Empirical Benchmarking of Traditional Machine Learning and DistilBERT-Based Zero-Shot Hierarchical Sentiment Analysis on Large-Scale Twitter Data

Bhumit Peshavariya[1], Sunny Nahar[2]
[1]Master of Computer Applications, Vivekanand Education Society's Institute of Technology, Collector's Colony, Chembur, Mumbai 400074, Maharashtra, India
[2]Vivekanand Education Society's Institute of Technology, Collector's Colony, Chembur, Mumbai 400074, Maharashtra, India
E-mail: 2025.bhumit.peshavariya@ves.ac.in, sunny.nahar@ves.ac.in

Keywords: DistilBERT, sentiment analysis, hierarchal text classification, zero-shot learning

Received: May 2, 2026

Text sentiment analysis of the social media text faces challenges posed by unstructured data and laborious human labeling for intent-driven, hierarchical classification. This work compares conventional ML models (SVM, Naïve Bayes, Logistic Regression) with contextual DL models (DistilBERT) in terms of their performance on Sentiment140 dataset (1.6 million tweets) where a balanced 300,000 tweets were selected (200,000 training, 50,000 validation and 50,000 test). As a solution to the bottleneck of human labeling for detailed topic classification, a zero-shot classification pipeline that uses Natural Language Inference (NLI) for classification of 1,000 tweets into a two-layer taxonomy of 30 parent topics and 330 subtopics has been created without any human-labeled samples. SVM is able to achieve 81.52% accuracy, while DistilBERT scores 84.44% on 50,000 tweet test set and 83.0% on a small 1,000 tweets sample.

Povzetek: Analiza sentimenta na družbenih omrežjih se sooča z nestrukturiranimi, šumnimi podatki in obsežnim ročnim označevanjem, potrebnim za hierarhično klasifikacijo glede na namen uporabnika. Ta študija primerja tradicionalne modele strojnega učenja (SVM, naivni Bayes, logistična regresija) z globokim kontekstualnim modelom (DistilBERT) na naboru Sentiment140 (1,6 milijona tvitov), iz katerega je bil izbran uravnotežen podnabor 300.000 tvitov (200.000 učnih, 50.000 validacijskih, 50.000 testnih). Za odpravo ozkega grla pri označevanju je bil razvit cevovod za brezvzorčno (''zero-shot'') klasifikacijo na osnovi sklepanja naravnega jezika (NLI), ki je 1.000 tvitov razvrstil v dvonivojsko taksonomijo 30 glavnih in 330 podkategorij brez ročno označenih primerov. Rezultati kažejo, da SVM doseže 81,52-odstotno natančnost, DistilBERT pa 84,44 % na testnem naboru 50.000 tvitov in 83,0 % na manjšem podnaboru 1.000 tvitov, uporabljenem tudi za zero-shot kategorizacijo tem.

1 Introduction

The difficulties faced by social media sentiment analysis in dealing with extremely noisy and unstructured texts along with constraints of only three labels: positive, negative, and neutral have been extensively discussed in literature [14, 12]. In order to capture user intents accurately, the use of Hierarchical Text Classification (HTC) is inevitable. However, the amount of data annotation required to train such models is enormous and results in a significant labeling problem [2, 13].

This paper proposes a two-fold solution to the above-mentioned issues. On one hand, it compares the performance of conventional models (SVMs, Naïve Bayes, Logistic Regression) to that of a context-aware deep learning architecture (DistilBERT) by analyzing 1.6 million tweets. On the other hand, it presents an NLI-driven zero-shot HTC framework which can automatically classify 1,000 tweets into 30 major categories and 330 minor categories without performing any labeling tasks.

These two elements of the study investigate separate research questions that should not be compared as two experimental procedures, based on a different scale and standard of evidence. This study seeks to (1) conduct an empirical comparison between the performance of traditional machine learning models and a deep contextual model (DistilBERT) in binary sentiment classification on a large, stratified ground truth labelled sample of Twitter data, and (2) test the potential of using NLI-based zero-shot approach in assigning social media messages to a large fine-grained topic taxonomy without any manual labelling. The research questions investigated in this study are the following: (a) How accurate are the traditional machine learning models and a deep contextual model in the task of sentiment classification on a large ground truth labelled Twitter sample? (b) Is it possible to use NLI-based zero-shot approach for assigning tweets to a fine-grained topic taxonomy without manual labelling, and what is the limitation of this pipeline

in practical applications?, this research question will be answered in a qualitative way by inspecting the results of the pipeline.

