Comparative Analysis of Textual Sentiment Analysis Models: ML, DL & TL

Dimple Tiwari[1], Bobbinpreet kaur[2]
[1]Post Doctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia
[2]Department of CSE, UIE, Chandigarh University, Mohali-140413, Punjab, India
Email ID : dimple.tiwari88@gmail.com

Abstract: Sentiment Analysis is a notable NLP task for identifying and classifying opinions and attitudes present in text. With the advent of machine learning, deep learning and transfer-learning based models, several different kinds of models were developed with distinct capability of capturing lexical patterns, sequential context and contextual meaning. In this experiment, a controlled comparative study of 7 different sentiment classification models: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), DistilBERT and RoBERTa was performed. The experiment was executed with the similar dataset and experimental protocol as implemented in the given notebook. The data after pre-processing has a total of 33,113 text samples divided into 5 classes: Negative, Neutral, Positive, Very Negative, Very Positive. The input text is a concatenation of Generated Caption and Sentiment Description. A stratified train-test split of 80:20 was applied while maintaining the class distributions. TF-IDF feature were used for classical ML models and token sequences were input for DL recurrent networks and transfer-learning based Transformer tokenization. From experiments, accuracies obtained for each model are RF:87.06, LR:82.03, SVM:84.70, LSTM:88.77, BiLSTM:87.06, DistilBERT:96.35 and RoBERTa:83.03, resulting in DistilBERT having the highest accuracy.

Keywords: sentiment analysis, NLP, machine learning, deep learning, LSTM, BiLSTM, Transformer, DistilBERT, RoBERTa, text classification.

1. Introduction

Sentiment analysis is the automated extraction of opinion and affective state information from text sources, and has been applied broadly to: opinion mining, customer feedback. Social-media analytics and intelligent decision support systems[1], [2]. Classic solutions often transform input documents into sparse lexical vectors, and feed this representation into a classifier based on statistical learning methods. These approaches are still very competitive due to their effectiveness, but highly sensitive to feature representations of the input text. Deep neural nets offer an approach of learning useful feature representations automatically from token sequences, alleviating the need of manual engineering[3], [4]. LSTM was then introduced to help with learning dependencies within long sequences. Two directions can be taken in recurrence: using bidirectional model to have information about both past and future. Transformer later takes the whole NLP field and represents documents using attention-based contexts [6], which, to the surprise of some, enables successful, large-scale, pre-training and fine-tuning with BERT [7].

And then RoBERTa offers an improved pre-training task and DistilBERT offers similar performance using a compressed Transformer model, with only minor performance degradations and with a much lower computational resource requirement.. The central research question examined in this report is what are the relative performance differences in a five-class sentiment classification task: (1) when using a basic ML model, (2) a standard DL model, and (3) a pretrained transfer-learning model? A comparison with this design using the same single dataset and a same stratified test portion derived using the model result reported in the notebook, for not using results from other works and benchmark tasks.

2. Literature Survey

Previous studies of sentence sentiment Classification are developed from basic supervised learning with bag-of-words features [1] through feature-engineered and lexicon-supported techniques [2], until neural representation

learning approaches [3],[4]. Word embeddings capture word semantics into dense vectors [10], while CNN/RNN architecture shows task-specified neural features can be learned directly from raw text [11],[12]. LSTM provides a good modeling approach for sequential information [5] whereas attention mechanism and transformer enhance modeling the contextual information [6]. Pretrained bidirectional representations are generated by BERT to facilitate downstream tasks [7], with an improved strategy for pre-training introduced by RoBERTa [8]. DistilBERT shows that language understanding of BERT can be largely retained by the knowledge distillation method for compression.

Table 1. Literature survey and research positioning.

