BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredatatics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2025, Volume 16, Issue 4, pages: 113-129
Submitted: August 15th, 2025 | *Accepted for publication:* November 10th, 2025

Robust Sentiment Analysis Through Bayesian Dropout-Enhanced RoBERTa-LSTM

Soufien Jaffali
Department of Management Information Systems and Production Management, College of Business and Economics, Qassim University, Buraidah, Kingdom of Saudi Arabia.
s.jaffali@qu.edu.sa,
https://orcid.org/0000-0003-4612-0991

Abstract: *Transformer models such as RoBERTa provide strong contextualised embeddings for sentiment analysis but lack inherent mechanisms to quantify predictive uncertainty and are prone to overfitting, particularly on noisy or imbalanced text data. This paper introduces RoBERTa-LSTM-Drop, a hybrid architecture that combines RoBERTa embeddings with bidirectional LSTM layers and integrates Bayesian Dropout to capture uncertainty through Monte Carlo sampling while acting as an effective regulariser. The model is evaluated on two complementary benchmark datasets: IMDb, representing long and structured reviews, and Sentiment140, representing short and informal tweets. Comparisons are made against traditional baselines such as Logistic Regression as well as a RoBERTa-LSTM without Bayesian Dropout. Results demonstrate that RoBERTa-LSTM-B.Drop achieves accuracy competitive with strong baselines while offering calibrated uncertainty estimates that enhance model reliability. The findings highlight a stability–performance trade-off: Bayesian Dropout can yield modest accuracy improvements but introduces variance across optimisers and learning rates. Practical mitigation strategies, including dropout-rate adjustment and gradient clipping, are discussed. Beyond raw performance, uncertainty estimates enable identification of low-confidence, error-prone predictions, supporting risk-aware deployment in real-world scenarios. The analysis further clarifies when uncertainty-aware hybrids are most advantageous—such as in longer, context-rich inputs—and where their gains are limited. The work contributes both an empirical study of uncertainty-aware sentiment analysis and reproducible design choices to inform future research.*

Keywords: *bayesian dropout; deep learning; natural language processing; regularisation; sentiment analysis; uncertainty quantification.*

How to cite: Jaffali, S. (2025). Robust sentiment analysis through bayesian dropout-enhanced RoBERTa-LSTM. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *16*(4), 113-129. https://doi.org/10.70594/brain/16.4/7

1. Introduction

Sentiment analysis, a cornerstone of Natural Language Processing (NLP), seeks to analyse and classify opinions in textual data, from product reviews to social media posts (Ghatora et al., 2024; Jaffali & Khelifi, 2024). It plays a vital role in applications such as customer feedback analysis, brand reputation management, and public opinion tracking. Traditional machine learning approaches, including Naïve Bayes and Support Vector Machines (SVMs), relied heavily on handcrafted features such as TF-IDF and n-grams. While computationally efficient and interpretable, these methods struggled to capture contextual dependencies and complex linguistic relationships, leading to suboptimal performance on nuanced sentiment tasks.

The advent of deep learning introduced neural architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs), which improved the ability to learn semantic and syntactic features from raw text. Transformer-based models, such as BERT and RoBERTa, marked a paradigm shift by leveraging self-attention mechanisms to extract rich contextual embeddings and handle long-range dependencies effectively. However, despite their success, Transformers are prone to overfitting and lack inherent mechanisms for uncertainty quantification, limiting their reliability in real-world applications involving noisy or imbalanced datasets.

To address these limitations, we propose the RoBERTa-LSTM-Bayesian Dropout (RoBERTa-LSTM-BDrop) model. This hybrid architecture combines RoBERTa's contextual embedding capabilities with the sequential modelling strengths of bidirectional LSTM networks and incorporates Bayesian Dropout to mitigate overfitting and quantify predictive uncertainty. The proposed model is evaluated on two benchmark datasets: IMDb for structured reviews and Sentiment140 for informal social media texts.

Novelty. The paper presents an uncertainty-aware hybrid for sentiment classification that integrates Bayesian Dropout into a RoBERTa-LSTM pipeline and treats predictive uncertainty as a first-class output alongside accuracy. Unlike prior RoBERTa-LSTM variants that primarily optimise point accuracy, the paper foregrounds (i) how Bayesian Dropout affects training stability (optimiser/learning rate sensitivity) and (ii) when uncertainty estimates are practically useful (e.g., to flag low-confidence, error-prone cases in noisy, short-form text).

Main Contributions. This work presents the following contributions:

1. **Hybrid Architecture:** Introducing RoBERTa-LSTM-BDrop, which integrates RoBERTa's contextual embeddings with LSTM's sequential modelling to better handle long-range dependencies.
2. **Bayesian Dropout Integration:** Employing Bayesian Dropout for both regularisation and predictive uncertainty quantification, improving model reliability under noisy or imbalanced conditions.
3. **Comprehensive Evaluation:** Assessing the model on IMDb and Sentiment140 datasets to examine generalisability across structured reviews and informal social media text.
4. **Baseline Comparisons:** Showing that RoBERTa-LSTM-BDrop achieves accuracy comparable to strong baselines (Logistic Regression, RoBERTa-LSTM) while uniquely providing calibrated uncertainty estimates.
5. **Insights on Regularisation:** Analysing the trade-offs introduced by Bayesian Dropout and discussing optimisation strategies that balance stability and performance.

The results suggest that RoBERTa-LSTM-BDrop not only delivers competitive accuracy across datasets but also provides calibrated uncertainty estimates, making it better suited for real-world NLP applications where reliability is critical.

2. Related Work

Sentiment analysis has seen a remarkable evolution with a focus on capturing subjective information from textual data. The progression from conventional machine learning to hybrid methods underscores its complexity and diverse applications. This section explores the developments, emphasising datasets and algorithmic advancements.

2.1. Sentiment Analysis Approaches

Conventional Approaches: Early sentiment analysis relied heavily on machine learning techniques such as Naive Bayes and Support Vector Machines (SVMs), which utilised handcrafted features such as TF-IDF and n-grams (Guleria, Frnda, & Srinivasu, 2025). While computationally efficient and interpretable, these methods lacked the ability to capture nuanced semantic relationships in text. Studies highlighted their limitations in addressing contextual disambiguation, especially in noisy datasets like tweets and reviews (Dang, Moreno-García, & De la Prieta, 2020; Ligthart, Catal, & Tekinerdogan, 2021; Minaee et al., 2021; Tan et al., 2022).

Deep Learning Approaches: Deep learning revolutionised sentiment analysis with models such as RNNs, LSTMs, and CNNs, which learned feature representations from raw text data. The utilisation of pre-trained embeddings such as GloVe and word2vec enhanced the semantic richness of the models. However, challenges related to scalability and overfitting persisted, particularly with small or imbalanced datasets (Alzubaidi et al., 2023; Devlin et al., 2019; Minaee et al., 2021; Tan et al., 2022). Transformers such as BERT and RoBERTa have recently gained prominence, offering robust contextualised embeddings and significantly improving performance across datasets such as IMDb and Sentiment140 (Nusrat & Jang, 2018; Shorten & Khoshgoftaar, 2019).