2 Literature review

2.1 Traditionally applied machine learning in sentiment analysis

Machine learning methods have been applied traditionally and are the fundamental techniques used in text classification and sentiment analysis. Some of the traditional machine learning techniques are the Naïve Bayes, SVM, and Logistic regression models that operate on handcrafted features, specifically, TF-IDF. From among all these techniques, SVM becomes one of the most efficient baseline models in terms of efficiency in maximizing margins between vectors that are sparse and high dimensional [4, 11]. Similar traditional approaches combining Naïve Bayes and Support Vector Classification have also been applied directly to Twitter sentiment datasets [15]. Comparative studies have further shown that deep learning models such as RNN, LSTM, and GRU tend to outperform classical Naïve Bayes and Logistic Regression baselines on textual sentiment tasks [1]. Despite being efficient and fast, such techniques have the inherent limitation of not being able to solve the problem of dependencies and semantics.

2.2 Deep learning and transformer models

Due to the incapacity of conventional models to address issues concerning syntactic complexity, the research domain shifted its focus towards deep learning and transformer models. In recent times, with the invention of Bidirectional Encoder Representations from Transformers (BERT) [5], there has been a breakthrough in the efficacy of machine learning models. With the help of self-attention layers where words are treated not only based on their connection with the subsequent word but also with respect to all the words in the sentence, DistilBERT manages to capture the bidirectional context of words [16, 10]. Comparing tests performed using a domain that is characterized by noisy textual information, such as Twitter, it can be seen that DistilBERT outperforms conventional machine learning methods. The present study employs DistilBERT, a compact, distilled variant of BERT that retains the same self-attention mechanism at substantially reduced computational cost, as detailed in Section 3.3.

2.3 Hierarchical text classification (HTC) and zero-shot learning

To avoid confusion, three separate tasks will be discussed in this paper that are all related to each other: sentiment analysis – assigning a polarity to a piece of text (positive or negative); topic classification – assigning text into a label from a flat list of labels; and finally Hierarchical Text Classification (HTC) – assigning text to a certain point inside a nested taxonomy of categories [18]. This paper considers both sentiment classification and hierarchical topic categorization as two separate complementary tasks, which are implemented by separate modules of the pipeline and cannot confirm each other's validity.

Whereas traditional sentiment analysis algorithms may be used to detect the sentiment of a piece of text, knowing the intention behind the message requires the application of Hierarchical Text Classification (HTC) in order to map the text to the intricate hierarchy of topic-based classifications [19]. Nevertheless, HTC is highly dependent on the availability of a massive quantity of annotated training samples to operate, a strategy fraught with significant challenges owing to the presence of a "labeling bottleneck" – i.e., the difficulty of labeling large volumes of data with multiple categories [2, 13]. Thus, ZS-HTC is one of the most essential fields of research.

2.4 Summary of related work

The summary of related work is presented in Table 1, showing how their datasets, approaches, and findings relate to the joint sentiment–zero-shot-HTC experiment in this paper.

2.5 Current literature gap

While both areas have seen great advancements separately, the current state of the literature shows a definitive divide between the two techniques. In current literature, scientists usually focus on evaluating benchmark performance for models' architectures in general sentiment analysis [6, 1, 3], or zero-shot hierarchical classification [2, 13, 17]. In comparison to the quantity of publications focused on either of the two topics separately, there isn't much literature about applying both methods simultaneously. This paper seeks to bridge this gap by implementing classical machine learning algorithms and DistilBERT models, utilizing 1.6 million tweets, alongside a zero-shot classifier for 30 parent categories and 330 subcategories.