Reference	Approach	Domain	Main contribution
Pang et al. [1]	Supervised ML	Movie reviews	Established ML sentiment classification baseline.
Liu [2]	Survey/book	Sentiment analysis	Taxonomy of lexicon and ML approaches.
Zhang et al. [3]	Deep learning	Sentiment analysis	Survey of neural methods.
Minaee et al. [4]	Deep text classification	Multiple tasks	Reviews neural architectures for text classification.
Hochreiter & Schmidhuber [5]	LSTM	Sequence learning	Long-range dependency modeling.
Vaswani et al. [6]	Transformer	NLP	Introduced attention-only Transformer architecture.
Devlin et al. [7]	BERT	NLP	Pretraining + fine-tuning paradigm.
Liu et al. [8]	RoBERTa	NLP	Optimized BERT pretraining.
Sanh et al. [9]	DistilBERT	NLP	Knowledge-distilled BERT.
Mikolov et al. [10]	Word2Vec	Embeddings	Dense distributed word representations.
Kim [11]	CNN	Sentence classification	Neural convolutional text classification.
Schuster & Paliwal [12]	BiRNN	Speech/NLP	Bidirectional sequence modeling.
Zhang & Wallace [13]	Survey	Sentiment analysis	Deep-learning sentiment analysis overview.
De et al. [14]	Survey	SA/LLMs	Evolution from lexicons to LLMs.
This study [15]	ML/DL/TL	Five-class task	Unified experimental comparison on supplied notebook dataset.

However, recently there were few reviews stating that trends seem to be slowly moving away from classic ML / DL to pretrained transformers and also more and more towards the recent LLM. [13], [14] Besides that, several papers claim the result that fine-tuned Transformers achieve better results as classical classifier approaches, although it depends on training setup, optimizer choice and the nature of the task and data.

3. Methodology

In this work, we propose the systematic pipeline for 5-class sentiment classification as follows. At first, the corpus has been downloaded from an Excel file with 33,142 original instances and then been preprocessed to discard duplicate entries and missing values, which makes it 33,113 instances. Combined the generated caption and sentiment description with each other to form the text input and then encoded sentiment classes into five class. The corpus is split by an 80:20 ratio of stratified split. The train size and test size are 26,490 and 6,623 respectively. Three types of text representation methods include: TF-IDF for ML, token sequences for DL, and token sequences

for transfer learning (Transformer tokens). Finally, use 7 kinds of machine learning/deep learning models to verify that: Random Forest, Logistic Regression, SVM, LSTM, BiLSTM, DistilBERT, and RoBERTa. Also use accuracy, precision, recall, F1-score and confusion matrix as the evaluation metric.

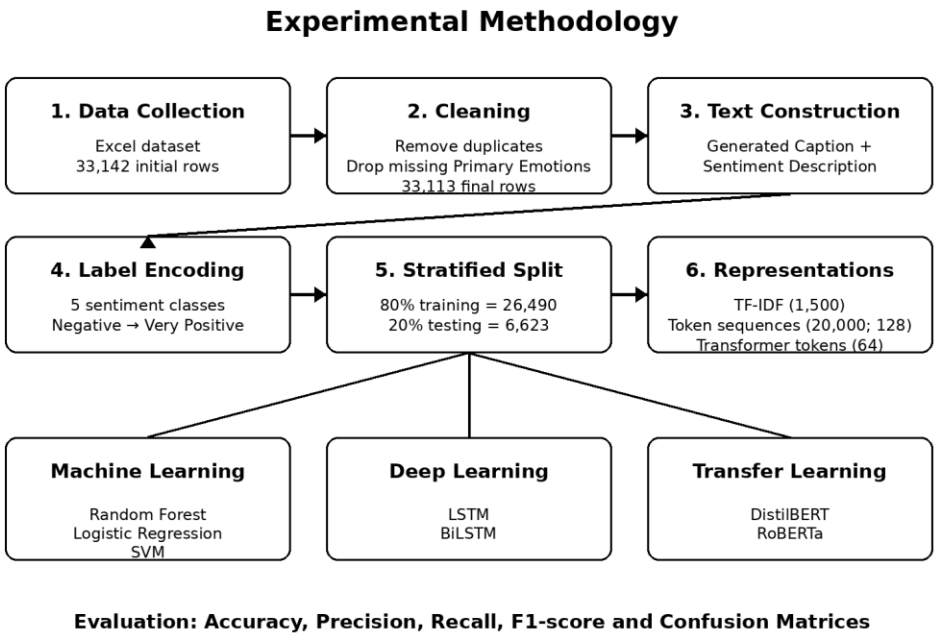

Fig. 1. Experimental methodology.

