Twitter Sentiment Analysis with Traditional and Transformer Models

Crescent Liew Qi En, Koa Le Yee, Shavonne Lim Jing Ning, Vivekanantham Bhuvaneswari Akshayaa

Introduction

Twitter sentiment analysis (SA) is a challenging task in natural language processing (NLP) due to tweets' short, informal, and often noisy nature. Regardless, SA remains valuable in applications such as public opinion monitoring and product feedback analysis.

Recent studies have examined both traditional ML and transformer-based deep learning (DL) models for SA. Qi and Shabrina (2023) found that traditional ML models struggled with class imbalance and informal language. Meanwhile, transformer-based DL models like DistilBERT have achieved up to 93.5% accuracy, surpassing traditional ML models (Singh & Kumar, 2023). However, prior studies tend to rely on pretrained models and overlook how different preprocessing strategies, particularly duplicate removal, affect model performance. Given that input noise and model complexity can affect robustness, these gaps warrant further analysis.

In addressing these gaps, this project framed SA as a binary supervised task: classifying whether a tweet is positive or negative. It benchmarked four widely used ML models – Logistic Regression (LR), Maximum Entropy (MaxEnt), Random Forest (RF), and Stochastic Gradient Descent (SGD) Classifier – against Pretrained and Custom DistilBERT variants. Each model was assessed with varying preprocessing depths, focusing on the effect of duplicate removal, which affects approximately 0.91% of the dataset. Although minimal, such redundancy can reinforce sentiment patterns during training, especially for context-sensitive models like transformers. This setup enabled a comparative analysis of how small-scale redundancy and model architecture influenced classification performance. Models were evaluated using F1-score and accuracy, rather than misclassification costs.

Dataset

The Twitter Sentiment Dataset consists of 100,000 English-language tweets labeled for binary sentiment (1 for positive, 0 for negative), balanced across both classes with no missing values and stored in CSV format. Unlike formal text, tweets often contain slang, abbreviations, hashtags, and informal expressions, presenting unique challenges for SA.

Despite being structurally clean, tweets' informal traits necessitate tailored preprocessing strategies. To guide preprocessing choices, exploratory analyses were conducted to examine the dataset's structural and lexical characteristics.

Exploratory Data Analysis (EDA)

To evaluate compatibility with transformer models, Figure 1 reveals that most tweets contain 10 to 40 tokens, with a long tail exceeding 100. Based on this, the input length was capped at 128 tokens to balance coverage and efficiency.

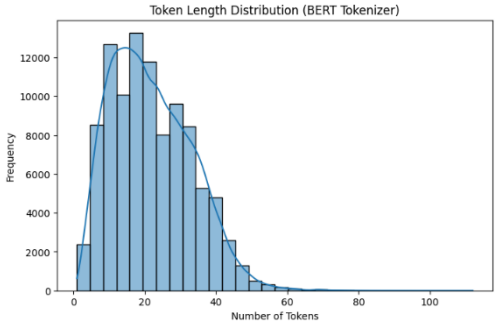

Figure 1: Token Length Distribution

Figures 2 and 3 display word clouds that show distinct sentiment cues – for example, *thank* and *love* for positives, and *work* and *sad* for negatives. Also, non-words like *quot* and *amp*, originating from HTML entities, appear frequently. This indicates the need for entity decoding, which otherwise fragments meaning during subword tokenization.

Figure 2: Word Cloud – Positive Tweets

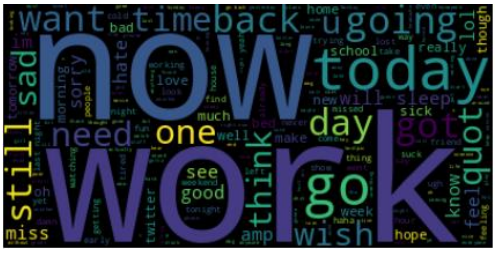

Figure 3: Word Cloud – Negative Tweets

Preprocessing Steps

Guided by insights from EDA and supporting literature, separate preprocessing pipelines for traditional ML models and DistilBERT were developed, as shown in Figure 4.