3 Theoretical framework & mathematical formulations

3.1 Text vectorization mathematics

The vectorization process with the help of mathematics becomes necessary for conducting an analysis of unstructured data like social media data. In order to calculate the relevance of terms in comparison to the other documents present in the corpus, traditional machine learning algorithms utilize the concept of TF-IDF. This process can be described mathematically as:

$$W_{t,d} = TF_{t,d} \times \log\left(\frac{N}{DF(t)}\right) \tag{1}$$

Table 1: Related work summary for sentiment analysis and zero-shot hierarchical text classification

Reference	Task (Data Set)	Approach	Achieved Result
Go, Bhayani & Huang (2009)	Binary sentiment analysis (Sentiment140)	NB, MaxEnt, SVM	>80% accuracy
Elmitwalli & Mehegan (2024)	Binary sentiment analysis (Sentiment140)	BERT, GPT-3, Bi-LSTM	F1 = 0.811 (BERT)
Chauhan et al. (2025)	Aspect-based sentiment analysis (SemEval)	RoBERTa	State-of-the-art (aspect-level)
Bongiovanni et al. (2023)	Zero-shot HTC (4 public data sets)	Z-STC + Upward Score Propagation	State-of-the-art on 3/4 data sets
Paletto et al. (2024)	Zero-shot HTC	Label augmentation	Increased accuracy of zero-shot HTC
Xia et al. (2025)	Zero-shot HTC (LLMs)	Ensembled prompting	Gains over single-prompt baseline
This paper	Sentiment + Zero-shot HTC (joint)	SVM/NB/LR/DistilBERT + NLI two-pass HTC	84.44% (DistilBERT); HTC: no ground truth

Where N is the number of total documents in the corpus, and $DF(t)$ represents the documents containing the term t.

3.2 Formulation of machine learning methods

3.2.1 Naïve bayes (NB)

The formulation of Bayes' theorem can be adopted in this machine learning method to predict the sentiments of the texts. In this NB algorithm, it is assumed that all attributes in the vector space model are independent:

$$P(c|d) = P(c)\prod P(w_i|c)P(d) \tag{2}$$

3.2.2 Support vector machine (SVM)

In the SVM model with the linear kernel, the aim is to discover the optimal hyperplane that maximizes the separation between the positive and negative sentiments. The objective function needs to be minimized by:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) \tag{3}$$

3.2.3 Logistic regression

The logistic function transforms the output to a probability value that decides whether the output is either 0 or 1 by logistic regression. The logistic function is depicted as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

where z is a function of weight and input values ($w^T x + b$).

3.3 Deep learning models

3.3.1 DistilBERT and attention model

DistilBERT, a distilled version of BERT obtained via knowledge distillation, retains BERT's self-attention mechanism while using roughly half the transformer layers, applying self-attention in processing words bidirectionally while taking into account the complete sentence context. The equation describing the transformer model can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

Where Q indicates the query matrix, K the key matrix, V denotes the value matrix, and d_k denotes the scaling constant.

3.4 Hierarchical classification and zero-shot learning

In order to classify the tweets according to their place in the two-layer taxonomy without using any labeled training data, a zero-shot NLI-based classifier is used. Each possible class for the classifier (parent category or child category), is described by a brief phrase; for example, parent category "Health" is expressed through the phrase "health and medical problems," while the child category "Mental Health" is described as "mental health and psychological wellbeing." For a particular tweet x and candidate label ℓ, the classifier assesses the combination of the tweet and the hypothesis generated from a predefined template:

$$h(\ell) = \text{"This tweet is about } p(\ell)\text{."} \tag{6}$$

where p(ℓ) denotes the descriptive phrase for label ℓ. The underlying NLI model produces an entailment probability for the tweet against each candidate hypothesis within the current label set, and these probabilities are normalized via softmax across the set to yield a similarity score S(x, ℓ) $\in$ [0,1].

Classification proceeds in two passes that mirror the two levels of the taxonomy. In the first pass, the tweet is scored against the full set of 30 parent-category phrases $C = \{c_1, \ldots, c_{30}\}$, and is assigned to the highest-scoring category, subject to a minimum-confidence threshold θ_{cat}:

$$c^* = \arg\max_{c \in C} S(x, c) \tag{7}$$

On the other hand, when $S(x, c^*) < \theta_{cat}$, the tweet will be allocated to a bucket corresponding to the second-level fallback "Others" bucket instead of being forced into a less-confident category.

