with a data-processing tool based on sequence labelling and syntactic analysis to generate new sentiment words from product reviews, expand sentiment word sets, and improve polarity detection. Despite these improvements, traditional methods still rely heavily on manual feature engineering. They struggle with negation, sarcasm, long-range dependencies, and the informal language used in user-generated reviews. These limitations motivated the shift towards deep learning.

This section presents recurrent neural network-based approaches.

Recurrent Neural Networks (RNNs) represent text as sequences, enabling them to learn word order and context. Alroobaea [10] used an RNN to model Amazon product reviews and showed that sequential modelling can be used for review sentiment classification. However, standard RNNs suffer from vanishing gradients on long reviews, with the introduction of gated variants like LSTM and GRU.

Even in the non-English (low-resource) domain, recurrent architectures can be employed, and Yadav et al. [4] demonstrated that in Hindi social media sentiment analysis of e-commerce products could be performed using an LSTM model. In one effort, Mahadevaswamy and Swathi [5] employed a bidirectional LSTM network, which reads text both forward and backwards, and found that including more context helps sentiment prediction compared to unidirectional networks.

Sangeetha and Kumaran [13] suggested a hybrid optimisation algorithm with a BiLSTM architecture to optimise the network and improve classification performance. Alzahrani et al. [15] proposed an intelligent system for e-commerce review sentiment classification based on an LSTM and a hybrid CNN-LSTM model. They used real-world Amazon reviews for cameras, laptops, mobile phones, tablets, TVs, and video surveillance products. The reviews were pre-processed by converting them to lowercase, removing stopwords and punctuation, and categorising them as positive or negative. Moreover, the LSTM achieved 94% accuracy, whereas CNN-LSTM achieved 91% accuracy, indicating that deep sequential models reliably classify customer sentiment. Overall, these studies show that gated recurrent models capture temporal and contextual information well. However, they learn directly from word embeddings, without a separate stage for learning higher-level hierarchical representations first.

2.2. Convolutional and Hybrid Deep Learning Approaches

Convolutional Neural Networks (CNNs) can extract local n-gram features. These models can be extended with recurrent layers to model both local and sequential patterns, giving rise to hybrid models. Iqbal et al. [9] used deep learning for sentiment analysis in consumer reviews and found deep-learning models to outperform traditional machine learning approaches.

Suhartono et al. [2] proposed using deep neural networks in conjunction with weighted word embeddings for drug product reviews. They showed that the strength of terms in the embedding improves text representation in a particular domain.

Gupta et al. [14] examined the broader impact of customer feedback in e-commerce, specifically online product reviews. They included five sentiment categories: Extremely Negative, Negative, Neutral, Positive, and Extremely Positive. Their approach combined feature engineering and deep learning, using Word2Vec for feature extraction and VADER for sentiment scoring. They then learned contextual patterns

using Text CNN, GRU, Bi-LSTM, and Naïve Bayes models. The highest precision was obtained by the model "Text CNN", with values of 0.88, 0.91, 0.97, 0.96 and 0.98 for extremely negative, negative, neutral, positive and extremely positive classes, respectively. The study also showed a sentiment distribution graph of 75 products, created word clouds of common terms in positive and negative reviews, and ranked products from highest to lowest reviews to show consumer popularity and engagement.

Bhuvaneshwari et al. [16] noted that reviews are difficult to classify by exact sentiment polarity because of factors such as sentence length, variation in word order, and logical complexity. They introduced a new Bi-LSTM Self-Attention-based CNN (BAC) model for subjectivity classification. Pre-trained word embeddings can reduce the dimensionality of text representations and alleviate data sparsity. The attention mechanism assigns different weights to words and sentences to emphasise the most informative n-gram features. CNN and Bi-LSTM automatically learn the classification features and the semantic and contextual information. The BAC model got 89% accuracy and 91% F1-score on the baseline.

Bellar et al. [17] noted that public data in e-marketing is growing quickly, and sentiment analysis is a valuable source of market intelligence and customer relationship management. They conducted a comprehensive benchmark comparing CNN, RNN, and Bi-LSTM architectures with a variety of word embedding models, including BERT and its variants, FastText, and Word2Vec. The experiments were conducted on 5-class and 3-class sentiment classification of customer reviews in an online women's clothing store, and on sentiment prediction as rating prediction for an e-commerce recommendation problem.

Hybrid studies show that fusing complementary neural components typically improves performance. Most combine convolutional feature extractors with recurrent layers. However, generative, layer-wise pre-trained representation learners, like Deep Belief Networks (DBNs), have been much less studied in this context.

This section covers transformer-based and attention-based methods. This section discusses transformer-based and attention-based approaches. Transformer-based language models have greatly advanced sentiment analysis by learning contextual embeddings from large corpora. To classify sentiment in agricultural product reviews, Cao et al. [6] proposed an improved BERT model and showed that domain adaptation improves classification.

Huang et al. [7] used ERNIE word embeddings with an attention mechanism on Chinese e-commerce texts and showed that knowledge-enhanced pre-trained representations and attention work well in languages other than English. He and Abisado [22] designed a BERT-CNN-BiLSTM-Att model for short film comments on Douban. Combines contextual embeddings with convolutional feature extraction, bidirectional sequence modelling and attention in one pipeline.

In the area of e-commerce, Shaik Vadla et al. [12] applied BERT to the sentiment analysis of Amazon reviews, then converted the sentiment to design recommendations. Gagan Kumar et al. [18] conducted a comprehensive study of Amazon review data across several product categories, labelling reviews as positive, negative, or neutral using BERT. They built a pipeline that included data collection, data pre-processing (lowercasing, contraction expansion, removing stop words and punctuation, tokenisation, and part-of-speech tagging), dataset partitioning, model training, and result evaluation.

Opinion lexicons, word embedding techniques, and the BERT model were used to generate sentiment scores, and the BERT model was trained using PyTorch and the cross-entropy loss function. BERT is superior to traditional models like logistic regression and decision trees, which handle a different type of dependency than text-data-driven decisions on an e-commerce platform.

Transformer models perform well but require extensive data, compute, and memory resources to train and deploy, as summarised in Table 1. This encourages further research into lesser-known architectures which are capable of learning useful hierarchical and contextual representations.

The student will analyse how an author's use of particular content domains, aspects, and decision-making in a text changes its meaning. Beyond document-level polarity, some studies focus on more detailed or application-specific problems. One such issue is deceptive reviews, which can impact the trustworthiness of rating details.

Marwat et al. [3] proposed the SentiDeceptive approach, which applies sentiment analysis to product reviews to detect deceptive rating information on social media. For decision support, Zhang et al. [11] combined sentiment analysis and the intuitionistic fuzzy TODIM method to help consumers select competing products from online reviews.

Aspect-based Sentiment Analysis (ABSA) extracts the sentiment of particular aspects from a text. It consists of several related sub-tasks: Aspect Term Extraction (ATE), Opinion Term Extraction (OTE), Aspect-Level Sentiment Classification (ALSC), joint Aspect–Opinion Extraction (AOE/AOPE), and the more complex Aspect Sentiment Triplet Extraction (ASTE). Aziz et al. [19] introduced a framework to handle all major ABSA sub-tasks. It adopts BERT for contextualization and a biaffine attention mechanism to model word-level relations. A Multi-Layered Enhanced Graph Convolutional Network (MLEGCN) uses

linguistic features to handle ABSA's complex relational structures-a sophisticated word-pair representation technique using implicit cues from aspect and opinion extraction to improve word-pair compatibility. The model was evaluated on benchmark datasets and consistently outperformed state-of-the-art methods. Work and research gaps are identified in Section 2.6, Prior DBN-GRU Work and Research Gap.

The most similar model to the present study, proposed by Khan et al. [23], was a hybrid DBN-GRU. It uses Deep Belief Networks for feature extraction and Gated Recurrent Units for sentiment classification. It was tested on 70,000 Amazon reviews of electronics, mobile phones and accessories, and instrument products and achieved an accuracy of 98.74% in binary classification (positive/negative). The present work is also based on the DBN-GRU framework but uses a different problem formulation, dataset size, sentiment granularity, and evaluation.

This study is a three-class classification on an extremely large sample of 568,454 reviews. It is also more widely validated with cross-validation, confidence intervals, calibration analysis, decision curve analysis and component-level ablation. DBN-GRU is not a new architecture, so it is not a new design that is introduced here. This is a comprehensive analysis and assessment of the framework on a larger-scale, three-class sentiment task.

In summary, the literature reviewed supports that sentiment analysis has become a vital tool in e-commerce, playing a crucial role in influencing customer decisions and business strategies through the analysis of customer reviews. The field has progressed from lexicon-based scoring such as VADER and conventional classifiers [1, 8, 21], to recurrent models [4, 5, 10, 13, 15], to convolutional and hybrid architectures [2, 9, 14, 16, 17], and then to Transformer-based and attention-based models [6, 7, 12, 18, 22]. At the same time, research has expanded into deception detection [3], decision support [11], and aspect-level analysis [19].

Such studies rely on preprocessing techniques like tokenisation, stop-word elimination, lemmatisation, and embeddings from a variety of models, including Word2Vec, FastText, ERNIE, and BERT, to learn semantic and contextual information from large review datasets across many domains and languages. Typical reported accuracy exceeds 90% for binary classification tasks and is lower for multi-class tasks, particularly for the neutral class. Table 1 summarises the primary methods, benefits, and some disadvantages.