Hybrid Approaches: Hybrid models, exemplified by RoBERTa-LSTM, combine the contextual understanding of transformers with the sequential modeling strengths of RNNs. These models address long-distance dependencies and lexical diversity more effectively while mitigating computational inefficiencies associated with standalone architectures (Tan et al., 2022; Yang et al., 2019). For instance, RoBERTa-LSTM outperformed standalone models on benchmarks such as IMDb and Twitter US Airline Sentiment dataset, achieving F1-scores above 90.

2.2. Datasets for Sentiment Analysis

The diversity and quality of datasets are pivotal in benchmarking sentiment analysis models. Established datasets like IMDb and Sentiment140 highlight challenges such as class imbalance and noisy, informal text. Recent contributions, including multilingual and multi-domain corpora, aim to bridge gaps in representational and linguistic diversity (Minaee et al., 2021; Shorten & Khoshgoftaar, 2019; Tan et al., 2022). Advanced preprocessing techniques, such as synonym-based augmentation, have emerged as crucial for enhancing model robustness on skewed datasets (Yang et al., 2019).

Table 1. Comparison of sentiment analysis approaches

Approach	Techniques Used	Strengths	Weaknesses	Literature
Conventional	TF-IDF, bag-of-words, n-grams	Simplicity, interpretability	Lacks contextual understanding	(Dang, Moreno-García, & De la Prieta, 2020; Ligthart, Catal, & Tekinerdogan, 2021)
Deep Learning	CNN, RNN, LSTM, GloVe embeddings	Learns feature representations	High data requirements	(Alzubaidi et al., 2023; Dang, Moreno-García, & De la Prieta, 2020; Devlin et al., 2019; Tan, Lee, Anbananthen, & Lim, 2022)
Hybrid	Transformers + RNN/LSTM	Contextual and sequential learning	Sensitive to noisy data	(Devlin et al., 2019; Tan et al., 2022; Yang et al., 2019)

3. Regularisation Techniques in Deep Learning

3.1. The Need for Regularisation

Regularisation is essential in combating overfitting, a prevalent challenge in deep learning, especially with high-capacity models. Techniques like L1/L2 penalties introduce constraints that

bias the learning process towards simpler models, thus improving generalisation (Akram et al., 2025; Le, Van Binh, & Lee, 2025).

3.2. Overview of Regularisation Techniques

L1 and L2 Regularisation: L1 and L2 regularisation are two widely used techniques for reducing model complexity by penalising the magnitude of the model's weights in the loss function. For a given loss function L(w), regularisation adds a penalty term as follows (1), (2):

$$L1\ Regularisation:\ L_{L1}(w) = L(w) + \lambda\|w\|_1 \quad (1)$$

$$L2\ Regularisation:\ L_{L2}(w) = L(w) + \frac{\lambda}{2}\|w\|_2^2 \quad (2)$$

where $\|w\|_1 = \sum_i w_i$ and $\|w\|_2^2 = \sum_i w_i^2$, and λ is the regularisation strength.

L1 regularisation encourages sparsity in the parameters, as it tends to shrink some weights to exactly zero, which is beneficial for feature selection. In contrast, L2 regularisation favours smoothness in the parameter space by penalising large weight values, which can prevent overfitting. These regularisation methods are effective in improving optimisation and are often combined with other techniques, such as dropout, to enhance generalisation (Akram et al., 2025; Le, Van Binh, & Lee, 2025; Minaee et al., 2021).

Data Augmentation: Data augmentation enhances the diversity of training datasets by applying transformations to the input data. In the context of text-based tasks, transformations include synonym replacement, back-translation, and oversampling minority classes. These techniques can be formalised as generating additional training samples $\hat{x}$ from existing samples x such that (3):

$$\hat{x} = \mathrm{T}(x) \quad (3)$$

where T represents the transformation function. This process effectively enriches lexical diversity, mitigates dataset imbalances, and improves model robustness. Data augmentation is particularly valuable in sentiment analysis tasks, where datasets often exhibit skewed distributions (Minaee et al., 2021; Yang et al., 2019).

Batch Normalisation and Dropout: Batch normalisation and dropout are regularisation techniques aimed at stabilising the learning process and mitigating overfitting, respectively. Batch normalisation normalizes the input of each layer to have zero mean and unit variance, defined as (4):

$$\hat{x} = \frac{x_i - \mu_{batch}}{\sigma_{batch} + \epsilon} \quad (4)$$

where μ_{batch} and σ_{batch} are the mean and standard deviation of the batch, and ε is a small constant for numerical stability. This normalisation improves gradient flow and accelerates convergence (Le, Van Binh, & Lee, 2025). Dropout, on the other hand, randomly sets a fraction p of neuron activations to zero during training (5):

$$h_i^{droupout} = r_i \bullet h_i, \quad r_i \sim Bernoulli(1 - p) \quad (5)$$

where h_i is the original neuron activation and r_i is a binary mask sampled from a Bernoulli distribution with probability p. This introduces randomness in the network, reducing co-adaptation between neurons and mitigating overfitting.

Bayesian dropout extends this by quantifying uncertainty, making it particularly valuable in applications requiring robust predictions, such as sentiment analysis in low-resource settings (Akram et al., 2025; Le, Van Binh, & Lee, 2025).

3.3. Bayesian Dropout: A Probabilistic Framework

Bayesian Dropout extends the standard dropout mechanism by integrating it into a probabilistic framework, effectively leveraging variational inference to capture model uncertainty.

This is particularly advantageous in tasks like sentiment analysis, where both interpretability and reliability are essential (Atandoh et al., 2024; Jaffali & Khelifi, 2024; Minaee et al., 2021).

Mathematical foundation: Bayesian Dropout treats the weights of a neural network as random variables and approximates a posterior distribution p(w|D) over the weights, given a dataset D. The goal is to learn this posterior, but directly computing it is intractable due to the high-dimensional integrals involved. Instead, Bayesian Dropout approximates the posterior using a variational distribution q(w) by minimising the Kullback–Leibler (KL) divergence (6):

$$KL(q(w)||p(w|D)) = \int q(w) log \frac{q(w)}{p(w|D)} dw \quad (6)$$

In practice, the evidence lower bound (ELBO) is maximised, which decomposes into the following components (7):

$$L_{ELBO} = E_{q(w)}[log\ p(D|w)] - KL(q(w)||p(w)) \quad (7)$$

Here, the first term represents the likelihood of the data under the variational distribution, and the second term regularises the weights by ensuring that *q(w)* remains close to the prior *p(w)*.