In the second run, the tweet will be scored again but using only the ten sub-category phrases that fall under the predicted parent category in the first stage ($Sub(c^*) = \{s_1, \ldots, s_{10}\}$) as follows:

$$s^* = \arg\max_{s \in Sub(c^*)} S(x, s) \tag{8}$$

under a threshold confidence level, θ_{sub}; otherwise, if the score is lower than the threshold, then the tweet will be classified in a second-level fallback bucket called "Others" belonging to c^*. In this experiment, the thresholds are set at $\theta_{cat} = 0.20$ and $\theta_{sub} = 0.25$. In such a manner, the candidate set at every stage of classification is kept small and topic-related by only considering the ten children of the predicted parent category instead of scoring all 330 subcategories directly for each tweet.

4 Approach

4.1 Datasets configuration and unbiasing approach

With regards to the empirical grounds of the research, Sentiment140, which is one of the most popular and frequently utilized datasets including 1.6 million labeled tweets, half of which are negative and half positive, was chosen [7]. Taking into account the possible biasness of social media data, much attention was paid to configuring the datasets properly. To start with, one needs to note that the natural configuration of the Sentiment140 [8] dataset presupposes sorting tweets chronologically according to their polarities (negative to positive). In other words, if some sample of the dataset is taken randomly without being shuffled, it would provide a very dangerous biasness for the classifier because the latter would be trained only on one polarity. To solve the issue, Data Shuffling and Stratification techniques were used. All 1.6 million tweets were shuffled, and a stratified split was carried out to preserve the near-50/50 class ratio.

Taking into account the time constraints of performing all experiments locally, a subsample of 300,000 tweets (instead of 1.6 million) was taken to conduct the sentiment classification experiment. It was distributed as follows:

- Training set – 200,000 tweets (approximately 100,000 negative / 100,000 positive);
- Validation set – 50,000 tweets (approximately 25,000 negative / 25,000 positive);
- Test set – 50,000 tweets (approximately 25,000 negative / 25,000 positive).

Each one maintains a balanced representation of the classes in the original dataset with minimal variance from a perfect 50/50 division due to stratified sampling.

4.2 Data preprocessing

The raw data that we obtained through social media channels was filled with many errors such as grammatical issues, emoticons, and features that are peculiar to the social media channel. For us to achieve consistency in input vectorization as expected in both algorithms, the following data preprocessing techniques were done: URL, HTML tag, Twitter handle, and special characters removal, lowercase conversion of the entire text, and contraction expansion.

4.3 Tokenization and vectorization

The text data underwent vectorization using techniques which varied according to the model chosen:

- **Classical Machine Learning Models:** The text data underwent tokenization using the n-gram technique before vectorization using TF-IDF in the case of support vector machines, Naïve Bayes, and logistic regression.
- **Deep Learning Models:** Wordpiece tokenization technique was employed in the DistilBERT framework as part of our efforts to retain rich contexts in the texts [5]. Wordpiece is vital for handling rare words in the text by breaking them down to meaningful subwords.

4.4 Pipeline for zero-shot hierarchical text classification

Although the first batch of 300,000 tweets served to benchmark polarity classifications in the four models discussed in Sections 3.2-3.3, such polarity-based analysis doesn't take into account the topic. For that reason, a smaller batch of 1,000 tweets was selected for zero-shot hierarchical text classification. Unlike in case of polarity classification, zero-shot hierarchical text classification takes much more computational resources, as the pipeline requires scoring each tweet twice using NLI; in this regard, processing the batch of 1,000 tweets locally took about ten minutes.

The 1,000 tweets were mapped onto two-level hierarchy containing 30 parent nodes and 330 subcategories (10 per category + Others as 11th category). In order to build the taxonomy, the set of 30 high-level civic and social discourse categories were identified based on an analysis of

existing taxonomies of topics and public discourse literature, followed by curating and refining the subcategories for each parent category.

