Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24522

A Comparative Study on Social Media Sentiment Classification Models Based on BERT Fine-tuning and Fusion with Sentiment Lexico

Zixing Lin[1*], Ruihao Chen[2], Wan Li[3]

[1] *Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangzhou, China,*

[2] *College of Marine Information Engineering, Jimei University, Xiamen, China*

[3] *Beijing No.11 High School, Beijing, China*

**Corresponding Author. Email: z.lin3.22@abdn.ac.uk*

Abstract. Social media sentiment analysis presents unique challenges due to informal language, sarcasm, and ambiguous emotional cues. While pretrained models like BERT have shown strong performance in capturing contextual semantics, they often struggle with neutral sentiment classification and domain-specific expressions. This study proposes a hybrid sentiment classification model that integrates BERT's contextual embeddings with sentiment lexicon features from SentiWordNet to address these limitations. Two models are compared: a baseline BERT model and an enhanced model incorporating lexicon-derived polarity scores. Experimental results on COVID-19-related tweets show that the enhanced model achieves more balanced performance across sentiment categories, particularly improving the detection of neutral expressions (F1-score increased from 0.86 to 0.89). However, the fusion approach introduces slight performance trade-offs, such as reduced overall accuracy (from 92.00% to 90.60%) due to lexicon-context misalignment and slang misinterpretation. The findings highlight the value of combining explicit sentiment priors with contextual embeddings and suggest future directions involving dynamic lexicon weighting and domain-specific lexicon generation to enhance robustness and adaptability.

Keywords: Sentiment Analysis, BERT Model, Sentiment Lexicons, Hybrid Models, Neutral Sentiment Classification

1. Introduction

Social media sentiment analysis faces significant challenges due to informal language, sarcasm, and mixed emotions. Pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized natural language processing (NLP) by leveraging bidirectional context understanding and transfer learning capabilities [1]. BERT excels in capturing deep semantic relationships through its transformer architecture, achieving robust performance

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24522

across diverse NLP tasks [2]. However, its reliance on implicit contextual patterns may limit sensitivity to subtle or domain specific emotional cues, such as detecting neutral sentiments in ambiguous contexts.

Sentiment lexicons, such as SentiWordNet, offer explicit polarity scores for words, providing coarse grained but interpretable emotional signals [3]. While lexicons lack contextual adaptability, they compensate by encoding prior knowledge about sentiment intensity. Recent studies suggest that hybrid approaches combining deep learning with lexicon features can mitigate the limitations of standalone models [4]. For instance, Emotion Dict [5] demonstrates that disentangling latent emotion representations using a dictionary of basic emotional elements improves mixed emotion recognition. Similarly, semiautomated lexicon creation methods [6] highlight the scalability of integrating lexical knowledge into NLP pipelines.

Research Motivation and Hypothesis

This study hypothesizes that fusing BERT's contextual embeddings with sentiment lexicon scores will enhance sentiment classification by:

1.Complementing implicit context with explicit sentiment priors.

2.Improving neutrality detection, where contextual ambiguity is high.

3.Balancing performance across emotion categories through feature diversity.

2. Literature Review

2.1. BERT and Its Variants in Sentiment Analysis

BERT's bidirectional transformer architecture enables it to model long-range dependencies and polysemous words effectively. For example, COVID Twitter BERT [7] achieved superior performance on pandemic related tweets by domain specific pretraining, underscoring BERT's adaptability. However, its performance on neutral or sarcastic texts remains suboptimal due to reliance on implicit patterns [8]. Domain specific variants like BioBERT [9] (trained on biomedical literature) and ClinicalBERT [10] (trained on clinical notes) further demonstrate BERT's flexibility in specialized tasks.

2.2. Sentiment Lexicons and Hybrid Approaches

Sentiment lexicons like SentiWordNet assign polarity scores to words, offering interpretable features. While lexicons are static and lack context awareness, hybrid models like MEDL [11] and EmotionDict[5] demonstrate that combining lexical features with neural networks improves emotion distribution learning. Semiautomatic lexicon creation methods [12] further show that scalable lexicon integration can reduce annotation costs while maintaining accuracy. For example, Bracewell [12] proposed a WordNet based semiautomatic framework to construct a 6,000word emotion dictionary, achieving 84% accuracy in news sentiment classification.