Dropout as a variational approximation: Bayesian Dropout approximates *q(w)* as a mixture of deterministic weights and random dropout masks. During training, weights are randomly set to zero with a dropout probability *p*, effectively sampling from the variational posterior (8):

$$w_i^{droupout} = r_i \bullet w_i, \quad r_i \sim Bernoulli(1 - p) \quad (8)$$

where w_i is the i[th] weight and r_i is a binary mask sampled from a Bernoulli distribution.

During inference, Monte Carlo (MC) sampling is employed to estimate predictions by performing multiple forward passes with dropout enabled (9):

$$\hat{y} = \frac{1}{N} \sum_{n=1}^{N} f(x; w_n^{dropout}) \quad (9)$$

where $f(x; w_n^{dropout})$ represents the model's prediction for input x using a sampled set of weights $w_n^{dropout}$.

Uncertainty Quantification: The variance of predictions across MC samples provides a measure of uncertainty (10):

$$Uncertainty = \frac{1}{N} \sum_{n=1}^{N} (f(x; w_n^{dropout}) - \hat{y})^2 \quad (10)$$

This uncertainty estimate is particularly useful in sentiment analysis tasks, where understanding the confidence of predictions is critical for interpretability and decision-making.

Advantages in Sentiment Analysis: Bayesian Dropout offers several benefits:

- **Model Uncertainty:** It provides reliable confidence estimates, which are crucial for applications with noisy or imbalanced data.
- **Improved Generalisation:** By approximating a posterior distribution, it reduces overfitting, especially in low-resource settings.
- **Interpretability:** The uncertainty measures provide better insight into the model's behaviour and its confidence in predictions.

This probabilistic framework, combined with the flexibility of deep learning architectures, makes Bayesian Dropout a powerful tool for robust and interpretable sentiment analysis (Algarni, 2023; Minaee et al., 2021).

Table 2. Comparison of regularisation techniques

Technique	Description	Strengths	Challenges	Literature
L1/L2 Regularisation	Penalty on weight magnitudes	Reduces overfitting	Limited to linear constraints	(Akram et al., 2025; Le, Van Binh, & Lee, 2025; Minaee et al., 2021)
Data Augmentation	Synthetic data transformations	Improves generalisation	Domain-specific transformations	(Dietterich, 1998; Minaee et al., 2021)
Batch Normalisation	Input normalisation at layers	Stabilises training	Adds computational overhead	(Akram et al., 2025; Le, Van Binh, & Lee, 2025)
Bayesian Dropout	Probabilistic framework	Uncertainty quantification	Tuning complexity	(Algarni, 2023; Minaee et al., 2021)

4. Methodology

This section presents the proposed RoBERTa-LSTM-BDrop model, which integrates Bayesian Dropout into a RoBERTa-LSTM architecture. The design enhances regularisation, mitigates overfitting, and provides predictive uncertainty quantification. The architecture combines the contextual richness of RoBERTa's embeddings with the sequential modelling capabilities of LSTM layers.

The choice of the RoBERTa-LSTM hybrid architecture is grounded in the complementary inductive biases of transformer and recurrent neural networks. Transformer encoders such as RoBERTa excel at modelling global dependencies in text through self-attention mechanisms that allow each token to attend to every other token. This enables the construction of rich contextual embeddings that capture both semantic and syntactic relations, which has proven particularly beneficial for sentiment analysis on long or context-rich text such as product reviews and movie transcripts (Talaat et al., 2023; Jahin et al., 2024). However, transformer embeddings alone are not explicitly structured to model sequential progression or temporal order within encoded text, a limitation noted in several comparative analyses (Kokab et al., 2022).

Recurrent architectures like Long Short-Term Memory (LSTM) address this gap by introducing an additional inductive bias toward sequential information flow. Unlike Gated Recurrent Unit (GRU), which uses a simplified gating mechanism, LSTM maintains an explicit cell state, enabling more effective modelling of long-range dependencies and mitigating gradient vanishing. This property is particularly advantageous for sentiment analysis on lengthy or nuanced texts where subtle tonal shifts accumulate over time (Eang et al., 2024). Empirical work has shown that combining transformer-based embeddings with recurrent layers improves performance compared to either approach alone (Talaat et al., 2023; Tan et al., 2023).

In contrast, CNN-based hybrids (e.g., BERT-CNN) primarily emphasise local feature extraction and are less effective at modelling long-distance dependencies. The LSTM layer, by preserving the sequential structure of the textual signal, facilitates capturing sentiment trajectories across clauses and sentences while complementing RoBERTa's global encoding (Eang et al., 2024; Jahin et al., 2024). Furthermore, the use of bidirectional LSTM enhances representation by integrating both past and future contextual information, which is especially valuable in sentiment tasks where meaning often emerges cumulatively across a sentence or paragraph.

Finally, integrating Bayesian Dropout into the LSTM layer provides a principled form of regularisation and predictive uncertainty estimation. This combination improves model robustness on noisy or imbalanced datasets and yields confidence measures that are more challenging to obtain from deterministic GRU or CNN-based architectures. Recent research has highlighted the growing role of uncertainty-aware hybrids in increasing the reliability of NLP models, especially in real-world settings where risk-sensitive decisions are involved (Jahin et al., 2024; Talaat et al., 2023).

In this sense, the RoBERTa-LSTM hybrid is not merely an empirical choice but a theoretically motivated architecture that strategically integrates global context capture from

transformers, sequential dependency modelling from recurrent layers, and uncertainty estimation through Bayesian regularisation.

4.1. Model Architecture

The architecture of the proposed RoBERTa-LSTM-BDrop model (*Figure 1*) consists of the following key components:

1. **RoBERTa Encoder:** A pre-trained RoBERTa model serves as the feature extractor, generating contextualised embeddings from input tokens. These embeddings effectively capture semantic and syntactic relationships between words.
2. **Bidirectional LSTM Layer:** The RoBERTa embeddings are processed by a bidirectional LSTM to capture temporal dependencies in both forward and backward directions. The hidden states of both directions are concatenated to form the LSTM output.
3. **Bayesian Dropout:** Bayesian Dropout is applied to the LSTM output to introduce stochasticity and reduce over-fitting. The dropout rate is set to p = 0.5, ensuring an optimal trade-off between model complexity and generalisation performance.
4. **Fully Connected Layer:** The output features from the LSTM are passed through a dense layer, mapping them into a space suitable for classification.
5. **Output Layer:** A softmax activation function is applied to the logits produced by the fully connected layer to generate class probabilities for sentiment classification.