Not using any manually labelled examples, Natural Language Inference (NLI) zero-shot classifier (MoritzLaurer/deberta-v3-xsmall-zeroshot-v1.1-all-33) [9] was used to score each tweet with respect to the taxonomy, applying the procedure of single hypothesis-template scoring and two-pass hierarchical filtering as discussed in Section 3.4. Thus, every tweet was first scored against the 30 parent categories and was allocated to the most probable one, or to the top-level 'Others' bucket if there is no category exceeding the confidence threshold; afterwards, the tweet was scored again only against the 10 subcategories of its assigned parent category.

For instance, in case the tweet expresses frustration concerning the long queues at the hospital's emergency department, it will be initially scored against 30 parent categories like "This tweet is about health and medical concerns" (Health) and "This tweet is about infrastructure and utilities" (Infrastructure), with the winning hypothesis assigned at the Health category that meets the specified confidence threshold. Further, the tweet will be re-scored against 10 subcategories under the winning parent category of Health, which may include "This tweet is about hospitals and medical access" (Healthcare Access) and "This tweet is about mental health and psychological wellbeing" (Mental Health).

4.5 Formulation of evaluation metrics

The prediction strength of all models is analytically estimated based on classification performance metrics involving true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

4.6 Implementation details and reproducibility

All classic models are trained on TF-IDF features with unigrams and bigrams (ngram range 1–2), max vocabulary **50,000**, min df **5** and max df **0.9** (terms occurring in more than 90% documents were discarded). The same vectorizer was fit only once on the entire 200,000-tweet training set and then kept unchanged for validation, testing, and all classic models.

4.6.1 Classical machine learning models

- **Support Vector Machine (SVM):** `LinearSVC` with default regularization ($C = 1.0$), squared hinge loss, and a fixed random state (**42**) trained in one go on the entire 200,000-tweet training matrix.
- **Naïve Bayes:** `MultinomialNB` with default Laplace smoothing ($\alpha = 1.0$) trained in four consecutive steps of 50,000-tweets each from the entire training matrix.
- **Logistic Regression:** Default L2 regularization ($C = 1.0$), `lbfgs` solver and a maximum of 1,000 iterations. **Because of the specific way the training loop is implemented, in case of this model the resulting parameters are based on training the last 50,000-tweet chunk processed, not the entire 200,000-tweet training matrix**, which we would like to point out here as a limitation of this study.

4.6.2 DistilBERT

DistilBERT (`distilbert-base-uncased`) is fine-tuned with the following configuration:

- Maximum sequence length: **128** tokens
- Effective batch size: **32** (batch size 16 with two-step gradient accumulation)
- Base learning rate: 3×10^{-5} with **layer-wise learning-rate decay** (decay factor 0.9 for each of the six layers)
- Learning rate schedule: **cosine** with 6% linear warmup
- Weight decay: **0.01** on all the parameters except bias and LayerNorm parameters
- Dropout: **0.2** for embeddings, attention layers and classification head
- Embeddings are frozen during epoch 1 and then are unfrozen
- Number of training epochs: up to **5** with early stopping (patience of 2 epochs without validation-accuracy improvement)
- Mixed-precision (FP16) training used when a CUDA-enabled GPU was available

4.6.3 Data splitting and environment

The data splitting and sampling processes involved a fixed random seed (**42**) throughout, with stratification of the training, validation, and test sets to ensure that they maintained the approximate balance of classes of the source dataset. Similarly, the same seed value was utilized during the training of the SVM classifier. **The DistilBERT model did not have any fixed global seed apart from the one used during data splitting**.

Training of the models was done through standard open-source libraries (scikit-learn for classical models; PyTorch and Hugging Face Transformer for DistilBERT).

5 Results & empirical analysis

5.1 Benchmarking of overall sentiment

The testing of the predictive capabilities of the baselines from classical ML techniques and deep learning neural networks was conducted on the strictly stratified dataset of 50,000 tweets to estimate their generalization capabilities accurately. The empirical findings, summarized in **Table 2**, reveal the clear superiority in performance of one group over another. The neural network based on the DistilBERT architecture proved itself to be the best solution in terms of test accuracy and F1-Score metrics, achieving 84.44% accuracy score and 0.84 of the Macro/Weighted Average F1-Score metric. Such performance highlights the role of bidirectional self-attention mechanisms in natural language processing of social media content. Among the fast and lightweight approaches, Support Vector Machines proved themselves to be the most accurate. The test accuracy of the model reached 81.52% and its F1-Score was 0.82, proving that it can be considered stable for use in classification of sparse high-dimensional TF-IDF vectors. At the same time, the Naive Bayes and Logistic Regression models performed worse in terms of accuracy, getting 79.66% and 78.60% accordingly. The test confusion matrix for DistilBERT, illustrating this high accuracy on the full 50,000-tweet test set, is shown in Figure 1.