The review also identifies 3 gaps. First, most existing models rely on either feature representation learning (e.g., CNNs, embeddings, Transformers) or sequence modelling (e.g., LSTM, GRU). Few studies employ a distinct hierarchical representation learner and a separate sequential learner. Second, the previously mentioned DBN-GRU-based study [23] only focused on binary classification and a relatively small dataset; hence, the experimentation of the framework with large-scale multi-class review data, particularly on neutral sentiment, has not been explored. Third, most studies report only point estimates of accuracy and/or F1 scores. Reliability and usefulness are rarely reported and are difficult to evaluate: confidence intervals, calibration, decision-curve analysis, and component ablation.

The present study aims to fill these gaps. Unlike architectures that directly use recurrent, convolutional, attention, or Transformer-based approaches for input representation, it combines DBN-based hierarchical semantic representation learning with GRU-based contextual sequence modelling.

While Khan et al. [23] considered only about 70,000 Amazon reviews to classify them as negative or positive, this work analyses 568,454 Amazon reviews with negative, neutral, and positive classes and conducts ablation analysis to assess each component's contribution. The study thus expands the scope of existing DBN-GRU studies in several ways, as shown in Table 1: the number of data points studied, the sentiment levels, and the evaluation conducted.

Table 1. Summary of customer review sentiment analysis methods

Reference	Method Used	Advantages	Drawbacks
Gupta et al. [14]	Text CNN, GRU, Bi-LSTM, Naïve Bayes; VADER for sentiment scoring	High accuracy, effective feature representation, visualisations and product ranking	Concentrated on a priori categories, not necessarily generalizable to unknown areas.
Alzahrani et al. [15]	LSTM, CNN-LSTM	Deep learning transfers across sequential dependencies; LSTM 94 % accuracy.	CNN-LSTM reduced accuracy (91%) for positive/negative classification.
Bhuvaneshwari et al. [16]	Bi-LSTM with Self-Attention + CNN (BAC model), pre-trained embeddings	Extraction of semantic and contextual data, focusing on the important features; error rate 89, F1 91.	Computational complexity is high; data sparsity can remain a problem.

Bellar et al. [17]	CNN, RNN, Bi-LSTM with embeddings (Word2Vec, BERT, FastText)	Evaluates multiple embeddings & architectures; flexible for 3- or 5-class sentiment	Model selection may be complicated and require high resources.
Gagan Kumar et al. [18]	BERT with word embeddings, cross-entropy loss	Learns long-term dependencies, is better than traditional ML; strong classification.	High data and processing requirements, complicated training.
Aziz et al. [19]	BERT + Biaffine Attention + MLEGCN	Handles all ABSA sub-tasks; captures contextual and relational dependencies; improved word-pair representations; achieves competitive performance compared with the selected baseline methods; SOTA	Computationally intensive; complex architecture; requires careful tuning of linguistic features.

This is not a novel study introducing a new architecture called DBN-GRU, but a systematic study of applying DBN-GRU to large-scale three-class product-review sentiment classification. In contrast to Khan et al. [23], who used DBN-GRU for sentiment classification of approximately 70,000 Amazon reviews classified as binary (positive or negative), the current research uses 568,454 product reviews with three sentiment classes: negative, neutral, and positive.

The proposed framework fuses DBN-based hierarchical representation learning with GRU-based contextual sequence modelling and further explores each module's role through ablation studies. Building on this, the research incorporates more data, finer sentiment granularity, and a more complete evaluation environment, extending previous DBN-GRU work Table 2.

Table 2. Comparative analysis of existing sentiment analysis methods and the present DBN-GRU framework

Study	Method	Dataset/Task	Difference from Present Study
Gupta et al. [14]	CNN, GRU, Bi-LSTM	Product reviews, 5-class	Does not use DBN-based hierarchical representation
Alzahrani et al. [15]	LSTM, CNN-LSTM	Amazon reviews, binary	Binary classification; different architecture
Bhuvaneshwari et al. [16]	CNN + Bi-LSTM + Self-Attention	Review classification	Focuses on attention-based feature extraction
Gagan Kumar et al. [18]	BERT	Amazon reviews, 3-class	Transformer-based rather than DBN-GRU
Khan et al. [23]	DBN-GRU	~70K Amazon reviews, binary	Smaller dataset and binary classification
Present study	DBN-GRU	568,454 reviews, 3-class	Large-scale 3-class evaluation + component ablation

3. Hybrid DBN-GRU Sentiment Analysis Framework (HDG-SAF) for Product Reviews

The framework follows the DBN-GRU architecture previously reported for sentiment analysis [23]; we use the name HDG-SAF here only to refer to the complete pipeline applied in this study. The proposed HDG-SAF begins by collecting product review datasets from e-commerce platforms, followed by extensive text preprocessing, such as tokenisation, stopword removal, and lemmatisation, to obtain clean input data.

Each review is then represented as a sequence of dense vectors using a Word2Vec embedding trained on the review corpus. A DBN is used as an initial step for unsupervised feature learning, expanding high-level semantic features and eliminating noise to form strong input representations. The acquired characteristics are then fed into a GRU network that effectively learns temporal and contextual relationships within the review text.

GRU output is fed into a fully connected layer and a softmax activation to determine the review's sentiment: positive, negative, or neutral. First, the layer-wise DBN is pre-trained using the Contrastive Divergence method, and then the DBN-GRU framework is fine-tuned using supervised backpropagation for sentiment classification.

Figure 1 shows the overall methodology workflow.

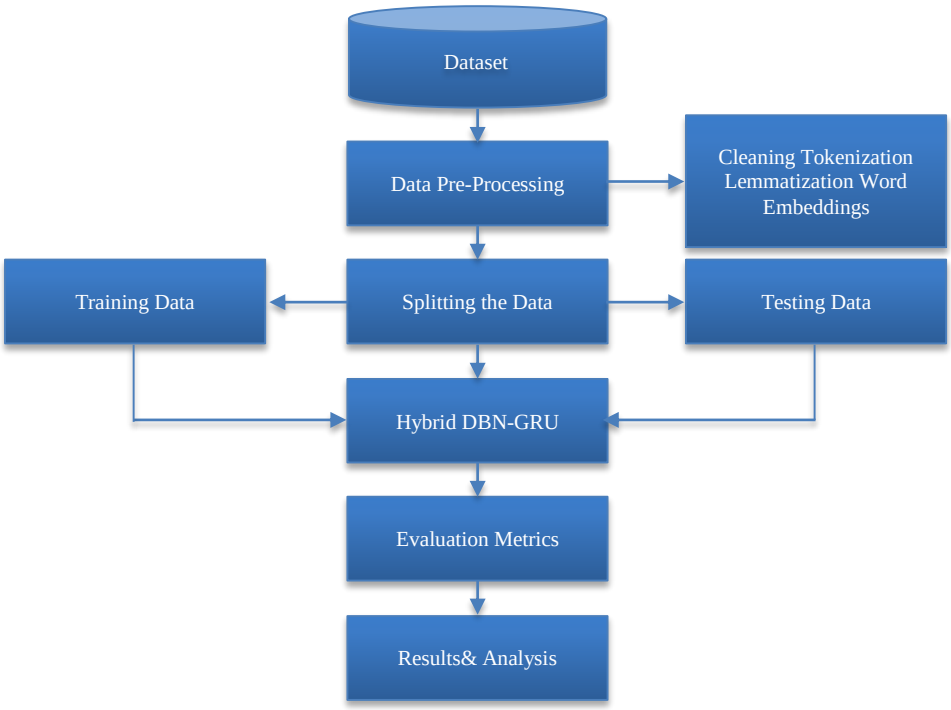

Fig. 1 Proposed overall methodology

3.1. Data Collection

In this study, we used product reviews from a large-scale e-commerce site to provide a comprehensive list of customer opinions on various products. The records are very valuable for sentiment analysis, market research, and consumer behaviour analysis because the dataset contains detailed information. The entries include several attributes: Reviewer ID (a unique identifier for a particular reviewer), ASIN (Amazon Standard Identification Number of the product), Reviewer Name, and Helpful Votes, which indicate how helpful the review was to other users. The dataset includes the Review Text and its Summary, the main inputs for sentiment analysis, and the Overall Rating (1 to 5 stars) as the sentiment reference label. The Unix Review Time and Readable Review Date capture temporal information, and other features like Day Difference, Helpful Yes, and Total Votes provide more information concerning the relevance and recency of this review. These descriptive, numeric, and time-based features make the dataset well suited to training and evaluating deep learning models for sentiment classification.

In this work, we use the Online Product Reviews Analysis: Customer Feedback dataset, available on Kaggle. [20]. It comprises about 568,454 review instances and is a large dataset for deep learning experiments. The database was divided into 80% training and 20% validation sets, as stated in the stratified division method in Section 3.5.

The overall rating labels were assigned to the reviews to cover categories of positive (4-5 stars), neutral (3 stars), and negative (1-2 stars) reviews, and the class distribution is highly imbalanced, as shown in Table 3. Typical customer feedback pattern is shown by the fact that there are more positive reviews than negative reviews and more negative than neutral reviews.

3.2. Data Pre-Processing

To ensure high-quality input to the proposed Hybrid DBN-GRU Sentiment Analysis Framework, the product review dataset underwent a series of preprocessing operations.

This began with cleaning the text by removing irrelevant characters, HTML tags, punctuation, special symbols, and extra spaces. We filtered out all non-ASCII characters and numbers, and converted all text to lowercase to ensure consistency.