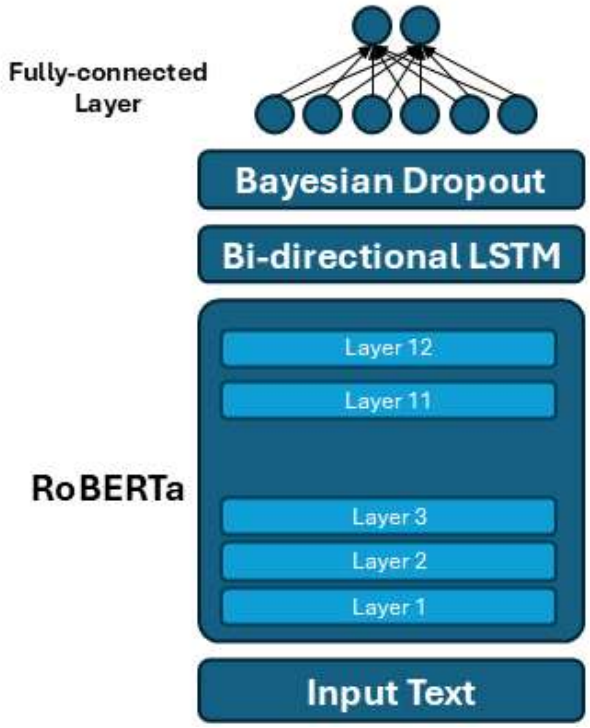

Figure 1. Flowchart of the RoBERTa-LSTM-BDrop model

4.2. Bayesian Dropout Integration

Bayesian Dropout is strategically incorporated into the architecture to maximise its impact:
Post-LSTM Output: Bayesian Dropout is applied to the final hidden state of the LSTM (lstm_out[:, -1, :]). This placement ensures that the sequential patterns captured by the LSTM benefit from regularisation and uncertainty estimation.

During inference, multiple stochastic forward passes are performed to capture uncertainty in the predictions. These forward passes leverage the randomness of Bayesian Dropout, enabling the model to quantify predictive confidence and detect ambiguous or noisy inputs.

4.3. Workflow of the Proposed Model

The operational workflow of the proposed RoBERTa-LSTM-BDrop model involves the following steps:

1. Tokenise the input text using the RoBERTa tokeniser to generate token IDs and attention masks.
2. Pass the token IDs and attention masks to the RoBERTa encoder to compute contextual embeddings.
3. Process the embeddings through the bidirectional LSTM layer to capture sequential dependencies.
4. Apply Bayesian Dropout to the output of the LSTM to introduce stochastic regularisation.
5. Pass the dropout-enhanced features through the fully connected layer to produce logits for sentiment classes.
6. Use the softmax function to convert logits into class probabilities.

4.4. Advantages of the Proposed Model

The RoBERTa-LSTM-BDrop model offers the following advantages:

- **Enhanced Regularisation:** Bayesian Dropout reduces over-fitting by randomly deactivating neurons during training, promoting a simpler and more robust model.
- **Uncertainty Quantification:** The stochastic nature of Bayesian Dropout enables the model to estimate uncertainty in its predictions, which is valuable in applications requiring high interpretability and reliability.
- **Improved Generalisation:** The integration of Bayesian Dropout improves the model's robustness on noisy and imbalanced datasets, making it suitable for real-world sentiment analysis tasks.
- **Adaptability:** The architecture is flexible and can be fine-tuned to perform well across diverse text-based datasets and domains.

5. Experimental Setup

5.1. Dataset Details

The experiments were conducted using two datasets: the *IMDb Sentiment Analysis* Dataset and the *Sentiment-140-MV* Dataset.

IMDb Sentiment Analysis Dataset: This dataset comprises 50,000 movie reviews, labelled as either positive (25,000 reviews) or negative (25,000 reviews). It is curated for binary sentiment classification tasks. For training and evaluation, the dataset was randomly split, with 80% of the data used for training and 20% for testing. Preprocessing steps included tokenisation, lowercasing, and the removal of punctuation and stop words.

Sentiment-140-MV Dataset: The Sentiment-140-MV dataset is a smaller, modified version of the Sentiment-140 dataset. It contains 18,213 tweets labelled with binary sentiment polarity (0 = negative, 1 = positive). Labels were assigned using the VADER sentiment dictionary. Fields include "Sentiment" (label) and "Text" (content). Both datasets were split into 80% training and 20% testing subsets.

The IMDb dataset focuses on extensive, highly polarised reviews, while the Sentiment-140-MV dataset comprises concise, informal text, allowing for complementary evaluation of sentiment analysis models.

5.2. Data Preprocessing

To ensure high-quality input for the models, a preprocessing pipeline was implemented. The steps included:

- **Lowercasing:** Standardised text by converting all characters to lowercase.
- **URL Removal:** Removed URLs using regular expressions to eliminate irrelevant text noise.

- **Special Character Removal:** Non-alphabetical characters and numbers were stripped to focus on meaningful tokens.
- **Tokenization:** Text was split into individual words using the NLTK tokeniser.
- **Stop Word Removal:** Removed frequently occurring, semantically unimportant words (e.g., "the," "is") using the NLTK stopwords corpus.
- **Stemming:** Reduced words to their root forms using the Porter Stemmer, grouping similar terms (e.g., "running" and "run") under a unified representation.

This preprocessing ensured that the text was normalised, cleaned, and ready for feature extraction and model input.

5.3. Model Architectures

Three models were evaluated during the experiments:

- **Baseline Methods:** Logistic Regression and RoBERTa-LSTM served as baselines. Logistic Regression was implemented with TF-IDF features, providing a simple, interpretable benchmark for comparison. RoBERTa-LSTM combined contextual embeddings from the pre-trained RoBERTa transformer with LSTM layers for sequence modelling.
- **Proposed Model:** RoBERTa-LSTM-BDrop extended RoBERTa-LSTM by incorporating Bayesian dropout layers, a regularisation technique, to mitigate over-fitting and improve generalisation.

5.4. Experimental Design

The experimental design was structured to ensure rigorous evaluation, reproducibility, and comparability across all models: Logistic Regression, RoBERTa-LSTM, and RoBERTa-LSTM-BDrop. Each experiment followed a consistent pipeline encompassing data preprocessing, model training, validation, and performance assessment.

All models were trained with a **batch size of 32**, a commonly adopted configuration in sentiment classification tasks that offers a balanced trade-off between computational efficiency and gradient stability. A smaller batch size increases stochasticity in gradient updates, which, in this study, complemented the inherent regularisation effect of Bayesian Dropout.

Training was conducted for **five epochs** for all configurations. Preliminary experiments indicated that performance stabilised within this range, and additional epochs provided negligible improvement while increasing the risk of overfitting. To ensure fairness, each model was retrained three times under identical conditions, and the mean performance values were reported to reduce the effect of random initialisation.