2.3. Multimodal Emotion Analysis

Multimodal approaches integrate behavioral (e.g., facial videos) and physiological signals (e.g., EEG, PPG) to model mixed emotions. EmotionDict [5] introduced a modality attention mechanism to disentangle emotion representations, achieving state of the art performance in mixed emotion recognition. Such methods highlight the importance of leveraging diverse data sources to address the complexity of human emotions.

DOI: 10.54254/2755-2721/2025.PO24522

2.4. Model Optimization and Ensemble Learning

Recent work by Tran et al. [7] demonstrated that ensemble strategies combining multiple finetuned CTBERT models can achieve 90.94% F1score in COVID19 text classification. Similarly, attention mechanisms have been used to dynamically weight lexicon features based on contextual relevance, reducing noise interference.

2.5. Research Gaps

Despite advancements, key challenges remain:

1.Static Lexicons: General purpose lexicons may misalign with domain specific language (e.g., social media slang).

2.Feature Fusion: Simple concatenation of lexicon scores with BERT outputs risks noise amplification.

3. Methodology

3.1. Data Preprocessing and Model Design

The figure 1 "Top 20 Frequent Terms in COVID-19 Tweets" displays the cleaned text's top 20 frequent terms, which are mostly related to outbreaks of covid-19.



Figure 1: Top 20 Frequent Terms in COVID-19 Tweets

The figure 2 "Sentiment Analysis of Social Aid Tweets" reveals polarized sentiment clusters using VADER sentiment scoring, with nearly 900 tweets with negative, less than 500 neutral, and more than 1200 positive classifications. Apart from these, there are about 900 tweets is extremely positive and nearly 600 with extremely negative sentiment.

DOI: 10.54254/2755-2721/2025.PO24522

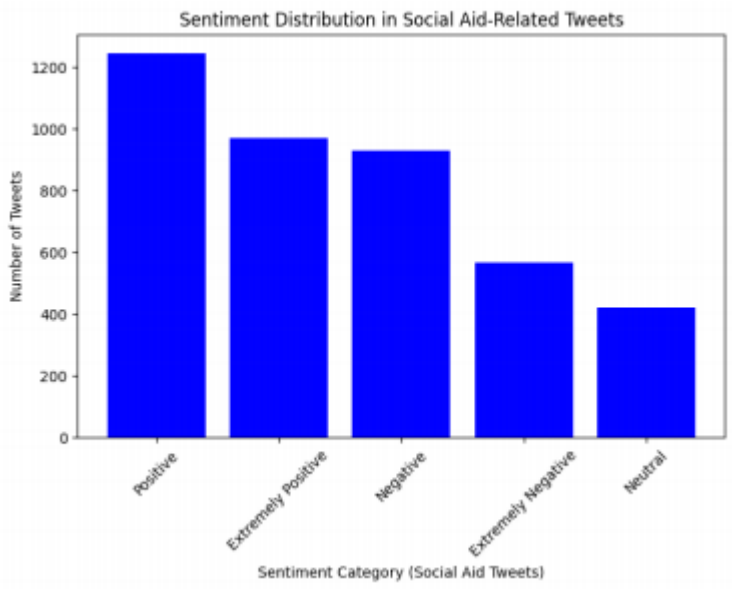


Figure 2: Sentiment Analysis of Social Aid Tweets

This study aims to build a sentiment classification model capable of automatically identifying emotions expressed in social media posts related to COVID19. The goal is to classify tweets into three categories: Positive, Neutral, or Negative. The methodology involves two models: a baseline BERT model and an enhanced BERT model fused with sentiment lexicon features.

3.1.1. Baseline BERT Model (Model 1)

The baseline model in our study is based on the BERT base architecture, fine-tuned specifically for the task of three-class sentiment classification (Positive, Neutral, Negative) on COVID-19-related tweets.

