

SentiLab: Comparative Evaluation of Classical, Sequential, and Transformer-Based Models for Three-Class Twitter Sentiment Analysis

Chandramani Srivastava
chandramani.sri@gmail.com

Dr. Pankaj Kr. Sharma
pankaj.sharma@abes.ac.in

Abstract- Sentiment analysis has become one of the most important applications of Natural Language Processing because social media generates a continuous stream of short, noisy, and opinion-rich text. This paper presents SentiLab, a comparative framework for three-class Twitter sentiment classification using Positive, Neutral, and Negative labels. The study evaluates three representative model families: TF-IDF with Logistic Regression, Bidirectional LSTM with Multi-Head Self-Attention, and fine-tuned DistilBERT. The original Sentiment140 dataset was extended into a balanced three-class dataset using confidence-based relabeling so that emotionally ambiguous tweets could be modeled as Neutral rather than being forced into binary polarity. The results show that DistilBERT achieved the best overall performance with 94.7% accuracy and a macro AUC-ROC of 0.988, followed by BiLSTM+Attention at 92.4% and TF-IDF+Logistic Regression at 76.3%. In addition to model comparison, the framework was deployed as a browser-based real-time sentiment analysis system using ONNX Runtime and Transformers.js. The paper also discusses the trade-off between classification accuracy, inference latency, and deployability in practical sentiment systems.

Keywords- Sentiment analysis, DistilBERT, BiLSTM, transformer models, TF-IDF, Logistic Regression, Twitter analytics, Sentiment140.

I. INTRODUCTION

Social media platforms have transformed the way public opinion is expressed and consumed. Twitter posts, product reviews, and discussion forums now serve as a large-scale source of emotional and factual content that can be mined for decision support. Sentiment analysis, also known as opinion mining, attempts to determine the polarity and emotional tone of text. In practice, however, social media text is difficult to model because it is short, informal, domain-dependent, and often filled with abbreviations, emojis, sarcasm, and mixed sentiment. A system that only predicts positive or negative labels is insufficient for many real-world applications, since a substantial amount of social content is neutral, objective, or unclear.

The thesis underlying this work identifies this gap directly and frames the core problem as a comparison of traditional TF-IDF-based methods, deep sequential architectures, and transformer

networks for three-class tweet classification. It also emphasizes that deployment matters as much as accuracy: a model that performs well offline may still be impractical if it is too large, too slow, or too difficult to run in a browser or edge environment. The thesis states that SentiLab was designed to compare these model families and to make real-time prediction available directly in the browser.

II. RELATED WORK

Early sentiment analysis approaches relied on lexicon-driven scoring systems and classical machine learning classifiers. Techniques such as Naive Bayes, Support Vector Machines, and Logistic Regression were typically paired with bag-of-words or TF-IDF features. These methods remain attractive because they are fast and interpretable, but they cannot capture deeper meaning or long-range dependencies.

Deep learning improved text representation through CNNs, RNNs, LSTMs, and BiLSTMs, while attention mechanisms allowed models to focus on sentiment-bearing words. Transformer architectures later replaced recurrence with self-attention and dramatically improved contextual modeling. BERT introduced bidirectional pretraining, and DistilBERT reduced the computational footprint while preserving strong performance.

III. DATASET AND PREPROCESSING

The primary dataset used in the thesis is Sentiment140, a widely used Twitter corpus containing 1.6 million tweets collected with distant supervision. Positive emoticons were used to label positive tweets, while negative emoticons were used to label negative tweets. The thesis reports that the original dataset contains 800,000 positive and 800,000 negative samples.

To make the benchmark more realistic, the binary dataset was extended into a three-class dataset by creating a Neutral category. A baseline Logistic Regression classifier was first trained on the binary labels, and tweets with uncertain predicted probability values between 0.35 and 0.65 were relabeled as Neutral. This produced a more practical three-way split that better reflects real social media usage. After balancing, the dataset contained approximately 960,000 Positive tweets, 480,000 Neutral tweets, and 960,000 Negative tweets.