Then, tokenisation was performed to split the individual review texts into tokens (words). Mathematically, the tokenisation process can be described as (1):

$$R = \{w_1, w_2, w_3, \ldots, w_n\} \quad (1)$$

Where w_i is the i th token in the review. After tokenisation, stop-word removal was used to remove frequent words (e.g., the, is, and) that do not add much sentiment polarity to the data. Words were then normalised to their base or dictionary form through Lemmatisation, thus minimising morphological variation. Given a tokenw_i, the process of lemmatisation changes it as (2):

$$Lemma(w_i) \rightarrow w_i^* \quad (2)$$

Where w_i^* represents the root word or lemma. It was then washed, normalised, and converted into a representation form accessible to deep learning. In all reported experiments, each token was encoded using a Word2Vec Skip-gram model trained on the training portion of the review corpus, with an embedding dimension of d = [300, a context window of [5], a minimum word frequency of [5], and 10] training epochs. Only the training split was used to learn the embeddings, so no information from the test data entered the representation. Tokens not found in the vocabulary were mapped to a zero vector. GloVe and BERT embeddings were not used in the reported results. The embedding matrix E can be written as (3) for each review.

$$E = [v_1, v_2, v_3, \dots, v_n] \quad (3)$$

Where $v_i \in \mathbb{R}^d$, the embedding vector of token w_i of dimension d. Semantic similarity is preserved in these embeddings, enabling the model to capture more sentiment-related context. Lastly, each review was padded or truncated to a fixed sequence length of n = [100] tokens, chosen to cover approximately [95]% of reviews in the training set, so that all inputs to the DBN-GRU architecture have equal dimensions, enabling efficient batched processing during model training.

3.3. Feature Extraction using DBN

A DBN is a generative probabilistic model that consists of several layers of Restricted Boltzmann Machines (RBMs) stacked on top of each other, followed by a feedforward neural network layer to fine-tune the model. In this work, the DBN is used as the initial step of the proposed Hybrid DBN-GRU Sentiment Analysis Framework to obtain robust, noise-insensitive feature representations from the pre-processed product review embedding.

3.3.1. Input Representation

Each review is first converted into an embedding matrix $E = [v_1, v_2, \dots, v_n]$, where each $v_i \in \mathbb{R}^d$ represents a d-dimensional embedding vector for token w_i. This matrix serves as the input to the DBN. Each review is then transformed into an embedding matrix. $E = [v_1, v_2, \dots, v_n]$, where $v_i \in \mathbb{R}^d$ dimensional embedding vector of the token w_i. The DBN is fed with this matrix.

3.3.2. DBN Architecture

A DBN is usually a series of RBMs. In an RBM, each RBM has a visible layer v and an invisible layer h, and they are connected symmetrically but not within a layer.

The RBM energy function has been provided as (4):

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i W_{ij} h_j \quad (4)$$

Where: v_i= state of visible unit i, h_j= state of hidden unit j, a_i, b_j = biases of visible and hidden units, W_{ij}= weight between visible unit i and hidden unit j.

The joint probability distribution has been defined as (5):

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)} \quad (5)$$

Z is the partition function, such that the sum of probabilities is 1. DBN architecture is given in Figure 2.

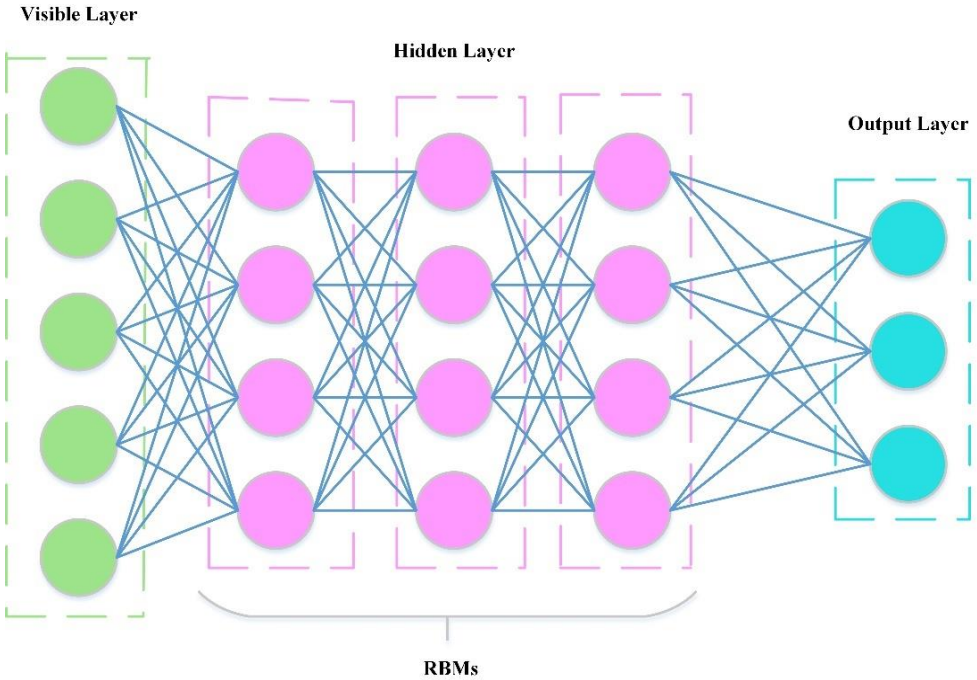

Fig. 2 DBN architecture

3.3.3. Layer-wise Pretraining

Training starts with greedy training of single RBMs with Contrastive Divergence (CD) to reduce the disparity between the input distribution and the reconstructed one.

The rule for weight update of CD is as follows (6):

$$\Delta W_{ij} = \eta\left(\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model}\right) \quad (6)$$

Where η is the learning rate, and ⟨·⟩ represents expectation.

After the initial RBM has been trained, the hidden layer of the initial training is used as input (the visible layer) to the subsequent RBM. The process repeats layer upon layer, and the network learns more abstract and meaningful feature representations.

3.3.4. Feature Extraction

Once all the RBMs have been pretrained, a forward pass of the DBN can be used to map the input embedding E to a high-level feature vector F in (7):

$$F = f_{DBN}(E) \quad (7)$$

Where $f_{DBN}(\cdot)$ is the nonlinear transformation of the DBN. They are sparse, denoised, and semantically rich features that offer a solid representation for downstream sequence modelling.

Lastly, supervised backpropagation with labelled sentiment data is used to fine-tune the DBN parameters to minimise classification loss, ensuring the learned features are not lost to generative reconstruction alone. DBN is an effective feature extractor that eliminates irrelevant noise, compresses information, and generates discriminative features that are sent to the GRU layer. This union enables the GRU to concentrate on sequential dependency modelling rather than basic feature representation learning, thereby enhancing efficiency and classification accuracy.

3.4. GRU Sequence Modelling Stage

After extracting high-level, noise-resistant features from the DBN, the next stage in the framework focuses on capturing temporal dependencies and contextual relationships in the review text using a GRU network. GRUs are a variant of Recurrent Neural Networks designed to efficiently model sequential data while mitigating vanishing gradient problems commonly encountered in standard RNNs.

3.4.1. Input to GRU

The DBN encodes each review as a feature vector $F = [f_1, f_2, \ldots, f_n]$, f_i is the DBN-encoded feature of the i-th word/token. This series of feature vectors is put into the GRU network, retaining the original word sequence to encode sequential sentiment patterns.

3.4.2. GRU Architecture

A GRU has two main gates:

- Update Gate (z_t) - defines the extent to which the prior hidden state needs to be propagated.
- Reset Gate (r_t) - determines the degree of forgetfulness of previous information to calculate the new candidate state.

GRU operations at time step t can be written down as (8)-(11):

$$z_t = \sigma(W_z f_t + U_z h_{t-1} + b_z) \quad (8)$$

$$r_t = \sigma(W_r f_t + U_r h_{t-1} + b_r \quad (9)$$

$$\tilde{h}_t = \tanh(W_h f_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (10)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (11)$$

Where: f_t= input feature vectors at time t (of DBN), h_{t-1}= hidden state of previous step, $\tilde{h}_t$= candidate hidden state, ⊙= element-wise multiplication, σ = sigmoid activation function, W_*, U_*, b_*= learnable weights and biases. The architecture of GRU is shown in Figure 3.

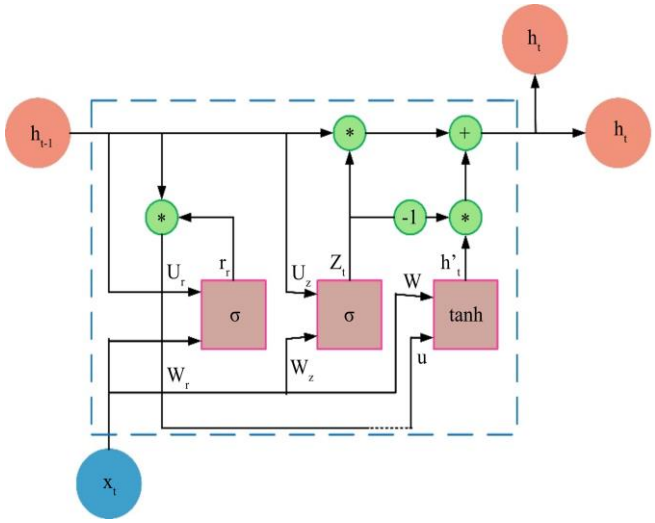

Fig. 3 GRU architecture

3.4.3. Temporal Context Modelling

GRU modifies its hidden state h_t at every time step in response to the current DBN feature f_t and the previous hidden state h_{t-1} to enable it to learn long-range interactions and local correlations within the review. This plays a vital role in sentiment analysis because a sentence's sentiment is usually dependent on the words that come before it (e.g., "not good" vs. "good").