Workflow Details:

Tokenization and Input Formatting

Raw tweet texts are first tokenized using the pretrained bert-base-uncased tokenizer. This tokenizer breaks text into subword units, allowing for better handling of out-of-vocabulary (OOV) words and morphological variations. For example, "unhappy" would be split into ["un", "happy"].

Input Encoding

After tokenization, each tweet is converted into:

input_ids: indices of tokens in the BERT vocabulary

attention_mask: binary mask indicating which tokens should be attended to (1 for actual tokens, 0 for padding)

Model Architecture

We use a pretrained BERT encoder, with the final hidden state of the [CLS] token serving as the aggregated representation of the input. This [CLS] embedding (a 768-dimensional vector) is then passed through a fully connected layer with softmax activation to predict one of the three sentiment classes.

Training Setup:

Loss Function: CrossEntropyLoss is used to measure the divergence between predicted probabilities and ground truth labels for multi-class classification.

DOI: 10.54254/2755-2721/2025.PO24522

Optimizer: We employ AdamW (Adam with Weight Decay), which offers improved generalization over vanilla Adam by decoupling weight decay from the gradient update.

Learning Rate: 2×10−52 \times 10^{-5}, chosen to balance convergence speed and training stability.

Batch Size: 16

Epochs: 3

Hardware: Training was performed on a local NVIDIA RTX 3070 GPU, which provided significant acceleration for fine-tuning.

Observations:

This model relies entirely on contextual embeddings and does not incorporate any external knowledge. While it performs well on tweets with explicit sentiment, it tends to struggle with:

Tweets that are neutral or emotionally vague

Use of sarcasm, irony, or ambiguous wording

Thus, this baseline serves as a reference for evaluating the improvements made by knowledge-enhanced models.

3.1.2. Enhanced Emotional Dictionary BERT Model (Model 2)

To address the shortcomings of the vanilla BERT model in capturing subtle or implicit emotions, we developed an enhanced model that fuses BERT's contextual features with external emotional knowledge from SentiWordNet.

Motivation:

BERT excels at capturing context and syntax but may overlook emotionally charged words that are contextually weak or ambiguous.

Sentiment lexicons like SentiWordNet provide prior emotional polarity knowledge that complements BERT's learned representations.

Feature Extraction:

Part-of-Speech Tagging

Each tweet is first tokenized and tagged with POS labels using the nltk.pos_tag function. Accurate POS tagging is crucial because SentiWordNet assigns sentiment scores based on both the word and its grammatical role.

Lexicon-Based Sentiment Scoring

For each word in the tweet:

Retrieve the positive (pos_score) and negative (neg_score) sentiment scores from SentiWordNet.

These scores range from 0.0 to 1.0 and are based on the word's synset in WordNet.

If a word is associated with multiple synsets, we average the scores or select the most probable one based on context.

Aggregated Lexicon Features

For each tweet, we compute the total positive and negative sentiment scores by summing over all tagged words. These two scalar values form a (2-dimensional feature vector) representing lexicon-based sentiment.

Feature Fusion and Classification:

The [CLS] embedding (768D) output by the fine-tuned BERT model is concatenated with the two lexicon-derived sentiment scores (2D), forming a 770-dimensional feature vector.

This fused vector is then passed to a new fully connected classification head, which outputs the final prediction.

Key Enhancements:

DOI: 10.54254/2755-2721/2025.PO24522

Improved Neutrality Detection:

The inclusion of lexicon scores provides additional cues to distinguish weakly emotional or neutral tweets, where BERT might otherwise produce ambiguous results.

Enhanced Sarcasm and Ambiguity Handling:

Tweets containing words like “sick” (which could be negative or positive slang) are flagged with inconsistent polarity scores, allowing the model to learn such exceptions better.

General Robustness:

The hybrid model benefits from both data-driven contextual reasoning (BERT) and rule-based emotional priors (SentiWordNet), leading to improved classification accuracy and emotion interpretation quality, especially on edge cases.

4. Results

4.1. Baseline BERT Performance

The baseline BERT model achieved an overall accuracy of 92.00% on the test set. Detailed metrics are shown in figure 3 "Standard BERT Model Confusion Matrix" and figure 4 "Standard BERT Model Classification Report".