The preprocessing pipeline handled Twitter-specific noise through a multi-stage workflow: URL removal, mention removal, normalization, contraction expansion, emoji

conversion to descriptive tokens, TweetTokenizer-based tokenization, and stopword filtering with negation words preserved. The thesis specifically notes that negation tokens such as not, never, and nor were retained because they directly affect polarity.

Table I. Extended Dataset Split

Split	Positive	Neutral	Negative
Training	672,000	336,000	672,000
Validation	144,000	72,000	144,000
Test	144,000	72,000	144,000



Fig. 1. Multi-stage tweet processing pipeline.

IV. METHODOLOGY

SentiLab compares three architectures that represent major paradigms in sentiment classification. The first is a classical baseline using TF-IDF vectorization and Logistic Regression. The second is a sequential deep learning model that combines BiLSTM with Multi-Head Self-Attention. The third is a transformer-based model obtained by fine-tuning DistilBERT.

The TF-IDF baseline uses unigram and bigram features, a maximum feature cap of 20,000, a minimum document frequency of 5, and sublinear term-frequency scaling. Logistic Regression is regularized using L2 penalty because it handles high-dimensional sparse features efficiently. The BiLSTM+Attention model uses 100-dimensional GloVe Twitter embeddings, 128 hidden units per direction, eight attention heads, a dense ReLU layer, and dropout for regularization. Attention improves the model's ability to focus on sentiment-bearing phrases, negations, and modifiers. DistilBERT is used as the strongest contextual model because it retains most of BERT's performance while using fewer parameters and less compute.

Table II. Model Configuration Summary

Model	Key Settings	Main Advantage
TF-IDF + LR	20k features, (1,2) n-grams, L2 regularization	Fast and interpretable
BiLSTM + Attn	GloVe Twitter, 128 hidden units, 8 heads	Captures sequence context
DistilBERT	Fine-tuned transformer, softmax classifier	Best contextual accuracy

V. EXPERIMENTAL RESULTS

The experiments were conducted on an NVIDIA Tesla T4 GPU with 32 GB RAM using TensorFlow 2.12 and HuggingFace Transformers 4.30. DistilBERT was fine-tuned for 3 epochs, while the BiLSTM model was trained for 10 epochs with early stopping. The thesis reports the following overall performance hierarchy: DistilBERT at 94.7% accuracy, BiLSTM+Attention at 92.4%, TF-IDF+Logistic Regression at 76.3%, and Naive Bayes at 71.2%. Neutral sentiment remained the most difficult class across all models. DistilBERT produced the strongest Neutral F1-score at 91.3%, while the TF-IDF baseline dropped to 61.2%. This indicates that contextual language understanding is especially important when the text is ambiguous or weakly polar. The thesis also reports macro AUC-ROC values of 0.847 for TF-IDF+LR, 0.971 for BiLSTM+Attention, and 0.988 for DistilBERT.

Table III. Comparative Performance on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	Macro AUC-ROC
Naive Bayes	71.2%	70.5%	69.8%	70.1%	-
TF-IDF + LR	76.3%	75.8%	74.9%	75.3%	0.847
BiLSTM + Attention	92.4%	91.8%	92.1%	91.9%	0.971
DistilBERT	94.7%	94.3%	94.5%	94.4%	0.988

Table IV. Browser Deployment Efficiency

Model	Approx. Size	Latency	Deployment Mode
TF-IDF + LR	~2.1 MB	<5 ms	Browser JS
BiLSTM + Attention	~12 MB	~45 ms	ONNX Runtime
DistilBERT	~67 MB	~280 ms	Transformers.js / WebGPU