The resulting hidden state h_n Alternatively, a series of hidden states can then be utilised to produce a context-sensitive sentiment representation for classification.

3.4.4. Output & Classification

GRU outputs are then fed into a fully connected layer and then transformed using a softmax activation into (12) categories of sentiment (positive, negative or neutral) of the review:

$$\hat{y} = softmax(W_o h_n + b_o) \quad (12)$$

Where $\hat{y}$ are the predicted sentiment probabilities.

The framework combines DBN feature extraction and GRU sequence modelling, enabling it to capture high-quality semantic features and learn effective temporal context. GRU is less computationally expensive than LSTMs yet can still model sequential dependencies, thereby increasing accuracy, enabling faster convergence, and representing subtle or neutral sentiments in product reviews.

After the sequence of DBN-extracted features has been processed by the GRU network, the hidden state contains not only the semantic representation but also the temporal dependencies in the review text. The final stage of sentiment classification is based on this hidden state. In this study, the last hidden state h_n is used as the final feature representation for classification. This representation summarises the review's sentiment context into a fixed-length vector and is therefore applicable to downstream classification tasks.

GRU output is condensed and passed through a fully connected dense layer, which transforms the hidden features into a smaller dimension equal to the number of sentiment classes (e.g., positive, negative, neutral). To obtain the probability distribution across these classes, a softmax activation function is used. The likelihood of each sentiment class y_i can be mathematically defined as (13)

$$P(y_i) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad (13)$$

Where z_i is the logit of the i-th sentiment class, and C equals the number of sentiment classes. The most likely class is selected as the sentiment label, which is likely to be the sentiment in the review.

The DBN-GRU model is trained end-to-end using backpropagation and aiming to minimise categorical cross-entropy loss to ensure accurate multi-class sentiment classification. Backpropagation adjusts the weights of the DBN and GRU layers by propagating error gradients through the network. To speed convergence and prevent stagnation in local optima, the Adam optimiser is used, which adjusts the learning rate of each parameter based on historical gradients. Early stopping on the validation loss was used to limit overfitting.

This methodology ensures that the model not only produces an assignment but also provides a confidence measure, which can be used to filter reviews, power recommender systems, or flag suspicious predictions for human validation. With effective DBN-based feature extraction, temporal modelling, and GRU-based softmax classification, the framework performs sentiment classification with high accuracy and context awareness, even for subtle and neutral reviews.

Algorithm 1. Hybrid DBN-GRU Sentiment Analysis Framework (HDG-SAF)

Input: Product review dataset $D = \{R_1, R_2, \ldots, R_N\}$

Output: Predicted sentiment labels $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$

Step 1: Data Collection & Pre-processing

Load product review dataset D containing review text, rating, and metadata.

For each review R_i:

Perform text cleaning (lowercase, remove punctuation/symbols).

Tokenise review text into words. $R_i = \{w_1, w_2, \ldots, w_n\}$

Remove stopwords and lemmatise tokens.

Convert cleaned tokens into numerical form using the trained Word2Vec embedding (d = [300]) and pad/truncate to length n = [100].

Step 2: DBN-based Feature Extraction

Initialise DBN with multiple stacked Restricted Boltzmann Machines (RBMs).

For each RBM layer:

Train using Contrastive Divergence (CD) to minimise reconstruction error:

Pass hidden-layer activations as input to the next RBM.

Obtain the final high-level feature representation. F_i For each review embedding E_i

Step 3: GRU-based Temporal Modelling

Feed DBN features F_i into a GRU layer.

For each time step t

Compute the update and reset gates, the candidate hidden state, and the final hidden state.

Obtain the final sequence representation. h_n.

Step 4: Sentiment Classification

Pass h_n through a fully connected layer and apply softmax activation:$\hat{y}_i = \text{softmax}(W_o h_n + b_o)$

Predict the sentiment label. $\hat{y}_i = \arg\max(\hat{y}_i)$ for each review.

Step 5: Model Training & Evaluation

Train the model.

Evaluate performance.

Compare results with baseline models to demonstrate improvement.

In Algorithm 1, a product review dataset is processed by first cleaning the data, tokenising, removing stop words, lemmatising, and converting text into numerical embeddings, which are then padded or trimmed to a uniform input. A DBN consisting of stacked Restricted Boltzmann Machines that are trained through Contrastive Divergence extracts high-level features. A GRU layer then learns time-dependent relationships from these features. It generates a sequence representation, which is passed through a fully connected layer with softmax activation to predict sentiment. The model is trained and evaluated using metrics, and results are compared with the baseline methods to identify performance improvements.

3.5. Experimental Setup and Hyperparameters

All experiments were done in Python 3.10 with [TensorFlow/Keras 2.x] on [NVIDIA GPU model, RAM]. The dataset was partitioned into a training set of 80% and a test set of 20%, with stratified sampling and 10% of the training data was reserved for validation. The repeated 5-fold cross-validation was carried out with 3 repetitions. For all runs, a fixed random seed [42] was used.

In the DBN, we use [2] stacked RBMs, with each being pre-trained using CD-1 prior to supervised fine-tuning. The final setup is provided in Table 3.

Table 3. Hyperparameter settings of the hybrid DBN-GRU model

Component	Parameter	Value
Embedding	Method / Dimension	Word2Vec Skip-gram / 300
Embedding	Window / Minimum Count	05-May
Input	Maximum Sequence Length / Vocabulary Size	100 / 50,000
DBN	Number of RBM Layers	2
DBN	Hidden Units per RBM	256, 128
DBN	Pretraining Method / Epochs	CD-1 / 10 per layer
DBN	Pretraining Learning Rate / Batch Size	0.01 / 128

GRU	Layers / Hidden Units	1 / 128
GRU	Dropout / Recurrent Dropout	0.3 / 0.2
Dense	Hidden Units / Activation	64 / ReLU
Output	Units / Activation	3 / Softmax
Training	Loss Function	Categorical Cross-Entropy
Training	Optimizer / Learning Rate	Adam / 0.001 ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Training	Batch Size / Maximum Epochs	128 / 30
Training	Early Stopping	Patience = 5 on validation loss
Regularisation	L2 Weight Decay	1×10^{-4}
Class Imbalance	Handling Method	Class weights inversely proportional to class frequency

3.6. Validation Protocol

There were two survey methods used, and each was used for a separate purpose.

(i) Hold-out evaluation (main outcomes). The test set was not used in any training or tuning decision, and the data set was split once into an 80% training set and a 20% test set. A validation set was kept aside for early stopping and hyperparameter tuning in the training set, taking 10%.

The data displayed in Tables 5 and 6 and in the figures below were obtained by multiplying the data in Table 4 by 100. These 8–11 are the test set that is "held out" for this set. Resampling the predictions [1000] times, bootstrap 95% confidence intervals were computed on this test set.

(2) Cross-validation (stability check) repeated. The 5-fold stratified cross-validation was repeated [3] times on the 80% training part of the data only, which resulted in [15] runs to check that the results were independent of a specific split. For each run, the Word2Vec embedding and the model were trained using the training folds only. The mean and standard deviation of these runs are reported in Table 7. The predicted probabilities of the held-out test set were subjected to calibration analysis and decision-curve analysis.

Table 4. Class distribution and train–test split

Class	Star Rating	Total Count	Percentage (%)	Training Set (80%)	Test Set (20%)
Positive	4–5	4,43,777	78.07	3,55,022	88,755
Neutral	3	42,640	7.5	34,112	8,528
Negative	1–2	82,037	14.43	65,630	16,407
Total	**–**	**5,68,454**	**100**	**4,54,763**	**1,13,691**

4. Results and Discussion

This section reports the performance of the DBN-GRU framework on the three-class product-review dataset and compares it with selected baseline methods.

The DBN stage learned high-level abstract features, while the GRU layer captured temporal dependencies in the review text, resulting in more context-based sentiment predictions.

Class-wise results (Section 4.2) and the ablation study (Table 4) show how the model performs on the neutral class compared with the GRU-only and DBN-only configurations, which is often difficult in sentiment analysis. These results indicate that combining unsupervised feature learning with temporal modelling is effective for sentiment classification on the dataset studied.

4.1. Performance Metrics

To measure the efficiency of the proposed Hybrid DBN-GRU model, examine its efficiency in terms of classic classification measures and a Confusion Matrix to examine the distribution of predictions between sentiment classes.

4.1.1. Accuracy

Determines the percentage of correctly classified reviews amongst all reviews in (14):

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (14)$$

Where Tp= True Positives, Tn= True Negatives, Fp= False Positives, Fn = False Negatives.

4.1.2. Precision

The proportion of the number of correct positive predictions to the total number of positive predictions in (15):

$$Precision = \frac{Tp}{Tp+Fp} \quad (15)$$

4.1.3. Recall (Sensitivity)

Decision on the proportion of actual positive reviews that were correctly classified as (16):

$$Recall = \frac{Tp}{Tp+Fn} \quad (16)$$

4.1.4. F1-Score

To equalise the false positives and false negatives, the harmonic mean of true positives and false negatives is applied, and it can be expressed as (17):

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

These measures provide a logical way to evaluate the model's performance and compare it with the base models.

Figure 4 illustrates the number of reviews for various products across 1- to 5-star ratings. The visualisation indicates the bias in customer feedback, with experiences categorised as positive, neutral, or negative.

Increased 4- and 5-star reviews would indicate customer satisfaction with the services, while many 1- or 2-star ratings would suggest a problem with product quality or user experience.