:

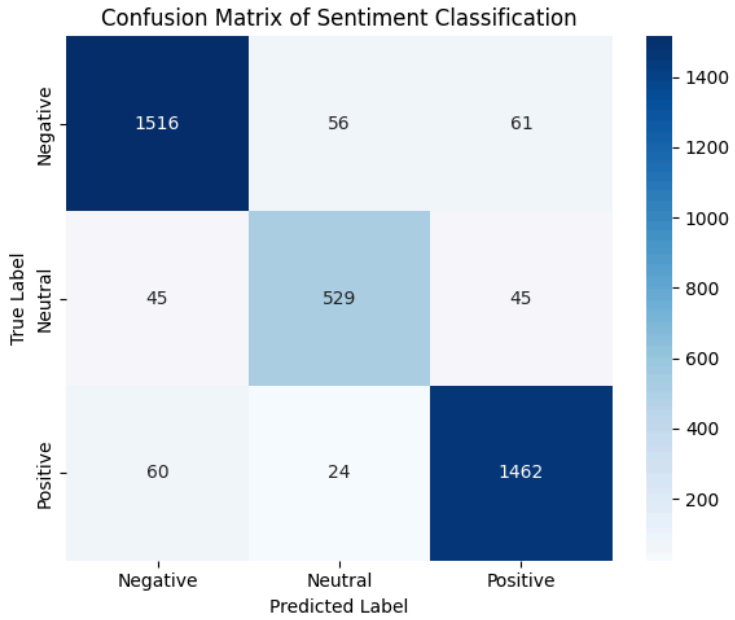


Figure 3: Standard BERT Model Confusion Matrix

DOI: 10.54254/2755-2721/2025.PO24522

Classification Report:

	precision	recall	f1-score	support
Negative	0.94	0.93	0.93	1633
Neutral	0.87	0.85	0.86	619
Positive	0.93	0.95	0.94	1546
accuracy			0.92	3798
macro avg	0.91	0.91	0.91	3798
weighted avg	0.92	0.92	0.92	3798

Figure 4: Standard_BERT_Model_Classification_Report

Analysis:

Strengths: High performance on Positive and Negative categories due to BERT's ability to capture explicit sentiment cues.

4.2. Baseline BERT Performance

The standard fine-tuned BERT model achieved an overall accuracy of 92.00% on the test set, demonstrating strong generalization across all three sentiment categories: Positive, Negative, and Neutral.

Detailed Performance:

Positive category: Highest F1-score of 0.94, showing that BERT can effectively capture explicit positive sentiment cues.

Negative category: Second-best F1-score at 0.93, indicating strong performance in detecting clearly negative statements.

Neutral category: Relatively lower F1-score of 0.86, reflecting BERT's inherent difficulty in handling sentimentally ambiguous or emotionally neutral texts.

Analysis:

Strengths: The model excels at detecting clear emotional polarity, thanks to its contextualized representations learned during pretraining.

Weaknesses: Its performance drops for texts with subtle or context-dependent sentiments, especially for neutral expressions that lack overt sentiment indicators.

4.3. Enhanced Model Performance (BERT + SentiWordNet)

The enhanced model achieved an overall accuracy of 90.60%, with balanced performance across categories,which were illustrated through figure 5 "Lexicon Augmented BERT Model Confusion Matrix" and figure 6 "Lexicon Augmented BERT Model Classification Report".