3.1 Data Collection

Dataset preparation and configuration shown in table. The original dataset consists of 33,142 records, of which 29 records are discarded, duplicates or absent and get 33,113 records for the study. The central input to the model is called combined_text where a synthesized caption and statement sentiment is concatenated to get deeper text for the classification of emotions. Five classes are used: Negative, Neutral, Positive, Very Negative and Very Positive. For the building and verification of the model, 26,490 training samples and 6,623 tests corresponding to an 80:20 stratified splitting were utilized in the last records. Stratified sampling used to prevent data skew and represent proportional ratios of classes throughout all records.

Table 2. Dataset description.

Item	Value	Description
Initial records	33,142	Notebook dataset
Final records	33,113	After cleaning
Input	Generated Caption + Sentiment Description	combined_text
Classes	5	Negative, Neutral, Positive, Very Negative, Very Positive
Train/Test	26,490 / 6,623	80:20 stratified

3.2 Data Preprocessing

Delete duplicate records and filter out observations where there is no 'Primary Emotions'. The generated 'Caption' and 'Sentiment Description' are converted to strings and then concatenated to create the 'combinedtext' feature. Target 'Sentiment Category' is then label-encoded. This notebook uses an 80:20 stratified train-test split. There's a setting on randomstate to=42[15] for this. For the ML model, TF-IDF's max vocab is 1,500, mindf=3, and maxdf=0.85. For the recurrent DL, tokenizer max words: 20,000 OOV token, sequence length: 128. Transformer's max length: 64 truncation and padding[15]. Imbalanced data set with Neutral being the majority class, class-level metric other than just accuracy must be considered.

3.3 Experimental Setup

An experimental setting has been set up in order to provide a fair comparison between all types of models (ML, DL, transfer-learning). The dataset has been split into training (80%) and testing (20%) sets through stratified sampling in order to retain the percentage of sentences for each of the five sentiment classes: 26,490 training and 6,623 testing data. Conventional ML models consist of a process that transforms textual data into TF-IDF features (1,500 features were used to allow models to learn most relevant word representations). DL models are implemented through a Keras tokenizer with a vocabulary of 20,000 words, with texts being padded or truncated to a maximum of 128 tokens. They have been trained with the Adam optimizer using an early-stopping strategy as set forth in the notebook. Transfer learning models make use of pre-trained Transformer models and their respective tokenizers, with a maximum length of 64 tokens allowing them to capture contextual relationships captured during the pre-training phase. Metrics (accuracy, precision, recall, F1-score) are used in order to Evaluate performance accuracy, precision and recall give you insight into the prediction rate and F1-score gives you a balance measure between precision and recall). To assess the performance of each of the Seven experimented models, they have all been compared using a set of metrics, namely accuracy, precision, recall, and F1-score, as well as their respective confusion matrices to detect commonly confusable class pairs.

3.4 Models Implementation

Five models, belonging to the three key ML/DL/TL learning approaches have been considered (Random Forest, Logistic Regression and SVM), learning on textual features extracted from TFIDF features. Specifically, parameters for random forest such as the number of trees or the maximum depth are the default parameters; random forest having 50 trees, a max depth equal to 10. Logistic Regression parameters have been tuned by default; C is 0.01, max_iter to 200 and penalty L2. DL models are LSTM and BiLSTM, that are models that run on the sequence of token. RoBERTa and distilbert belong to the Transfer learning approaches. DistilBERT consists of a compression of the BERT model, namely ‘distilbert-base-uncased’. This models the same concept than the standard Transformer architecture with more lightweight processing;RoBERTa is a deep learning transfer learning model that uses Transformer-based tokenized word representations. The use of RoBERTa implies using a model already trained to capture textual nuances in context.