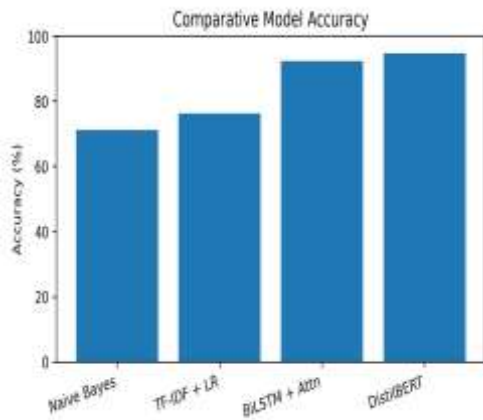


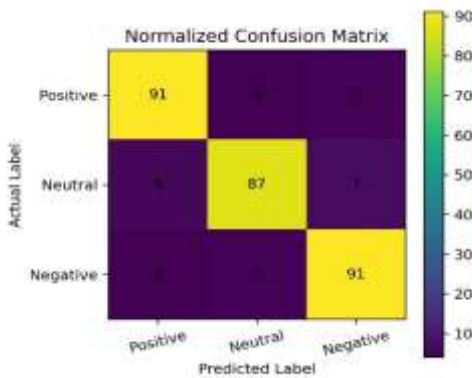
Fig. 2. Comparative model accuracy.

Fig. 3. Normalized confusion matrix.

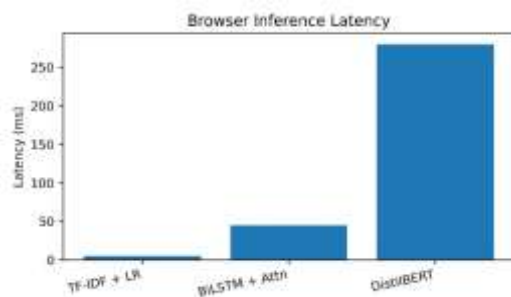
Fig. 4. Browser inference latency comparison.

VI. DISCUSSION

The results show a clear advantage for transformer-based sentiment analysis. TF-IDF+Logistic Regression is fast and easy to deploy, but sparse vector representations cannot fully capture context, sarcasm, or semantic ambiguity.



BiLSTM+Attention improves performance by modeling sequence order and focusing on important words, but it still



depends on embedding coverage and does not match the contextual richness of transformers. DistilBERT gives the best overall performance because its bidirectional transformer encoder learns deeper sentence-level relationships. The thesis further explains that the Neutral class is the most problematic

category and that false polarization is a frequent failure mode for the classical baseline. It also notes that DistilBERT performs best but requires greater latency and memory than the lightweight models, making deployment context an important part of model selection.

VII. WEB APPLICATION DEPLOYMENT

SentiLab was deployed as a browser-based single-page application using HTML5, CSS3, vanilla JavaScript, Chart.js, ONNX Runtime, and Transformers.js. The application supports three inference modes: a lightweight TF-IDF mode, an ONNX-based BiLSTM mode, and a DistilBERT mode for higher accuracy. This design removes the need for a backend inference server and demonstrates the feasibility of real-time browser-side sentiment analysis. The thesis describes the deployment layer as a modular frontend architecture intended for practical real-time use.

VIII. CONCLUSION

The study shows that DistilBERT achieves the strongest overall performance for three-class Twitter sentiment analysis, BiLSTM+Attention offers a strong middle ground, and TF-IDF+Logistic Regression remains a useful low-cost baseline. Future work may include multilingual sentiment analysis, aspect-based sentiment classification, quantized transformer deployment, and streaming analysis of live social media data.

REFERENCES

- [1] A. Go, R. Bhayani, and L. Huang, 'Twitter Sentiment Classification using Distant Supervision,' Stanford University, 2009.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,' Proc. NAACL-HLT, 2019.
- [3] V. Sanh et al., 'DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,' arXiv preprint arXiv:1910.01108, 2019.
- [4] A. Vaswani et al., 'Attention Is All You Need,' Proc. NIPS, 2017.
- [5] Y. Kim, 'Convolutional Neural Networks for Sentence Classification,' Proc. EMNLP, 2014.
- [6] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory,' Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
- [7] D. Bahdanau, K. Cho, and Y. Bengio, 'Neural Machine Translation by Jointly Learning to Align and Translate,' arXiv preprint arXiv:1409.0473, 2015.
- [8] J. Pennington, R. Socher, and C. Manning, 'GloVe: Global Vectors for Word Representation,' Proc. EMNLP, 2014.
- [9] C. Hutto and E. Gilbert, 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,' Proc. ICWSM, 2014.
- [10] M. Hu and B. Liu, 'Mining and Summarizing Customer Reviews,' Proc. ACM SIGKDD, 2004.