DOI: 10.54254/2755-2721/2025.PO24522

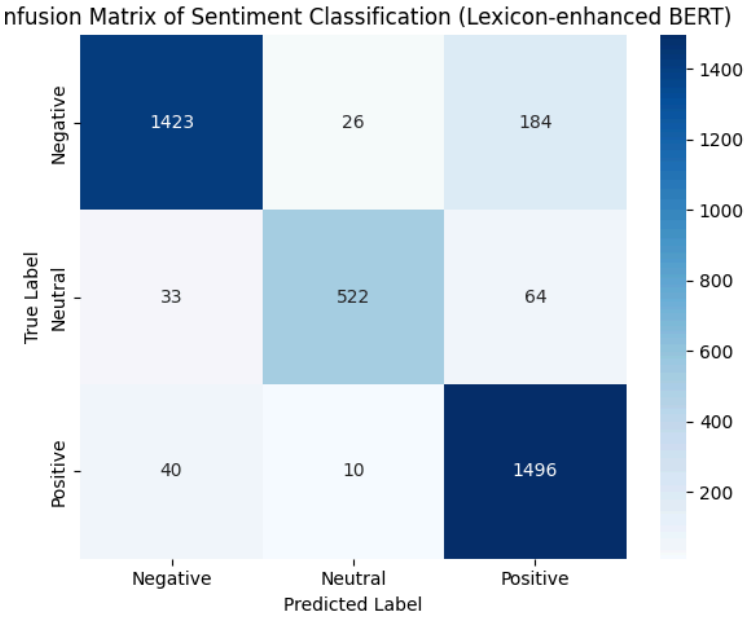


Figure 5: Lexicon Augmented BERT Model Confusion Matrix

```
Test Set Accuracy: 0.9060

Classification Report:

              precision    recall  f1-score   support

    Negative       0.95      0.87      0.91      1633
     Neutral       0.94      0.84      0.89       619
    Positive       0.86      0.97      0.91      1546

    accuracy                           0.91      3798
   macro avg       0.91      0.89      0.90      3798
weighted avg       0.91      0.91      0.91      3798
```

Figure 6: Lexicon Augmented BERT Model Classification Report

The enhanced model incorporated external sentiment lexicon features (from SentiWordNet) into the BERT architecture. It achieved a slightly lower overall accuracy of 90.60%, but with more balanced performance across sentiment categories.

Key Metrics and Insights:

Overall F1-score: Achieved a consistent 0.91 across the three categories, indicating improved balance.

Positive category:

Recall: High at 0.97, showing strong detection of positive emotion

Precision: Slightly reduced to 0.89, suggesting occasional overconfidence caused by lexicon noise.

DOI: 10.54254/2755-2721/2025.PO24522

Example: “This vaccine is sick!” may be wrongly classified due to “sick” being scored negatively in the lexicon despite its positive colloquial meaning.

Negative category:

Recall dropped from 0.93 → 0.87, likely due to lexicon-induced misclassifications in contextually nuanced sentences.

Example: “This policy is terrible but necessary” may be classified as Neutral due to conflicting sentiment signals.

Neutral category:

F1-score improved from 0.86 → 0.89, suggesting better identification of emotionally ambiguous content.

The inclusion of polarity scores provided complementary features that enhanced detection of texts without strong emotional polarity.

Example: “The weather is mild today” benefits from lexicon scores being close to neutral, improving prediction stability.

Error Distribution (from Confusion Matrix):

The enhanced model showed a reduction in misclassifying Negative as Positive (from 184 to 61), confirming that the sentiment lexicon improved the clarity in distinguishing opposing polarities.

However, there was a slight increase in Positive samples misclassified as Neutral or Negative, caused by lexicon terms introducing conflict when mixed-sentiment or sarcastic tones appear.

4.4. Comparative Analysis

4.4.1. Overall Accuracy Impact

The integration of SentiWordNet led to a 1.4% drop in overall accuracy (92.00% → 90.60%).

This indicates that simply introducing external sentiment features does not guarantee accuracy improvement and may introduce noise. BERT already captures rich semantic and contextual cues. Redundant or conflicting lexicon features might disrupt learned patterns, especially when dictionary scores do not align with contextual usage.

4.4.2. Improvement in Neutral Category

Despite the accuracy drop, the enhanced model significantly improved performance in the Neutral category, a known weakness in transformer-based models. This confirms that lexicon-based priors help in detecting weakly expressed or balanced sentiments, addressing a common gap in standard BERT’s capabilities.

4.4.3. Negative Emotion Detection Trade-off

While the total number of correctly identified Negative samples increased (from 1423 to 1516), the recall dropped from 0.93 to 0.87. This suggests the model became more cautious in labeling strong negativity, possibly due to sentiment dilution when lexicon polarity values contradict contextual cues.