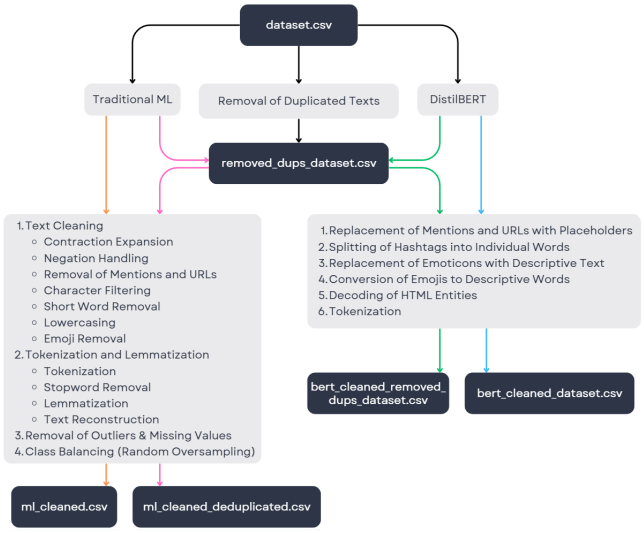

Figure 4: Overview of Preprocessing Pipelines for Traditional ML Models and DistilBERT

The ML pipeline performed standard text cleaning, handled outliers and missing values, and addressed class imbalance via random oversampling. In contrast, the DistilBERT pipeline applied a more targeted preprocessing approach, including normalization of mentions and URLs into placeholders, and transformation of hashtags, emojis, and emoticons into semantically meaningful tokens, which were shown to improve model performance by reducing textual noise and revealing sentiment-rich features (Pota et al., 2021). To evaluate the effect of redundancy, two dataset variants – one with and another without duplicate tweets – were prepared for both ML and DistilBERT models.

Methods

This project investigates how preprocessing depths – specifically, duplicate removal – and model architecture affect sentiment classification performance on tweets.

Method Pipeline

To ensure fair and consistent evaluation, all models followed a standardized pipeline (see Figure 5). After preprocessing and feature extraction, the dataset was split using a stratified 80-20 train-test split to preserve class balance. Baseline performance was assessed using 5-fold cross-validation on the training set, followed by final evaluation on the held-out test set. Hyperparameter tuning was conducted via cross-validation, and the best configuration was used for final testing. F1-score was used as the primary metric due to its ability to balance precision and recall in the absence of domain-specific misclassification costs, while accuracy served as a secondary reference.

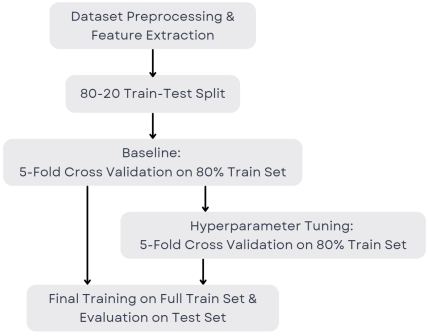

Figure 5: Overview of Method Pipeline

Traditional ML Models

To establish strong non-contextual baselines, four ML models were selected: LR, MaxEnt, RF, and SGD Classifier. They are computationally efficient, and well-suited for high-dimensional, sparse text representations such as n-grams and Term Frequency-Inverse Document Frequency (TF-IDF) features.

- **LR** estimates class probabilities using weighted features and performs reliably on sparse, high-dimensional text data (Abia & Johnson, 2024). It is computationally efficient and serves as a strong baseline for SA.
- **MaxEnt** extends LR by maximizing entropy to model the most uniform class distribution consistent with observed features. It captures interactions between correlated

features, which is particularly useful for noisy, short-form text (Atillo et al., 2023). Prior work has shown that it consistently outperforms ML models like Naïve Bayes and Support Vector Machine in SA (Wang et al., 2010).

- **RF** is an ensemble method that constructs multiple decision trees on bootstrapped data subsets and aggregates predictions. It helps reduce overfitting and improves generalization, especially for datasets with non-linear relationships such as informal tweets (Salman et al., 2024; Srivani et al., 2025).
- **SGD Classifier** trains linear models using SGD, updating weights per sample. This enables fast and memory-efficient training, making it well-suited for large-scale text classification tasks.

These models require explicit feature engineering and cannot be trained directly on raw tweets, which lack the structured representations necessary for vectorization. Thus, they were trained on the *cleaned* and *cleaned + deduplicated* datasets (see Figure 4, orange and pink pipelines), where appropriate text normalization and feature extraction were applied. Hyperparameter tuning was conducted using GridSearchCV for LR, MaxEnt, and SGD, while RandomizedSearchCV was used for RF to efficiently explore a larger parameter space.

Together, they formed a comparative baseline for evaluating the effects of preprocessing and redundancy against complex transformer-based models like DistilBERT.