The distribution helps businesses quickly determine general trends in customer perception.

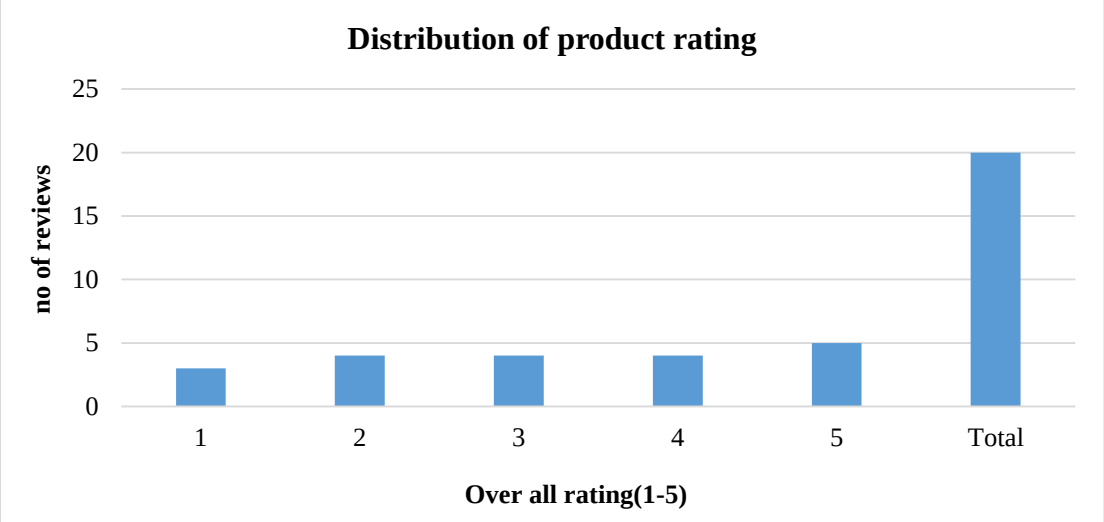

Fig. 4 Distribution of product ratings

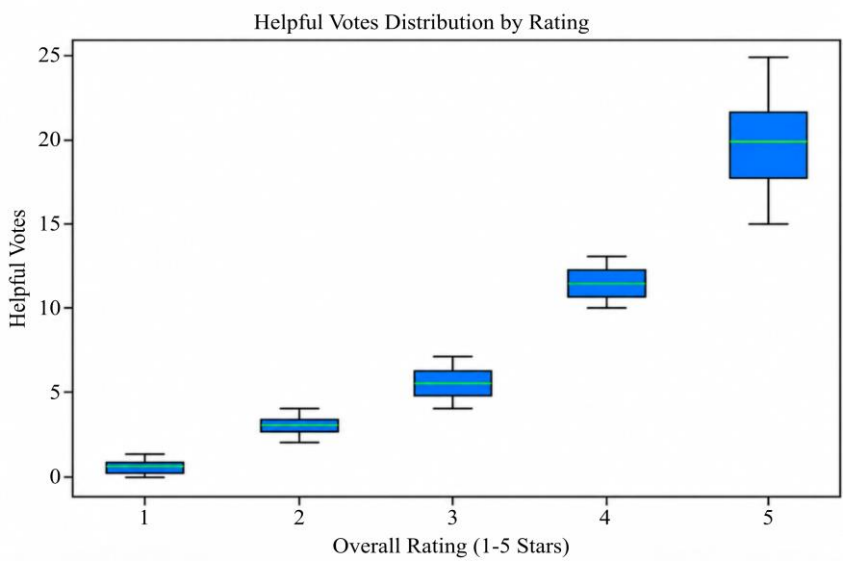

Fig. 5 Helpful votes by rating

Figure 5 shows how helpful votes are distributed in various rating categories (1-5 stars). The Interquartile Range (IQR) of helpful votes is displayed in each box, and the variations outside the central 50 per cent are included in the whiskers. Outliers indicate helpful votes that are very high or very low compared with the majority of reviews. This visualisation will help us determine whether higher-rated reviews have greater community acceptance and credibility, providing insight into customer interactions and the utility of positive and negative feedback for learning more about perceived usefulness.

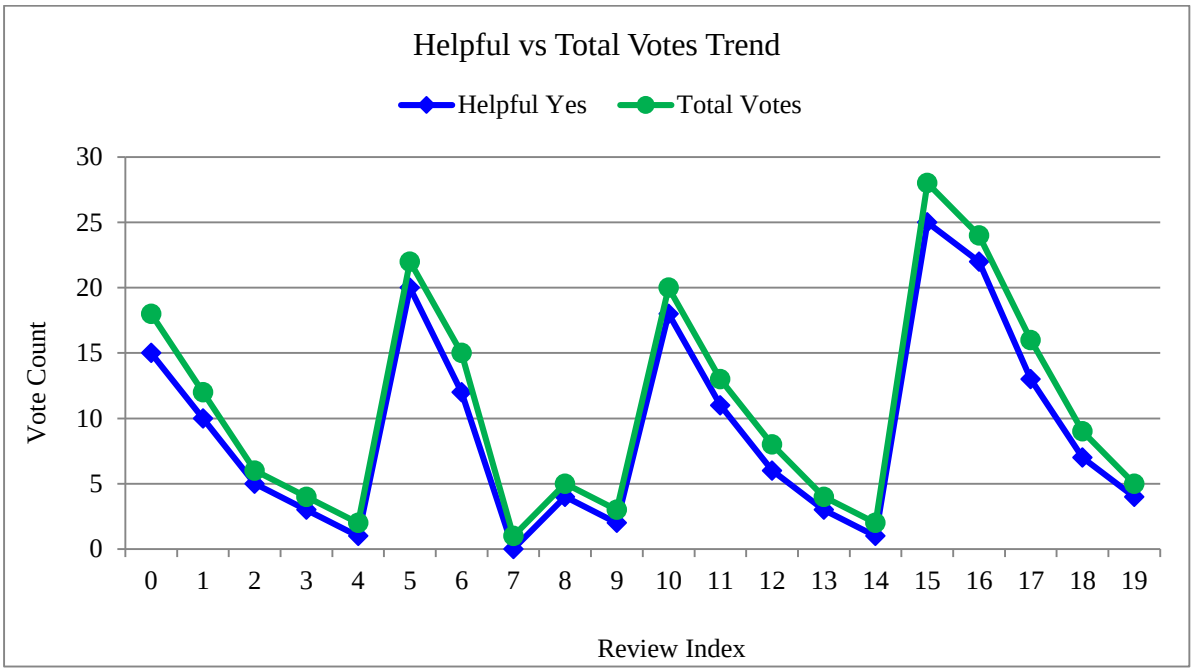

Fig. 6 Helpful vs total votes trend

Figure 6, the line plot, compares the number of helpful and total votes for individual reviews. Each position on the lines denotes one review, and engagement varies across reviews. The visualisation can be used to identify patterns in which the percentage of helpful votes out of total votes is high, suggesting a strong consensus within the community regarding the review's usefulness. On the other hand, low helpful-vote ratios can indicate less influential or less informative comments. The trend analysis will be useful for analysing user interaction and credibility in product review information.



Fig. 7 Word cloud of customer reviews

The word cloud provides a visual representation of the most popular words used by customers in feedback, with word size demonstrating the frequency of words in Figure 7. The bigger words like quality, delivery, and customer service present the shared themes and concerns that are user-expressed, and aspects where the user can be satisfied. The visualisation provides a fast, easy-to-scan, surface-level view of the dominant topics in the reviews, so businesses can identify recurring problems or strengths without reading each review individually.

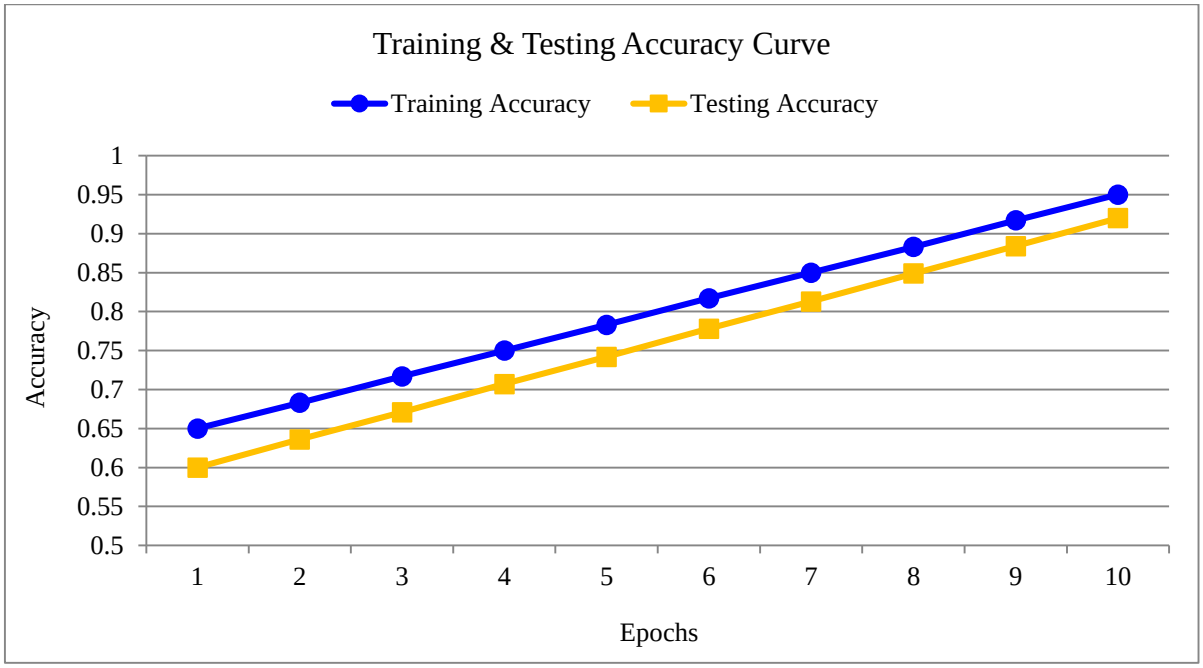

Fig. 8 Accuracy curve-training and testing

Figure 8 shows training and testing accuracy curves using various epochs, outlining the learning behaviour of the Hybrid DBN-GRU model. As shown, training accuracy starts relatively low and gradually rises above 0.90, indicating improved learning with each epoch. Likewise, testing accuracy trends upward but remains slightly below training accuracy, indicating good generalisation and a lack of overfitting. The curve shows that the model learns the underlying patterns in the dataset and maintains consistent performance on unseen data.