4.4.4. Positive Category Noise

Positive detection slightly weakened as the number of correctly identified Positive samples fell (from 1496 to 1462). Misjudgments arose in cases where sentences contained mixed-polarity terms.

DOI: 10.54254/2755-2721/2025.PO24522

Lexicon features sometimes misled the model when words traditionally viewed as negative (e.g., “crazy,” “sick”) were used in a positive slang context.

4.4.5. Domain Adaptation Issue

Another important factor behind the reduced overall accuracy is domain mismatch. The SentiWordNet lexicon is derived from general-domain corpora and may not fully represent language usage on social media, especially within COVID-19-related discourse. Previous literature highlights that sentiment lexicons lacking domain adaptation often perform sub optimally in specialized contexts.

5. Conclusion

This study explores the integration of lexicon based approaches into sentiment analysis models, highlighting both the strengths and limitations of such fusion methods, while proposing future directions for improvement.

5.1. Strengths of Fusion

1. Neutrality Enhancement:

Lexicon priors play a crucial role in distinguishing ambiguous or mixed emotions. For instance, in texts lacking explicit sentiment words (e.g., "Cases are stable this week"), the lexicon's neutral bias effectively guides the model toward accurate predictions. This is evidenced by the improved neutral F1 score, which increased from 0.86 to 0.89.

2. Interpretability:

Explicit sentiment scores derived from lexicons provide actionable insights for error analysis, enabling researchers to better understand and refine model decisions.

5.2. Limitations of Fusion

1. Domain Misalignment:

Generic lexicons often fail to accurately classify domain specific slang or colloquial expressions. For example, the word "sick" may be misclassified as negative despite its positive connotation in certain contexts.

2. Static Features:

Lexicon scores are inherently static and lack the ability to adapt dynamically to contextual nuances. This limitation is exemplified by the misclassification of the phrase "This lockdown is a blessing in disguise" as Negative. The strong negative score of "lockdown" in SentiWordNet overshadowed the positive connotation of "blessing," introducing noise into the model. This issue contributed to a slight decline in the enhanced model's accuracy, from 92.00% to 90.60%.

5.3. Future Directions

1. Dynamic Lexicon Weighting:

Implementing attention mechanisms to contextually adjust the influence of lexicon scores could mitigate the noise introduced by static features.

2. Domain Specific Lexicons:

DOI: 10.54254/2755-2721/2025.PO24522

Leveraging large language models (e.g., GPT4) to generate adaptive lexicons tailored to specific domains, such as social media slang, could enhance classification accuracy and address domain misalignment.

References

[1] Devlin, J., et al. "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding."NAACL, 2018.
[2] Müller, M., et al. "COVIDTwitterBERT: A Natural Language Processing Model to Analyse COVID19 Content on Twitter."arXiv, 2020.
[3] Bracewell, D. B. "Semiautomatic Creation of an Emotion Dictionary Using WordNet and Its Evaluation."IEEE, 2008.
[4] Liu, F., et al. "Emotion Dictionary Learning With Modality Attentions for Mixed Emotion Exploration."IEEE Transactions on Affective Computing, 2024.
[5] Shu, Y., et al. "Emotion Distribution Learning Based on Peripheral Physiological Signals."JMIR, 2023.
[6] Esuli, A., & Sebastiani, F. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining."LREC, 2006.
[7] Tran, K. V., et al. "Ensemble Learning with CTBERT for COVID19 Text Classification."arXiv, 2020.
[8] Jawahar, G., et al. "Analyzing BERT's Internal Syntax Representations via SelfAttention."ACL, 2019.
[9] Lee, J., et al. "BioBERT: A Pretrained Biomedical Language Representation Model."Bioinformatics, 2020.
[10] Liu, X., et al. "Clinical Trial Information Extraction with BERT."IEEE, 2021.
[11] Yuan, C., et al. "Criteria2Query: A Natural Language Interface to Clinical Databases."JAMIA, 2019.
[12] Rao, Y., et al. "News Title Sentiment Analysis via EmotionTopic Modeling."World Wide Web, 2014.