Transformer-Based DL Model (DistilBERT)

Transformer-based models like DistilBERT are designed to capture contextual word representations, making them highly effective for informal text where semantics may shift based on surrounding words. While BERT was initially considered, DistilBERT was chosen for its computational efficiency – retaining 97% of BERT's accuracy while being 60% faster and 40% smaller (Sanh et al., 2019).

This project evaluates two DistilBERT variants:

- Pretrained DistilBERT: A baseline model used without freezing any transformer layers.
- Custom DistilBERT: A partially fine-tuned variant where only the final two transformer layers and a custom classification head were updated during training. This strategy, inspired by Ingle et al. (2022), has been shown to improve training stability and efficiency in few-shot learning contexts on BERT-like architectures.

To isolate preprocessing effects, both DistilBERT variants were tested on three dataset variants: *raw* (tokenization only), *cleaned* (see Figure 4, blue pipeline), and *cleaned + deduplicated* (see Figure 4, green pipeline).

To optimize model performance, hyperparameter tuning was conducted using Optuna, a flexible and efficient framework for automated optimization (Akiba et al., 2019). Its support for Bayesian optimization and pruning of underperforming trials makes it well-suited for computationally intensive transformer models. Both variants were tuned for training hyperparameters such as learning rate, batch size, and warm-up steps, while Custom DistilBERT further explored architectural hyperparameters like hidden layer size and dropout rate.

To balance reliability and computational cost, each Optuna trial was evaluated using the first fold of a 5-fold cross-validation split. The best-performing configuration from each search was then retrained on the full training set and evaluated on the held-out test set.

This controlled setup enabled a focused comparison between the two DistilBERT variants under varying preprocessing depths. By isolating the effects of both architectural complexity and input redundancy, the project assessed whether performance gains stem primarily from model design, data preparation, or both.

Results & Discussions

This section presents the performance of traditional ML models and DistilBERT variants on sentiment classification, focusing on the impact of preprocessing depth, particularly duplicate removal, and model architecture.

Baseline Performance

Baseline results are summarized in Table 1.

Dataset	Raw		Cleaned		Cleaned + Deduplicated	
Model	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**
LR	-	-	0.7363	0.7366	0.7506	0.7509
MaxEnt	-	-	0.7350	0.7352	0.7500	0.7503
RF	-	-	0.7450	0.7400	0.7600	0.7602
SGD Classifier	-	-	0.6904	0.6892	0.7003	0.7044
Pretrained DistilBERT	0.8293	0.8294	0.8299	0.8299	0.8231	0.8231
Custom DistilBERT	0.8207	0.8207	0.8197	0.8197	0.8167	0.8167

Table 1: Baseline performance of traditional ML and transformer-based models across different preprocessing depths.

Among ML models, RF achieved the best F1-score (0.7600) on the *cleaned + deduplicated* dataset, likely due to its ability to capture non-linear patterns in sparse feature representations. LR and MaxEnt followed closely, while

SGD Classifier underperformed, possibly due to sensitivity to feature variance and a more rigid optimization process.

Notably, all ML models showed improved performance with duplicate removal, suggesting that non-contextual models benefit from reduced data redundancy, as they rely heavily on surface-level lexical features and are more susceptible to overfitting repeated patterns.

Pretrained DistilBERT consistently outperformed Custom DistilBERT across all dataset variants, with its best performance observed on the *cleaned* dataset. This aligns with Pota et al. (2021), who found that preprocessing steps such as emoji and hashtag transformation enhance transformer performance. However, removing just 0.91% of duplicate tweets led to a measurable decline in performance, suggesting that even minor redundancy can reinforce sentiment signals in transformer models.

Moreover, Custom DistilBERT, though slightly lower-performing, demonstrated more stable training behavior and was less affected by deduplication, consistent with Ingle et al. (2022), who found that partial layer freezing improves model robustness. This is further supported by training loss curves in Appendix A's Figures A1 and A2, which illustrate early overfitting in Pretrained DistilBERT and greater training stability in its custom variant.

Fine-Tuned Performance

Fine-tuned results are summarized in Table 2.