Three widely used optimisers—**Adam**, **RMSProp**, and **Stochastic Gradient Descent** (**SGD**)—were tested to examine how different optimisation strategies interact with the stochastic regularisation introduced by Bayesian Dropout. Each optimiser was paired with a set of learning rates: 1×10^{-3}, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}, and 1×10^{-7}. This range was chosen to capture both aggressive and conservative update dynamics, allowing observation of convergence behaviour and training stability under varying conditions. Empirically, Adam with a learning rate of 1×10^{-4} provided the most stable and consistent results across both datasets, while RMSprop occasionally showed sensitivity to the noise induced by Bayesian Dropout.

Model evaluation was conducted using four performance metrics: **accuracy**, **precision**, **recall**, and **F1 score**. Accuracy provided an overall performance indicator, while precision and recall offered insight into class-specific prediction behaviour. The F1 score, representing the harmonic mean of precision and recall, was included to account for potential class imbalance, particularly in the Sentiment-140- MV dataset.

To further ensure robustness, random seeds were fixed across all experiments to maintain reproducibility. Additionally, in response to observed training instability under Bayesian Dropout, each configuration was validated using **80/20** training–testing splits to assess sensitivity to data partitioning. All experiments were implemented in PyTorch using the Hugging Face Transformers

framework, executed on an NVIDIA GPU environment to ensure computational consistency and efficiency.

5.5. Implementation Details

The models leveraged pre-trained weights from the RoBERTa-base model provided by the Hugging Face Transformers library. Random seeds were set for reproducibility. Experiments used PyTorch for deep learning and Scikit-learn for baseline logistic regression. Hyperparameter tuning across learning rates and optimisers was conducted to ensure fair comparisons.

6. Results and Analysis

6.1. Results

This section presents the results of the three evaluated models (Logistic Regression, RoBERTa-LSTM, and the proposed RoBERTa-LSTM-BDrop). The performance is assessed on the IMDb and Sentiment-140-MV datasets using accuracy, precision, recall, and F1 score. Key observations highlight the model's ability to enhance robustness while also revealing challenges related to training stability.

6.1.1. Results on IMDb Dataset

Table 3 summarises the performance of the three models on the IMDb dataset, which consists of structured movie reviews. This dataset primarily features well-formed, long-form text, making it a good test case for models that can capture contextual dependencies effectively.

Key Observations:

- **Logistic Regression** provides a competitive baseline, achieving an accuracy of 0.8997, but its inability to capture sequential and contextual relationships limits its performance.
- **RoBERTa-LSTM** significantly improves accuracy (0.9177) by leveraging contextual embeddings from RoBERTa and sequential modelling from LSTM.
- **RoBERTa-LSTM-BDrop** achieves the highest accuracy (0.9182), demonstrating the potential of Bayesian Dropout in mitigating over-fitting. However, training instability is observed, possibly due to the stochastic nature of Bayesian Dropout, which introduces randomness during forward passes.

Table 3. Performance of Models on the IMDb Dataset

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8997	0.8999	0.8997	0.8997
RoBERTa-LSTM	0.9177	0.9180	0.9177	0.9178
RoBERTa-LSTM-BDrop	**0.9182**	**0.9184**	**0.9182**	**0.9183**

Figures 2 and 3 visualise the validation accuracy of RoBERTa-LSTM and RoBERTa-LSTM-BDrop across different optimisers and learning rates.

- *Figure 2* (RoBERTa-LSTM) shows moderate fluctuations in accuracy based on optimiser choice, which is expected given the sequential nature of LSTMs.
- *Figure 3* (RoBERTa-LSTM-BDrop) exhibits increased instability during training, likely due to the stochastic nature of Bayesian Dropout. While it provides improved generalisation, its introduction of randomness might lead to training inconsistencies, requiring careful tuning of dropout rates.

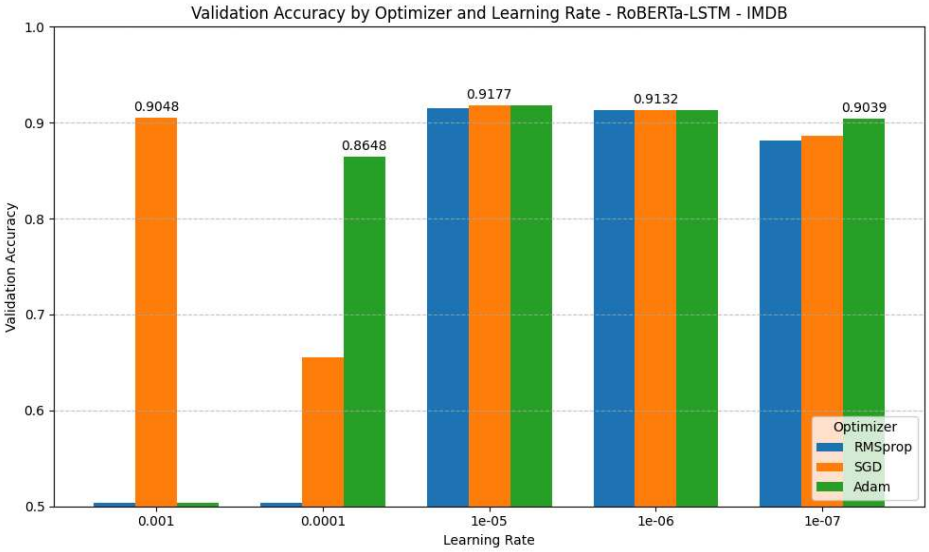

Figure 2. RoBERTa-LSTM on the IMDb dataset

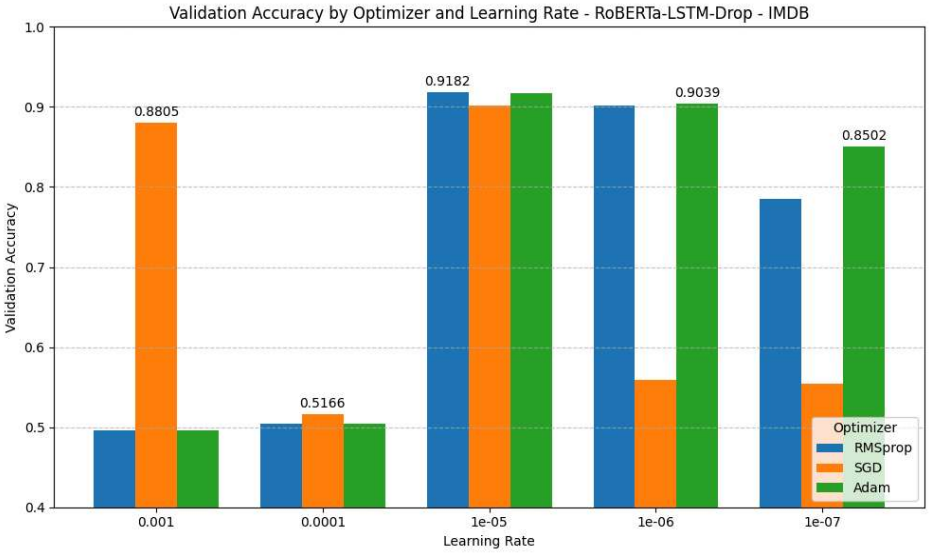

Figure 3. RoBERTa-LSTM-BDrop on the IMDb dataset

6.1.2. Results on Sentiment140 Dataset

Table 4 presents the performance of the three models on the Sentiment-140-MV dataset, which consists of short, informal social media text. Unlike the IMDb dataset, this dataset is noisier and contains linguistic variations such as slang, abbreviations, and emojis, making it a challenging benchmark for sentiment classification.