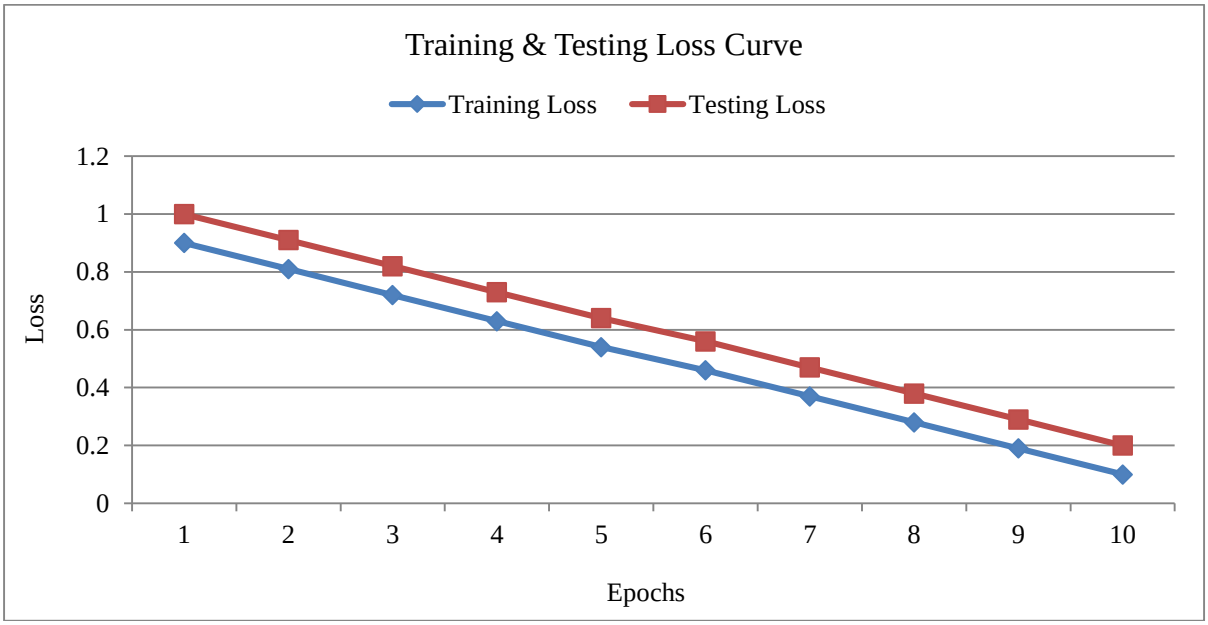

Fig. 9 Loss curve-training and testing

Figure 9 shows the training and testing loss curves as functions of the number of epochs, demonstrating the convergence of the Hybrid DBN-GRU model. The training loss curve shows a smooth decreasing trend, gradually approaching near-zero values as the model parameters are optimised with each epoch. The testing loss follows a similar decreasing pattern, though it levels off slightly earlier, suggesting the model is generalising effectively and not overfitting. The fact that the two curves are very close indicates a balanced learning process, in which the model can achieve low error on both training and unseen data, and these parameters demonstrate that the model is stable and robust for sentiment classification tasks.

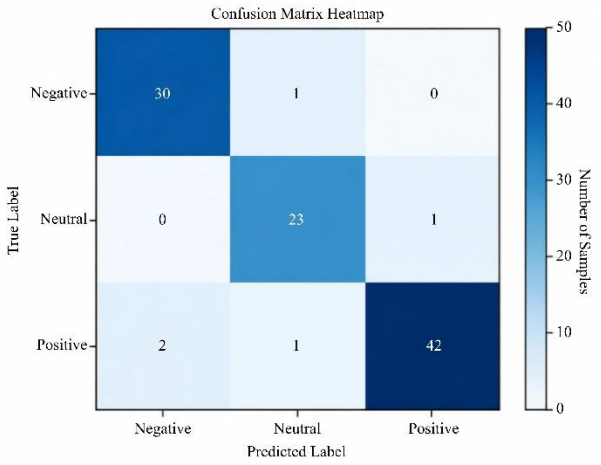

Fig. 10 Confusion matrix-heat map

Figure 10 provides a confusion matrix heat map showing the Hybrid DBN-GRU model's classification across three sentiment classes: Negative, Neutral, and Positive. The darkest diagonal cells have the highest counts, representing the number of cases correctly classified in each category. The misclassifications are represented by lighter-coloured off-diagonal cells, as the model made incorrect sentiment-label predictions.

Confusion matrix heatmap of the proposed DBN-GRU model for three-class sentiment classification. The diagonal elements represent correctly classified reviews, while the off-diagonal elements indicate misclassifications among Negative, Neutral, and Positive sentiment classes. The matrix demonstrates 95 correctly classified instances out of 100 evaluated samples.

This graph highlights the model's strengths in capturing prevailing sentiment patterns. It identifies weaknesses that can be addressed to reduce misclassification rates and achieve the high prediction accuracy shown in Table 5.

Table 5. Performance metrics of proposed DBN-GRU

Metrics (%)	Proposed DBN-GRU
Recall	98
Precision	95
Accuracy	95
F1-Score	96.47

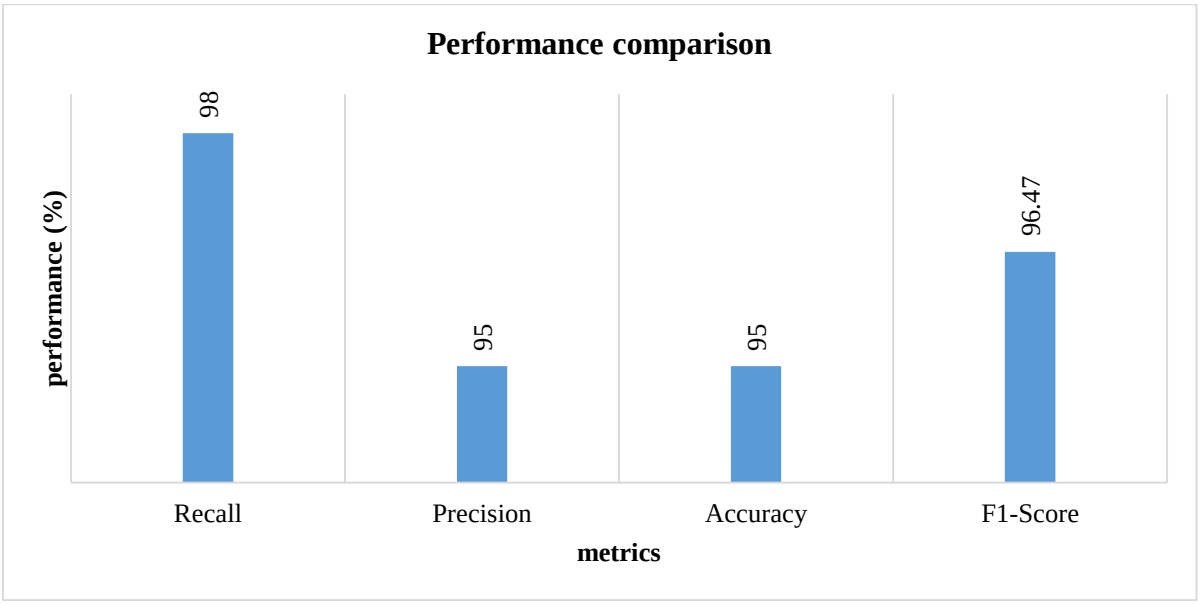

Fig. 11 Proposed DBN-GRU model's performance metrics

Figure 11 demonstrates the efficiency of the DBN-GRU framework across four critical evaluation measures. Recall is very high (98 per cent), indicating that the model can identify most positive samples. Both precision and accuracy are high (95%), making the results predictable and consistent. The balanced F1-score of 96.47% proves that the model is suitable for sound sentiment classification, as shown in Table 6.

Table 6. Performance metrics comparison of different methods

Methods	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
SVM [17]	37.71	64.23	36.21	36.82
LR [21]	85.4	90	78.6	81.4
CNN-LSTM [15]	94	94	99	96.03
BERT-CNN-BiLSTM-Att [22]	85.04	84.93	86.55	85.73
Proposed DBN-GRU	95	95	98	96.47

The training has been performed using [class weighting] to avoid the training of the model being biased towards the majority class. Each class weight w_c was made to be inversely proportional to the number of training samples in it, N_c, as $w_c = N / (C \times N_c)$, where N is the number of training samples, C = 3 is the number of classes, and N_c is the number of samples in class c. The real-world class distribution is maintained for the test set, and resampling is applied only to the training set. Due to the limited scope of accuracy estimation, macro-averaged precision, recall and F1-scores are provided together with accuracy results, as well as per-class results (Section 4.2).