4. Comparative Results

The following are the test accuracies obtained reported in the notebook [15]. It seems DistilBERT outperform all the others significantly. The best among RNNs and conventional ML methods are LSTM and Random Forest respectively. DistilBERT scored a high of 96.35%, which is 7.58% and 9.29% higher than LSTM and Random Forest respectively. This is main result and argument of the paper. It indicates our hypothesis was correct that a pretrained contextual representation would be a more effective form of semantic discrimination than sparse lexical representations or a learned recurrent representation. The fact that LSTM could score a 88.77% indicates that a sequential approach would be beneficial. However, BiLSTM could only reach 87.06%, tying with RF, suggesting that there was no additional gain from bidirectional representations in the current setup and architecture. In terms of conventional ML methods RF is leading at 87.06% followed by SVM and LR with accuracy 84.70% and 82.03% respectively. RoBERTa scored 83.03%. This is significantly lower than the score that DistilBERT was able to obtain. It is likely not an intrinsic limitation of RoBERTa; in published work, it is known that the performance of Transformers relies heavily on pre-training and fine-tuning choices [8], which we may not have sufficiently exploited in the notebook.

Table 5. Comparative test accuracy of all evaluated models

Model	Family	Accuracy (%)
DistilBERT	Transfer Learning	96.35
LSTM	Deep Learning	88.77
Random Forest	Machine Learning	87.06
BiLSTM	Deep Learning	87.06
SVM	Machine Learning	84.70
RoBERTa	Transfer Learning	83.03
Logistic Regression	Machine Learning	82.03

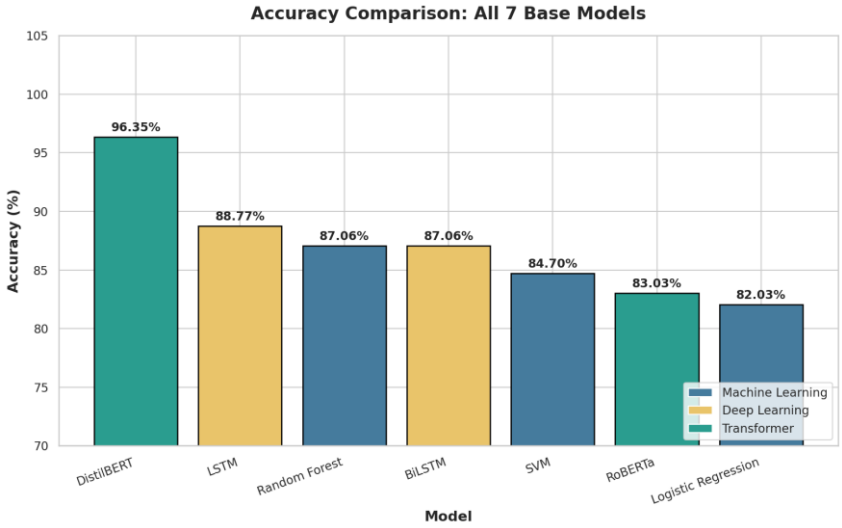

Fig. 2. Accuracy comparison.

4.1 Class-Level Interpretation

The notebook's classification reports show that the dominant Neutral class can strongly influence aggregate accuracy. For example, Logistic Regression has Neutral recall of 1.00 but Positive recall of only 0.20 [15]. This demonstrates why macro-F1, per-class recall and confusion matrices should accompany accuracy in a robust sentiment-analysis evaluation.