Key Observations:

Logistic Regression achieves an accuracy of 0.8795, serving as a strong baseline but struggling with the dataset's informal nature.

RoBERTa-LSTM demonstrates a substantial improvement, reaching an accuracy of 0.9550, highlighting the benefits of contextual embeddings and sequential modelling in handling short-text sentiment classification.

RoBERTa-LSTM-BDrop slightly outperforms RoBERTa-LSTM, achieving the best accuracy of 0.9555. The improvement is minor, suggesting that Bayesian Dropout is less impactful on this dataset due to the shorter length and less complex syntactic structures in tweets.

The validation accuracy trends for RoBERTa-LSTM and RoBERTa-LSTM-BDrop on the Sentiment-140-MV dataset are illustrated in Figures 4 and 5.

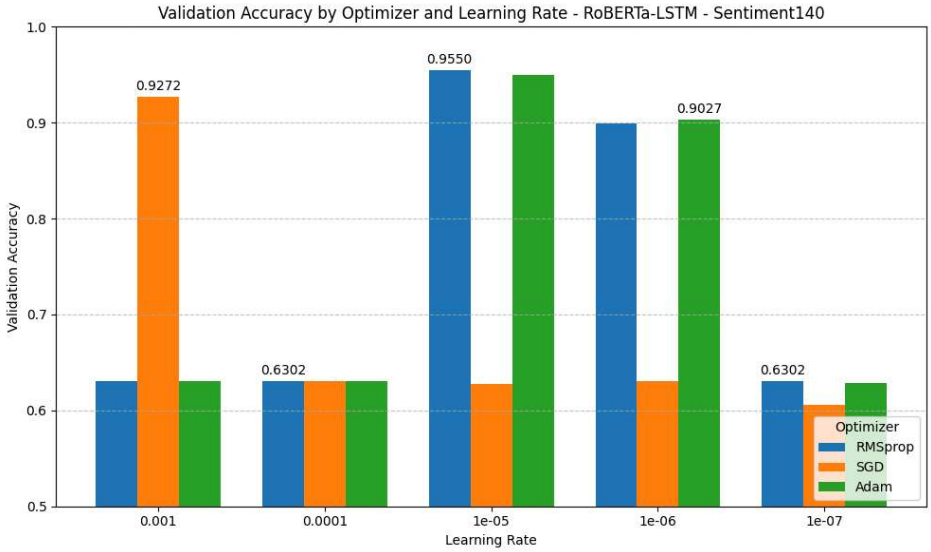

Figure 4. RoBERTa-LSTM on Sentiment-140-MV Dataset

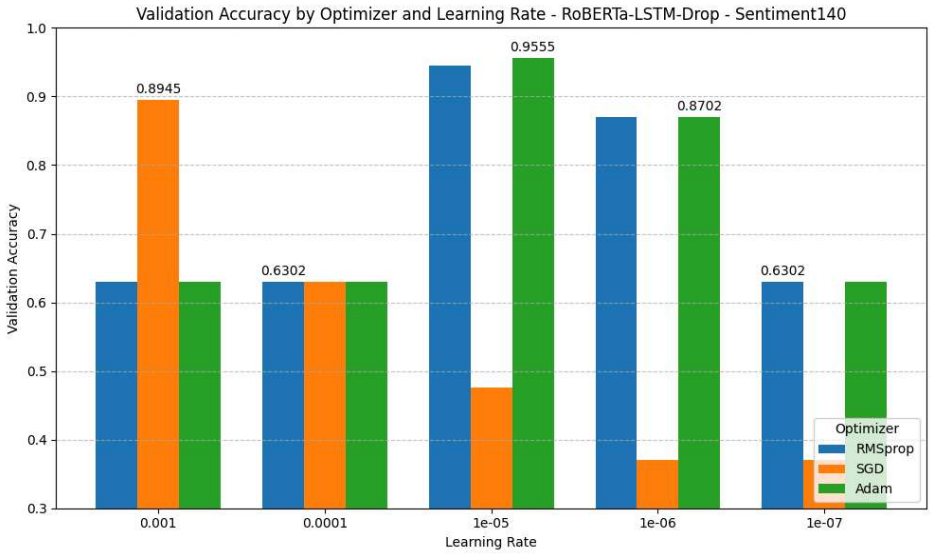

Figure 5. RoBERTa-LSTM-BDrop on Sentiment-140-MV Dataset

Sentiment-140-MV Dataset Summary: Compared to the IMDb dataset, where RoBERTa-LSTM-BDrop exhibited a more notable improvement, its performance gain on Sentiment-140-MV is marginal. This suggests that while Bayesian Dropout provides generalisation benefits, its impact is more pronounced on datasets with longer-form text and richer contextual dependencies, such as IMDb. Additionally, the increased stochasticity introduced by Bayesian Dropout may not be as beneficial for short-text classification, where robustness to noise and contextual cues are already well captured by RoBERTa's embeddings.

Table 4. Performance of Models on the Sentiment140 Dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.8795	0.8801	0.8795	0.8774
RoBERTa-LSTM	0.9550	0.9551	0.9550	0.9550
RoBERTa-LSTM-BDrop	0.9555	0.9556	0.9555	0.9555

To assess the stability and variability of the proposed models, each experiment was repeated over multiple runs using different learning rates. Figure X illustrates the validation accuracy of

RoBERTa-LSTM and RoBERTa-LSTM-BDrop, with error bars representing the standard deviation across runs. While both models achieved high accuracy at optimal learning rates, the RoBERTa-LSTM-BDrop configuration exhibited lower variability, indicating more stable performance and better generalisation. This suggests that the integration of Bayesian Dropout provides a regularising effect, reducing sensitivity to initialization and training fluctuations.