Dataset	**Raw**		**Cleaned**		**Cleaned + Deduplicated**	
Model	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**
LR	-	-	0.7480	0.7481	0.7592	0.7593
MaxEnt	-	-	0.7455	0.7457	0.7552	0.7554
RF	-	-	0.7400	0.7400	0.7600	0.7600
SGD Classifier	-	-	0.7238	0.7191	0.7403	0.7392
Pretrained DistilBERT	0.8320	0.8320	0.8293	0.8293	0.8237	0.8237
Custom DistilBERT	0.8230	0.8230	0.8214	0.8215	0.8172	0.8172

Table 2: Fine-tuned performance of traditional ML and transformer-based models across different preprocessing depths

Among ML models, RF again led with an F1-score of 0.7600 on the *cleaned* + *deduplicated* dataset. However, improvements were modest, indicating that traditional approaches may be reaching a performance ceiling.

Pretrained DistilBERT achieved the highest overall F1-score (0.8320) on the *raw* dataset, but its performance steadily declined with increased preprocessing. This trend supports the earlier hypothesis that even small amounts of redundancy may reinforce sentiment signals in transformers.

Custom DistilBERT, although marginally behind its pretrained variant, demonstrated greater robustness to preprocessing changes and milder overfitting. As shown in Appendix B's Figures B1 and B2, the pretrained variant experienced steadily rising validation loss, while the custom variant maintained a more stable trajectory – though not without a slight increase at the final epoch. These trends suggest that partial fine-tuning helps mitigate overfitting, in noisy or redundant datasets, aligning with findings by Ingle et al. (2022).

Practical Implications & Societal Impact

This project shows that even subtle design decisions, such as removing less than 1% of duplicate tweets, can tangibly impact model performance. These findings highlight the importance of careful data preparation and model design in real-world SA pipelines.

While these models cannot match human nuance, they offer significant advantages in terms of speed, scalability, and consistency. For instance, a simple model like LR can classify over 100,000 tweets in under a minute, but not a human. Importantly, models need not outperform humans – they only need to be fast, reasonably accurate, and consistent; a hybrid approach, where models handle general cases and humans focus on ambiguous or high-stakes instances, can strike a balance between efficiency and accuracy.

However, the use of models on user-generated content like tweets raises several societal and ethical considerations:

- Privacy: Tweets may contain sensitive or identifiable information that requires secure handling.
- Fairness: Informal expressions and dialects can be misinterpreted by models trained on biased data, reinforcing stereotypes or marginalizing certain groups.
- Interpretability: Transformer-based models are often opaque, which limits traceability and accountability, reducing trust in their predictions.
- Job Impact: While automation may reduce manual tagging needs, it can also support professionals by providing fast, high-level sentiment insights for marketing or customer engagement.

These concerns emphasize the importance of responsible model development, where choices in preprocessing, dataset design, and post-deployment monitoring are integral to building fair and explainable systems.

Appendices

Appendix A: Training Dynamics – Pretrained vs Custom DistilBERT on *raw* dataset

To examine architectural differences beyond performance metrics, training and validation loss curves were analyzed for both Pretrained and Custom DistilBERT variants on the *raw* dataset as a baseline.

Figure A1 shows that Pretrained DistilBERT achieved strong initial learning but began to overfit early, with validation loss increasing noticeably after epoch 2. This suggests that the model was highly sensitive to noise and redundancy in the raw tweets, potentially memorizing sentiment patterns reinforced by duplicated samples.

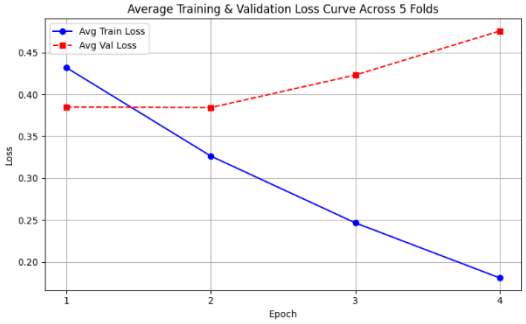

Figure A1: Training and validation loss curves – Pretrained DistilBERT (*raw* dataset)

In contrast, Figure A2 shows that Custom DistilBERT maintained a flatter and more stable validation loss trajectory. This indicates stronger generalization and training stability, likely due to reduced capacity and fewer trainable parameters. Despite slightly lower overall performance, Custom DistilBERT was more resilient to overfitting, reinforcing the benefit of partial fine-tuning in low-signal or noisy conditions.