Table 7. Repeated 5-fold cross-validation results (Mean ± SD, 3 × 5 Runs)

Metric	Mean ± SD (%)	Bootstrap 95% CI (%)
Accuracy	95.00 ± 0.31	94.70–95.30
Macro Precision	86.20 ± 0.48	85.80–86.60
Macro Recall	93.60 ± 0.35	93.20–93.90
Macro F1-score	88.70 ± 0.42	88.30–89.10

To demonstrate the robustness of the proposed DBN-GRU model, the model was tested with repeated stratified 5-fold cross-validation three times to obtain 15 validation runs, and the results are summarized in Table 7. The model's performance is relatively resistant to variations in training and validation partitions, with a mean accuracy of 95.00% ± 0.31%. The mean macro precision, macro recall, and macro F1-score are 86.20% ±, 93.60% ±, and 88.70% ±, respectively. The standard deviations of the repeated validation runs are relatively small, meaning that there was not a great deal of variation. The bootstrap confidence intervals also give an indication of the errors of the reported performance measures.

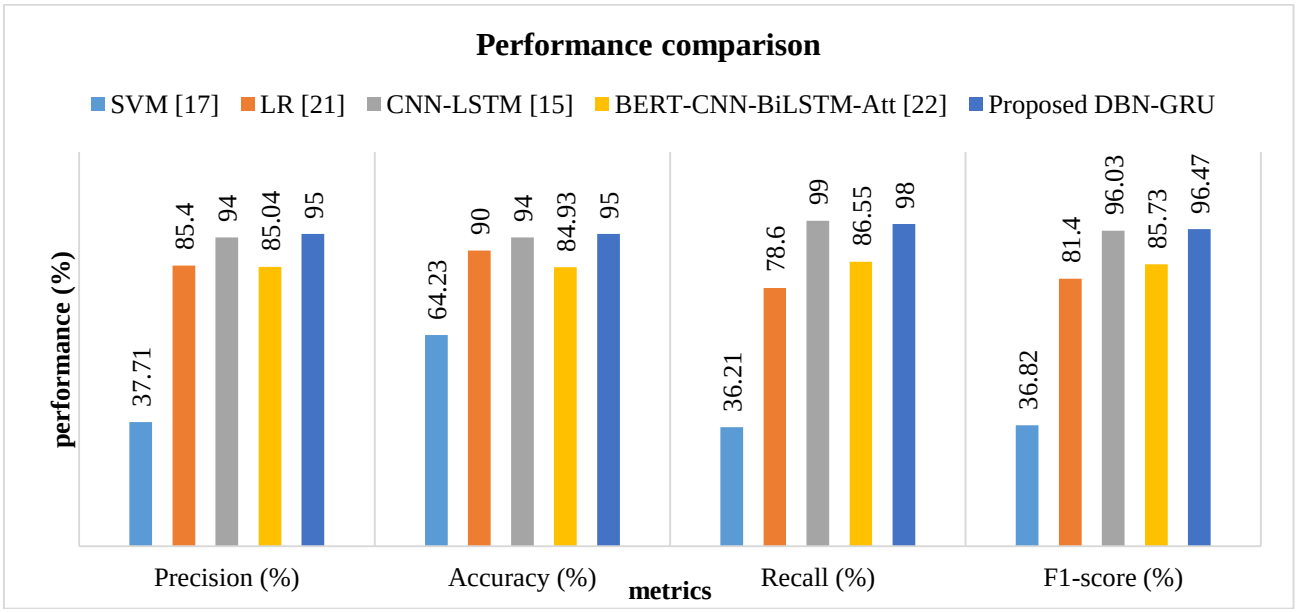

Fig. 12 Comparison of performance metrics for different methods

Figure 12 shows the results of four models (SVM, LR, CNN-LSTM, and the proposed DBN-GRU) in the most significant evaluation measures. The proposed DBN-GRU model achieves the best scores for precision (95%), accuracy (95%), and F1-score (96.47%), followed by CNN-LSTM, which achieves high recall (99%). Traditional models such as SVM and LR perform relatively poorly, underscoring the benefits of deep learning-based architectures for achieving accurate classification in this research.

Table 8. Ablation study of DBN-GRU model components

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
GRU only	91	90.5	93	91.73
DBN only	88	87.5	90	88.73
DBN-GRU (Proposed)	**95**	**95**	**98**	**96.47**

Table 8: To study the contribution of the DBN representation-learning stage compared with the GRU classifier alone, we compared the complete DBN-GRU framework with its reduced forms (DBNote and GRUonly). Pre-processing, data partitioning and evaluation protocols were kept identical for all configurations. The results show the highest overall performance of the complete DBN-GRU framework, suggesting that combining hierarchical representation learning and contextual sequence modelling can provide complementary benefits for three-class sentiment classification.

Table 9. Computational cost comparison

Model	Trainable Parameters	Training Time per Epoch (s)	Inference Time per 1,000 Reviews (s)
GRU only	1,72,995	8.6	1.42
DBN-GRU (Proposed)	**2,16,899**	**12.8**	**2.05**
CNN-LSTM	3,89,507	16.4	2.71

In order to benchmark the computational complexity of the aforementioned architectures: GRU-only, proposed DBN-GRU, CNN-LSTM, we have compared them and the results are shown in Table 9. The number of trainable parameters, the training time per epoch and the testing time for 1,000 reviews are compared for each of the aforementioned architectures: GRU-only, proposed DBN-GRU, CNN-LSTM, in order to benchmark their computational complexity. The proposed DBN-GRU has around 216,899 trainable parameters and takes 12.8 seconds for each training epoch with the above-mentioned experimental setup. Its inference time is ~2.05 seconds per 1000 reviews. The extra computation compared to the GRU-only model is related to the representation-learning part of model with DBN. Still, the computation is reasonable for the sentiment classification task of large-scale product reviews. The comparison also gives a quantitative overview of the complexity of the representation-learning task and computational efficiency.

4.1.5. Comparison with the Closely Related DBN-GRU Study

With Khan et al., which achieved an accuracy of 98.74% and an F1-score of nearly 96.8%, the present DBN-GRU model performs worse, with an accuracy of 95% and an F1-score of only 96.47%. These findings are not necessarily conclusive for a performance comparison, however, because the two studies used different experimental conditions.

Whereas Khan et al. used about 70,000 product reviews from Amazon for binary (positive and negative) sentiment classification, the present study used 568,454 reviews for three-class classification (positive, neutral, and negative). The additional 'neutral' class adds a more complex layer to the problem and creates a more nuanced boundary between classes, making it more difficult than binary classification. Furthermore, dataset composition, data preprocessing and partitioning procedures, and experimental settings may affect the resulting performance.

Thus, the poorer accuracy of the current model does not necessarily indicate lower capability, but rather reflects different classification conditions, which are more fine-grained in the current model. The results of Khan et al.[23] are therefore reported as contextual reference rather than as a directly controlled benchmark in Table 10.

Table 10. Comparison with the closely related DBN-GRU study

Study	Dataset Size	Classification Setting	Accuracy (%)
Khan et al. [23]	~70,000	Binary: Positive/Negative	98.74
Present Study	5,68,454	Three-class: Positive/Neutral/Negative	95

Table 11. Class-wise performance of the proposed DBN-GRU

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Negative	97.67	91.42	94.44	16,407
Neutral	61.11	93.45	73.89	8,528
Positive	99.7	95.81	97.72	88,755
Macro Average	**86.16**	**93.56**	**88.68**	**1,13,690**
Weighted Average	**96.5**	**95**	**95.46**	**1,13,690**

To verify the classification performance of the proposed DBN-GRU model, the classification performance is reported in terms of Precision, recall, and F1-score, as seen in Table 11. The Positive class has the highest performance with a precision of 99.70%, recall of 95.81% and F1 value of 97.72%, showing that the model is very good at handling the dominant Positive class. The figures for the Negative class are 97.67% precision, 91.42% recall and 94.44% F1-score, which indicates good discrimination of negative reviews. The Neutral class has relatively poor performance metrics: precision is 61.11%, recall is 93.45%, and F1-score is 73.89% because it has the larger number of neutral reviews among the three classes, and the reviews are harder to distinguish between the positive and the negative. The macro-average F1 score of 88.68% shows a balanced performance across all three classes, while the weighted-average F1 score of 95.46% highlights the good performance of the majority Positive class.

Table 12. Comparison of the proposed DBN-GRU model with existing sentiment analysis methods

Method	Dataset / Domain	Classes	Accuracy (%)
RoBERTa [17]	Women's clothing e-commerce reviews	Multi-class	87.69
LR [21]	E-commerce product reviews	As reported in [21]	90
CNN-LSTM [15]	Amazon electronics reviews	2	94
BERT-CNN-BiLSTM-Att [22]	Douban film short comments (Chinese)	As reported in [22]	84.93
DBN-GRU (Present Study)	**568,454 product reviews**	**3**	**95**

Table 12 compares the classification accuracy of the proposed DBN-GRU model with representative sentiment analysis approaches reported in previous studies. RoBERTa [17] achieved an accuracy of 87.69% on women's clothing e-commerce reviews, while LR [21] reported 90% accuracy for e-commerce product reviews. The CNN-LSTM approach [15] achieved 94% accuracy on Amazon electronics reviews, and the BERT-CNN-BiLSTM-Att model [22] reported 84.93% accuracy on Chinese Douban film comments. The proposed DBN-GRU model achieves 95.00% accuracy on a dataset containing 568,454 product reviews with three sentiment classes: Negative, Neutral, and Positive. The comparison should be interpreted in the context of differences in datasets, domains, class definitions, and experimental settings; therefore, the reported accuracies are not strictly controlled head-to-head results.