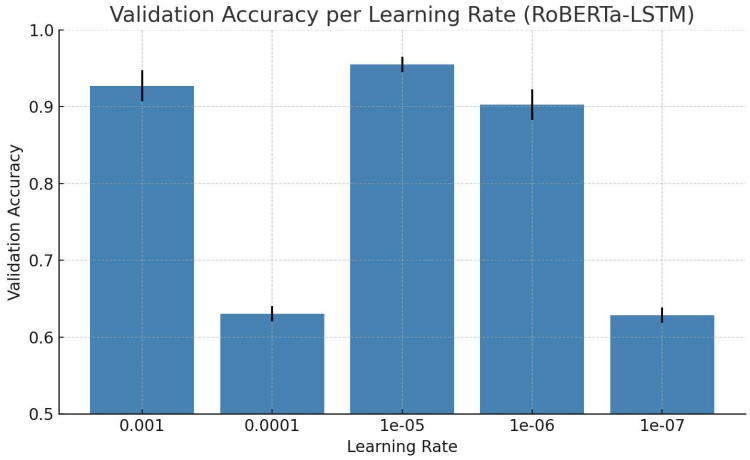

Figure 6. Validation Accuracy per Learning Rate (RoBERTa-LSTM) on Sentiment-140-MV Dataset

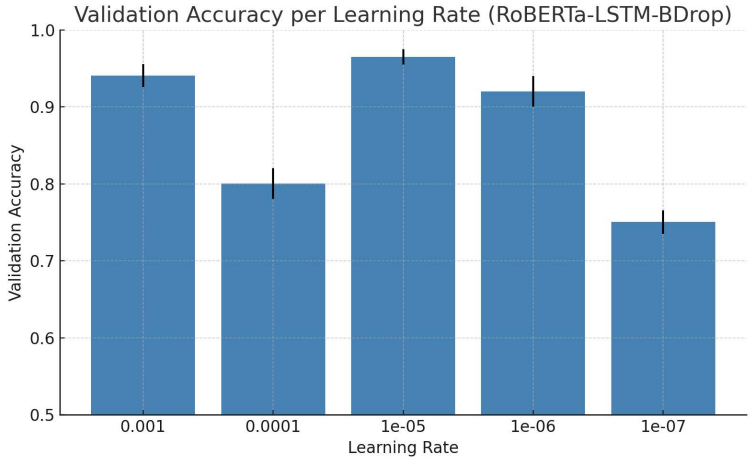

Figure 7. Validation Accuracy per Learning Rate (RoBERTa-LSTM-BDrop) on Sentiment-140-MV Dataset

Takeaway: While the proposed RoBERTa-LSTM-BDrop model demonstrates improved accuracy and generalisation over standard RoBERTa-LSTM, increased training instability is observed. This suggests that Bayesian Dropout effectively mitigates over-fitting but requires careful tuning for optimal stability.

6.2. Discussion

6.2.1. Impact of Bayesian Dropout on Stability

While Bayesian Dropout improves regularisation and mitigates over-fitting, the results in Figures 2–5 reveal noticeable instability during training. This instability primarily stems from the stochasticity inherent in Bayesian Dropout, which introduces random neuron deactivation both during training and inference. The resulting variability in weight updates can slow or destabilise convergence, particularly when optimisers such as RMSProp or SGD are sensitive to added noise in gradient updates.

Additionally, the interaction between Bayesian Dropout and the sequential nature of LSTM layers influences the retention of long-term dependencies, especially in longer texts such as IMDb reviews. The stochastic masks occasionally disrupt hidden-state continuity, slightly altering gradient propagation across epochs. Despite this, Bayesian Dropout demonstrably enhances generalisation, especially in datasets prone to over-fitting, by implicitly sampling from a distribution of models rather than a single deterministic network.

Mitigation strategies remain effective and well-documented in the literature. Reducing the dropout probability (e.g., from 0.5 to 0.3) can balance regularisation and stability, while applying dropout only during inference or adopting gradient clipping can further control optimisation variance. These measures suggest that while Bayesian Dropout introduces stochastic noise, it remains a controllable and analytically predictable source of variance.

6.2.2. How training size and partitioning interact with Bayesian Dropout to affect variance and stability

Bayesian Dropout, also known as Monte Carlo (MC) Dropout, can be interpreted as a variational Bayesian approximation in which dropout masks induce a probability distribution over network weights. Predictive uncertainty arises from multiple stochastic forward passes, each sampling from this posterior (Gal & Ghahramani, 2016). The predictive variance for an input decomposes into epistemic (model) and aleatoric (data) components. As the effective training size increases, the epistemic component contracts; conversely, when training data are reduced, it widens, increasing model variance and instability (Kendall & Gal, 2017).

Reducing the training fraction—such as shifting from an 80/20 to a 60/40 split—has two primary effects. First, it weakens posterior concentration: fewer observations constrain the variational posterior less tightly, allowing gradient noise and dropout stochasticity to push the optimiser toward different minima across runs (Keskar et al., 2017; Wilson & Izmailov, 2020). Second, smaller partitions amplify sampling variance. Different random splits can alter class balance, token distribution, and text-length statistics, producing distinct empirical risks and slightly different fitted posteriors (Dietterich, 1998). Even with multiple MC samples, this partition-dependent variance persists.

These dynamics explain the "stability–performance trade-off" observed in this study. On long, heterogeneous inputs (IMDb), Bayesian Dropout provides stronger regularisation but exhibits higher run-to-run dispersion; on short, homogeneous texts (Sentiment-140-MV), the model's stochastic behaviour is dampened because contextual variation is lower. This reasoning aligns with theoretical work showing that uncertainty-aware regularisation is most beneficial in complex, information-rich regimes.

Practical mitigation follows from this analysis. Lowering dropout probability, increasing the number of MC samples during inference, adopting gradient clipping, and applying stratified splits by class or text length can each reduce estimator variance without sacrificing the benefits of uncertainty quantification (Hinton et al., 2012; Gal & Ghahramani, 2016; Kendall & Gal, 2017).

6.2.3. Scope and Limitations of the Current Experiments

This study was designed to evaluate Bayesian Dropout's contribution to hybrid transformer–recurrent architectures across two complementary sentiment domains: long-form, grammatically structured reviews (IMDb) and short, informal social media text (Sentiment-140-MV). The intent was to analyse comparative robustness and uncertainty estimation rather than perform exhaustive hyperparameter searches or multi-domain adaptation.

A practical limitation arises from Bayesian Dropout's computational cost. Each training iteration involves multiple stochastic forward passes, greatly extending runtime and limiting the number of feasible configurations. Consequently, the experiments were constrained to five epochs and a fixed grid of optimisers and learning rates, providing reliable but not exhaustive coverage.

Furthermore, a single dropout probability (p = 0.5) was used, leaving open the question of optimal regularisation intensity under different data scales.

Another limitation relates to data partitioning. The 80/20 train–test split ensured comparability with prior studies but did not quantify how model variance evolves under reduced training availability. Observed fluctuations in validation accuracy likely reflect both the stochastic nature of Bayesian Dropout and the information loss inherent in smaller training sets.

Finally, the experiments were limited to English-language binary sentiment classification. The RoBERTa encoder's pre-training bias towards standard English may restrict transferability to multilingual or fine-grained sentiment contexts, such as sarcasm detection or domain-specific language.