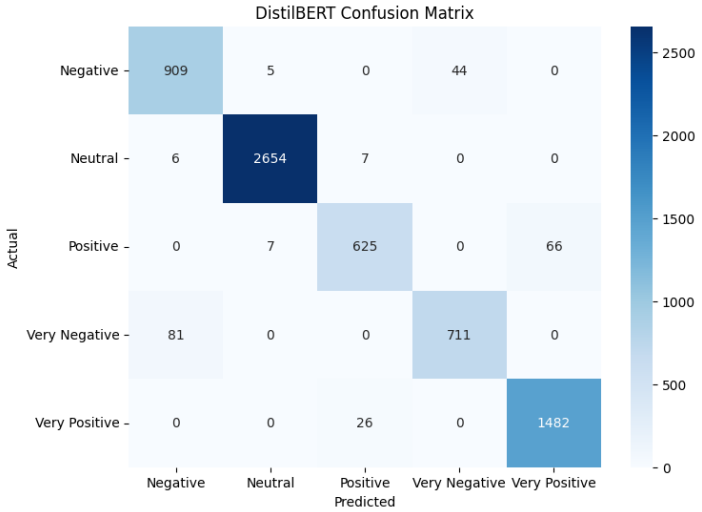

Fig. 3. DistilBERT confusion matrix.

However there appear to be errors on some semantically similar classes, for instance there are 44 incorrectly classified examples between Negative and Very Negative (predicted Very Negative, actual Negative) and 81 examples between Very Negative and Negative (predicted Negative, actual Very Negative), implying these sentiments are confused at some level. 66 Positive examples are misclassified as Very Positive, and 26 Very Positive examples as Positive. Overall the clearly large values along the diagonal of the confusion matrix and small number of errors in off-diagonal cells shows that DistilBERT achieves clear separation between all five classes and backs up its high accuracy score of 96.35% on the experiments.

5. Conclusion and Future Work

In this study, seven sentiment classification models encompassing traditional ML as well as recurrent DL and Transformer transfer learning were compared. The experiment was conducted with a common (five-class) dataset and stratified test set so models could be compared directly to one another within this study. DistilBERT achieved the highest score of 96.35% accuracy. The next highest was LSTM with 88.77%; Random Forest and BiLSTM were tied with 87.06%; SVM with 84.70%; RoBERTa with 83.03%; and Logistic Regression with 82.03% [15]. It is clear from the results that pre-trained contextualized representations provide an overwhelming benefit to this particular task. However, the relatively low RoBERTa accuracy shows that the benefits of transfer learning are far from a guarantee and are configuration sensitive. The unbalanced data set also illustrates that accuracy does not give a complete picture of the results, that recall of minority classes and macro-F1 should be emphasized for this task. Future directions of research include (i) systematic tuning of Transformer hyperparameters, (ii) statistically significant and repeated stratified runs, (iii) using macro-F1 and/or balanced accuracy as the primary evaluation metric, (iv) ablations separating Generated Caption and Sentiment Description, (v) class rebalancing and focal loss, (vi) explainability with methods like SHAP/LIME or token attribution, and (vii) independent external evaluations.

6. References

[1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” EMNLP, pp. 79–86, 2002.

[2] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.

[3] L. Zhang, S. Wang, and B. Liu, “Deep Learning for Sentiment Analysis: A Survey,” WIREs Data Mining and Knowledge Discovery, 2018.

[4] S. Minaee et al., “Deep Learning–Based Text Classification: A Comprehensive Review,” ACM Computing Surveys, 2021.

[5] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 9(8), 1735–1780, 1997.

[6] A. Vaswani et al., “Attention Is All You Need,” NeurIPS, 2017.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019.

[8] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT,” arXiv:1910.01108, 2019.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781, 2013.

[11] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” EMNLP, pp. 1746–1751, 2014.

[12] Tiwari, D., Bhati, B. S., Nagpal, B., Al-Rasheed, A., Getahun, M., & Soufiene, B. O. (2025). A swarm-optimization based fusion model of sentiment analysis for cryptocurrency price prediction. Scientific Reports, 15(1), 8119.

[13] Tiwari, D., Nagpal, B., Bhati, B. S., Mishra, A., & Kumar, M. (2023). A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques: D. Tiwari et al. Artificial Intelligence Review, 56(11), 13407-13461.

[14] Tiwari, D., & Nagpal, B. (2022). KEAHT: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis. New generation computing, 40(4), 1165-1202.

[15] Supplied experimental notebook, “Aug_Sentimental_Transformer (1).ipynb,” dataset preparation, model implementation and evaluation outputs.