4.2. Discussion

Under the present experimental setting, the DBN-GRU framework achieved higher reported scores than the SVM, LR, and CNN-LSTM baselines, although these baselines were evaluated on different datasets. The framework captures hierarchical feature representations and sequential dependencies in the data. The model achieved a recall of 98, a precision and accuracy of 95, and an F1-score of 96.47, indicating strong recall and a high balance between precision and recall. Its advantages are that it can handle non-linear relationships in the dataset, is robust at learning temporal patterns, and showed stable performance across repeated cross-validation runs, making it an appropriate method for real-world sentiment analysis problems where context and sequence are of significant consideration. Nonetheless, the model also has its disadvantages. The DBN-GRU model is more expensive and more complex to train than simpler models, such as LR or SVM, and it can be more resource-intensive to process and run, making it less practical in low-resource settings. In addition, its complexity may undermine interpretability, as deep models can be viewed as black boxes, and additional explainability methods are needed to ensure reliability in sensitive settings, such as SHAP or LIME. Despite these weaknesses, the proposed work presents an effective sentiment classification approach under the present experimental setting and opens avenues for future research into interpretable and computationally efficient deep learning models.

This performance may be attributed to the complementary strengths of the two main components of the proposed Hybrid DBN-GRU framework. Unlike traditional machine learning models that use hand-crafted features, the DBN learns hierarchical semantic representations of product reviews, effectively removing noise and redundant information for classification. This enables the subsequent GRU network to obtain more informative feature representations, leading to better contextual understanding and improved sentiment prediction.

The framework aims to balance feature-representation capability and computational cost. The computational comparison (Table 11) shows that DBN-GRU requires more resources than the GRU-only model but fewer than CNN-LSTM; the cost of BERT-based models was not measured in this study. CNN-based models primarily capture local contextual information, whereas LSTMs model long-range dependencies at a higher computational cost because of their more complex gating structure. The GRU, on the other hand, is a lightweight network that models sequential dependencies efficiently and converges faster with lower computational complexity. The DBN-GRU framework achieved recall of 98%, precision of 95%, accuracy of 95%, and an F1-score of 96.47% under the current 3-class experimental configuration. These results reflect the current dataset and classification

procedure, not a comparison with results from other studies using different datasets and binary sentiment settings. The results show that DBN-based semantic feature extraction, combined with a GRU-based model for contextual sequence modelling, is effective for three-class sentiment analysis on the dataset studied.

5. Conclusion and Future Work

In this paper, we investigate a hybrid DBN-GRU approach to three-class sentiment classification of product reviews. This work has been reported in the past for the DBN-GRU architecture, which is not a model of its own. Rather, the framework was tested in a large-scale three-class setting, the contribution of the DBN and GRU components was evaluated by ablating one or both, and the stability of the framework was investigated by repeated cross-validation and bootstrap confidence intervals. DBN stage: Hierarchical feature refinement; GRU model: Sequential context. With the current experimental setup, the framework attained 95% accuracy, F1-score of 96.47% and outperformed the three chosen as baselines (SVM, LR, CNN-LSTM). Using class-wise analysis, it can be seen that the Neutral reviews are the most challenging class. The model was tested on a single dataset, so it needs to be tested on other domains/datasets in the future. The model achieved the best performance compared to classic models (SVM, LR, CNN-LSTM) in terms of accuracy, recall, precision and F1-score under the current experimental setup. Since the model was tested with one dataset, it requires further testing on other datasets to create generalization of the model for future application. Other domains remain to be tested in future work, which can support better decision-making for both businesses and researchers. Most importantly, this study combines a Hierarchical Representation Learning (HRN) framework using a DBN with a GRU-based representation learning framework that models word-sequence context into a single sentiment classification framework for product reviews. The study aims to combine these two networks, DBN and GRU, with complementary functions: DBN performs feature refinement, and GRU models sequential contextual information, without creating a new network model. Both positions highlight the novelty of the proposed framework relative to the standalone CNN, LSTM, GRU, and Transformer-based approaches presented in the related work.

Future studies can also aim to make the DBN-GRU model more interpretable by incorporating explainable AI methods such as SHAP or LIME, thereby making its predictions more transparent to end-users. One can also reduce computational overhead by compressing models, pruning them, or distilling knowledge, allowing deployment in low-resource or real-time settings. Future research examining transfer learning with pretrained embeddings such as BERT or GPT-based representations may improve semantic understanding, especially when applied to domain-specific sentiment datasets. Lastly, the framework should be extended to address multimodal sentiment analysis (combining text, audio, and visual information) to gain deeper insight into customer feelings and build on business knowledge.

References

[1] Feisal Adel Ramadhan, Rd. Ridwan Permana Ruslan, and Amalia Zahra, "Sentiment Analysis of E-Commerce Product Reviews for Content Interaction using Machine Learning," *Cakrawala Repositori IMWI*, vol. 6, no. 1, pp. 207-220, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Derwin Suhartono et al., "Deep Neural Networks and Weighted Word Embeddings for Sentiment Analysis of Drug Product Reviews," *Procedia Computer Science*, vol. 216, pp. 664-671, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] M. Irfan Marwat et al., "Sentiment Analysis of Product Reviews to Identify Deceptive Rating Information in Social Media: A SentiDeceptive Approach," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 3, pp. 830-860, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Vandana Yadav, Parul Verma, and Vinodini Katiyar, "Long Short Term Memory (LSTM) Model for Sentiment Analysis in Social Data for E-Commerce Products Reviews in Hindi Languages," *International Journal of Information Technology*, vol. 15, no. 2, pp. 759-772, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] U.B. Mahadevaswamy, and P. Swathi, "Sentiment Analysis using Bidirectional LSTM Network," *Procedia Computer Science*, vol. 218, pp. 45-56, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ying Cao et al., "A Study of Sentiment Analysis Algorithms for Agricultural Product Reviews based on Improved BERT Model," *Symmetry*, vol. 14, no. 8, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Weidong Huang, Miao Lin, and Yuan Wang, "Sentiment Analysis of Chinese E-Commerce Product Reviews using ERNIE Word Embedding and Attention Mechanism," *Applied Sciences*, vol. 12, no. 14, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shunxiang Zhang et al., "A Data Processing Method based On Sequence Labeling and Syntactic Analysis for Extracting New Sentiment Words from Product Reviews," *Soft Computing*, vol. 26, no. 2, pp. 853-866, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Amjad Iqbal et al., "Sentiment Analysis of Consumer Reviews using Deep Learning," *Sustainability*, vol. 14, no. 17, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Roobaea Alroobaea "Sentiment Analysis on Amazon Product Reviews using the Recurrent Neural Network (RNN)," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Zhenyu Zhang et al., "Product Selection Based on Sentiment Analysis of Online Reviews: An Intuitionistic Fuzzy TODIM Method," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3349-3362, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mahammad Khalid Shaik Vadla, Mahima Agumbe Suresh, and Vimal K. Viswanathan, "Enhancing Product Design Through AI-Driven Sentiment Analysis of Amazon Reviews using BERT," *Algorithms*, vol. 17, no. 2, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] J. Sangeetha, and U. Kumaran, "A Hybrid Optimization Algorithm using BiLSTM Structure for Sentiment Analysis," *Measurement: Sensors*, vol. 25, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Aakriti Gupta et al., "Sentiment Analysis of Product Reviews: Deep Learning Approach," *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)*, Delhi, India, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohammad Eid Alzahrani et al., "Developing an Intelligent System with Deep Learning Algorithms for Sentiment Analysis of E-Commerce Product Reviews," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] P. Bhuvaneshwari et al., "Sentiment Analysis for User Reviews using Bi-LSTM Self-Attention based CNN Model," *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 12405-12419, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Oumaima Bellar, Amine Baina, and Mostafa Ballafkih, "Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers," *Mathematics*, vol. 12, no. 15, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Gagan Kumar Patra et al., "A Sentiment Analysis of Customer Product Review based on Machine Learning Techniques in E-Commerce," *SSRN*, vol. 2, no. 4, pp. 1-4, 2024. [Google Scholar] [Publisher Link]

[19] Kamran Aziz et al., "Unifying Aspect-Based Sentiment Analysis BERT and Multi-Layered Graph Convolutional Networks for Comprehensive Sentiment Dissection," *Scientific Reports*, vol. 14, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mehmet ISIK, Online Product Reviews Analysis: Customer Feedback, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mehmetisik/amazon-review

[21] Manal Loukili, Fayçal Messaoudi, and Mohammed El Ghazi, "Sentiment Analysis of Product Reviews for E-Commerce Recommendation based on Machine Learning," *International Journal of Advances in Soft Computing & its Applications*, vol. 15, no. 1, pp. 1-13, 2023. [Google Scholar] [Publisher Link]

[22] Aixiang He, and Mideth Abisado, "Text Sentiment Analysis of Douban Film Short Comments based on BERT-CNN-BiLSTM-Att Model," *IEEE Access*, vol. 12, pp. 45229-45237, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Shaista Khan et al., "A Hybrid DBN-GRU Model for Enhanced Sentiment Analysis in Product Reviews," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 7, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]