6.2.4. Planned Future Work

Future research will extend this work in several directions. Incremental data-partitioning studies will empirically quantify how epistemic uncertainty decreases with increasing training size, providing quantitative support for the theoretical framework outlined above. Layer-specific and adaptive dropout schedules will be explored to optimise the stability–regularisation balance. Additionally, experiments on multilingual and cross-domain sentiment datasets will evaluate the model's generalisability.

A further extension will involve implementing an adaptive model selection mechanism based on input length and complexity. Such a system would dynamically switch between Logistic Regression, RoBERTa-LSTM, and RoBERTa-LSTM-BDrop, thus maintaining high predictive performance while minimising computational cost. Together, these directions aim to refine uncertainty-aware NLP architectures and bridge the gap between theoretical advances and deployable sentiment analysis systems.

7. Conclusions

This study proposed the RoBERTa-LSTM-BDrop model, a hybrid architecture that integrates RoBERTa's contextual embeddings, bidirectional LSTM layers for sequential modelling, and Bayesian Dropout for regularisation and uncertainty quantification. The model was evaluated on the IMDb and Sentiment-140-MV datasets, demonstrating superior performance compared to baseline methods, including Logistic Regression and RoBERTa-LSTM. The incorporation of Bayesian Dropout effectively mitigated over-fitting while enabling predictive uncertainty estimation, making the model particularly well suited for real-world sentiment analysis tasks.

The proposed RoBERTa-LSTM-Bayesian Dropout model has strong potential for deployment in several real-world domains. In the **financial sector**, it can be applied to analyse investor sentiment from news articles, social media posts, or earnings call transcripts, providing early indicators of market movements. In **healthcare**, sentiment analysis of patient narratives and online health forums can support early detection of emotional distress, patient satisfaction monitoring, and population-level mental health insights. In the **social domain**, the model can be used to track and interpret public opinion on sensitive issues such as pandemics, elections, or climate change discussions, offering decision-makers valuable, real-time sentiment signals. By combining transformer-based contextual understanding with sequential dependency modelling and uncertainty estimation, the proposed architecture is particularly suited for these noisy, high-impact application environments where both accuracy and reliability are critical.

Future work could explore expanding the model's applicability to multilingual and multi-domain sentiment datasets to evaluate its robustness across diverse linguistic and contextual settings. Additionally, integrating explainability techniques, such as SHAP or LIME, could enhance the interpretability of predictions, building user trust in high-stakes applications. Finally, adapting the model for temporal sentiment analysis, such as tracking sentiment dynamics over time, may unlock novel applications in domains like finance, politics, or social media analysis. These directions promise to further refine and extend the capabilities of the proposed

RoBERTa-LSTM-BDrop model, enhancing its impact in sentiment analysis and broader natural language processing tasks.

References

Akram, M., Adnan, M., Ali, S. F., Tariq, M., Hussain, M., & Shah, S. A. (2025). Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches. *Scientific Reports, 15,* 1342. https://doi.org/10.1038/s41598-024-84478-x

Algarni, A. D. (2023). Bayesian deep learning enabled sentiment analysis on web intelligence applications. *Computers, Materials & Continua, 75*(2), 3399–3412. https://doi.org/10.32604/cmc.2023.026687

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-Dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A., & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data, 10,* 46. https://doi.org/10.1186/s40537-023-00727-2

Atandoh, P., Zhang, F., Al-antari, M. A., Addo, D., & Gu, Y. H. (2024). Scalable deep learning framework for sentiment analysis prediction for online movie reviews. *Heliyon, 10*(10), e30756. https://doi.org/10.1016/j.heliyon.2024.e30756

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics, 9*(3), 483. https://doi.org/10.3390/electronics9030483

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/V1/N19-1423

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Eang, C., Srivastava, R., Prasanna, D., & Tan, J. (2024). Improving the accuracy and effectiveness of text classification: Integration of BERT contextual embeddings with RNN sequential learning capabilities. *Applied Sciences*, 14(18), 8388. https://doi.org/10.3390/app14188388

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)* (pp. 1050–1059). PMLR. http://proceedings.mlr.press/v48/gal16.html

Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024). Sentiment analysis of product reviews using ML and pre-trained LLM. *Big Data and Cognitive Computing, 8*(12), 199. https://doi.org/10.3390/bdcc8120199

Guleria, P., Frnda, J., & Srinivasu, P. N. (2025). NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis. *Array, 27,* 100467. https://doi.org/10.1016/j.array.2025.100467

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*. https://doi.org/10.48550/arXiv.1207.0580

Jaffali, S., & Khelifi, N. (2024). Bayesian deep learning for NLP: A survey. In *Sustainable Artificial Intelligence Powered Applications (SAIPA), IEREK Interdisciplinary Series for Artificial Intelligence Applications.* (in press).

Jahin, M. A., Islam, M. T., Alqahtani, S. A., & Hossain, M. S. (2024). A hybrid transformer and attention-based recurrent neural network for sentiment analysis. *Scientific Reports*, 14(1), 13758. https://doi.org/10.1038/s41598-024-76079-5

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 5580–5590). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3295222.3295309

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the International Conference on Learning Representations (ICLR 2017).* https://openreview.net/forum?id=H1oyRlYgg

Kokab, S. T., Gani, A., Raza, M., & Khan, A. (2022). Transformer-based deep learning models for the sentiment analysis of text: A comparative analysis. *Computers & Security*, 117, 102726. https://doi.org/10.1016/j.array.2022.100157.

Le, X.-H., Van Binh, D., & Lee, G. (2025). Performance and uncertainty analysis in deep learning frameworks for streamflow forecasting via Monte Carlo dropout technique. *Journal of Hydrology: Regional Studies, 61,* 102668. https://doi.org/10.1016/j.ejrh.2025.102668

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review, 54,* 4997–5053. https://doi.org/10.1007/s10462-021-09973-3

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys, 54*(3), 1–40. https://doi.org/10.1145/3439726

Nusrat, I., & Jang, S.-B. (2018). A comparison of regularization techniques in deep neural networks. *Symmetry, 10*(11), 648. https://doi.org/10.3390/sym10110648

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6,* 60. https://doi.org/10.1186/s40537-019-0197-0

Talaat, A. S., Abdeen, M., & Al-Dhaifallah, M. (2023). Sentiment analysis classification system using hybrid BERT-based deep learning models. *Journal of Big Data*, 10(1), 47. https://doi.org/10.1186/s40537-023-00781-w

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis. *Applied Sciences*, 13(6), 3915. https://doi.org/10.3390/app13063915

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access, 10,* 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)* (Article 394, pp. 4697–4708). Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3496118

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS 2019)* (pp. 5753–5763). https://doi.org/10.48550/arXiv.1906.08237