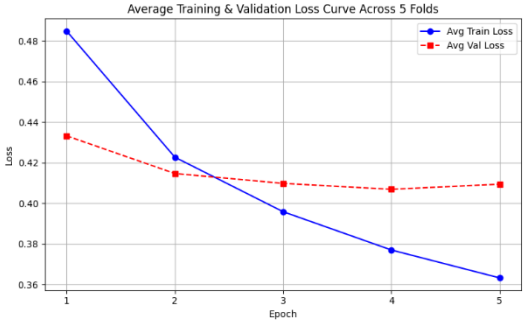

Figure A2: Training and validation loss curves – Custom DistilBERT (*raw* dataset)

Appendix B: Fine-Tuned Training Dynamics – Pretrained vs Custom DistilBERT on *raw* dataset

To better understand model behavior after fine-tuning, training and validation loss curves were analyzed for both Pretrained and Custom DistilBERT variants on the *raw* dataset as a baseline.

Figure B1 illustrates the training dynamics of Pretrained DistilBERT, which shows consistent overfitting: while training loss steadily decreases, validation loss rises continuously across epochs. This suggests that the model over-learned from the training data, including potential redundancies, at the expense of generalization.

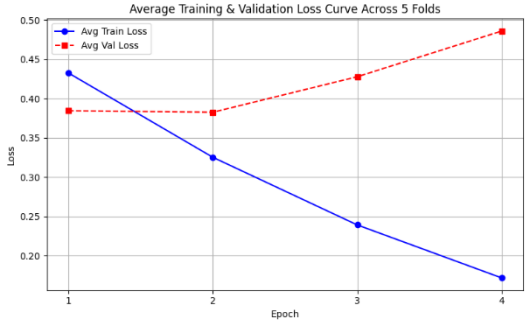

Figure B1: Training and validation loss curves – Fine-tuned Pretrained DistilBERT (*raw* dataset)

In contrast, Figure B2 shows that Custom DistilBERT exhibited milder overfitting. Validation loss decreased initially and remained relatively stable until a slight increase in the final epoch. The smaller divergence between training and validation loss curves suggests improved training stability and generalization, likely a result of its reduced model capacity and partial layer freezing.

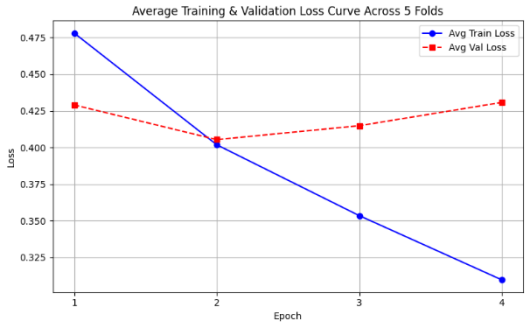

Figure B2: Training and validation loss curves – Fine-tuned Custom DistilBERT (*raw* dataset)

References

Abia, V. M., & Johnson, E. H. (2024). Sentiment Analysis techniques: A comparative study of logistic regression, random forest, and naive bayes on general English and Nigerian texts. *Journal of Engineering Research and Reports, 26*(9), 123–135. https://doi.org/10.9734/jerr/2024/v26i91268

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631. https://doi.org/10.1145/3292500.3330701

Atillo, G. N. A., Gerardo, B. D., & Medina, R. P. (2023). Sentiment analysis in product reviews with maximum entropy and naive bayes using n-gram method. *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, 522-526. https://doi.org/10.1109/icoiact59844.2023.10455843

Ingle, D., Tripathi, R., Kumar, A., Patel, K., & Vepa, J. (2022). Investigating the characteristics of a transformer in a few-shot setup: Does freezing layers in RoBERTa help? *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 238–248. https://doi.org/10.18653/v1/2022.blackboxnlp-1.19

Pota, M., Ventura, M., Fujita, H., & Esposito, M. (2021). Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications, 181*, 115119. https://doi.org/10.1016/j.eswa.2021.115119

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01030-x

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview (Trans.). *Babylonian Journal of Machine Learning, 2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108

Singh, A., & Kumar, S. (2023). A comparison of machine learning algorithms and transformer-based methods for Multiclass sentiment analysis on Twitter. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-9. https://doi.org/10.1109/icccnt56998.2023.10306507

Srivani, P., Sharma, H., Porwal, R., Nagalakshmi, T., Mercy, P., Adudhodla, M., & Parveen, N. (2025). Integrating natural language processing with AdaBoost, random forest, and logistic regression for an advanced ensemble-based network intrusion detection model. *Journal of Information Systems Engineering and Management, 10*(3s), 264-283. https://doi.org/10.52783/jisem.v10i3s.386

Wang, H., Wang, L., & Yi, L. (2010). Maximum entropy framework used in text classification. *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, 828-833. https://doi.org/10.1109/icicisys.2010.5658639