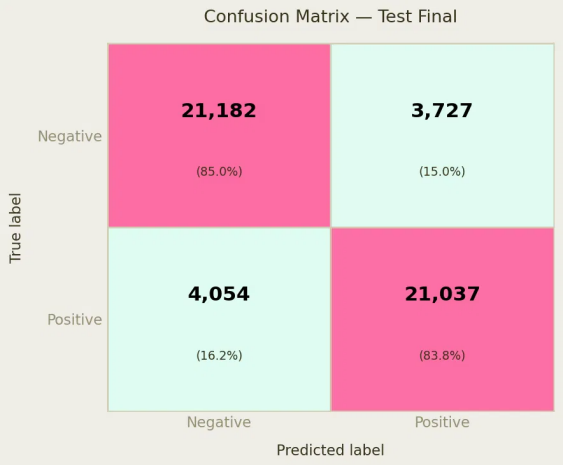

Figure 1: Test confusion matrix for the best performing model (DistilBERT), evaluated on the full 50,000-tweet test set.

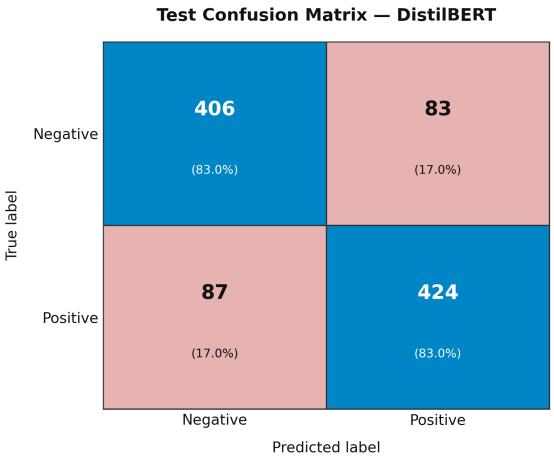

Figure 2: Test confusion matrix for the best performing model (DistilBERT), evaluated on the 1,000-tweet subset.

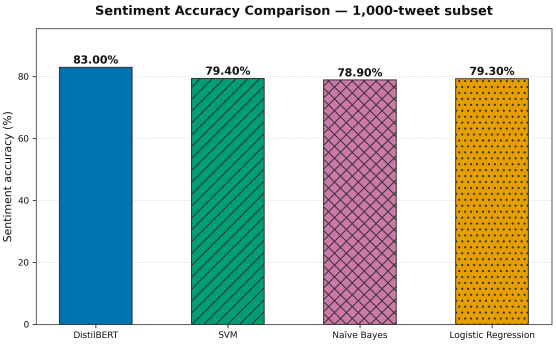

Figure 3: Sentiment accuracy comparison on the 1,000-tweet subset.

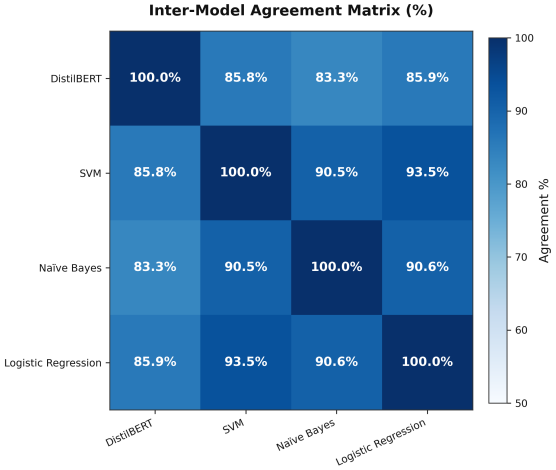

Figure 4: Inter-model agreement matrix highlighting prediction consensus, computed on the 1,000-tweet subset.

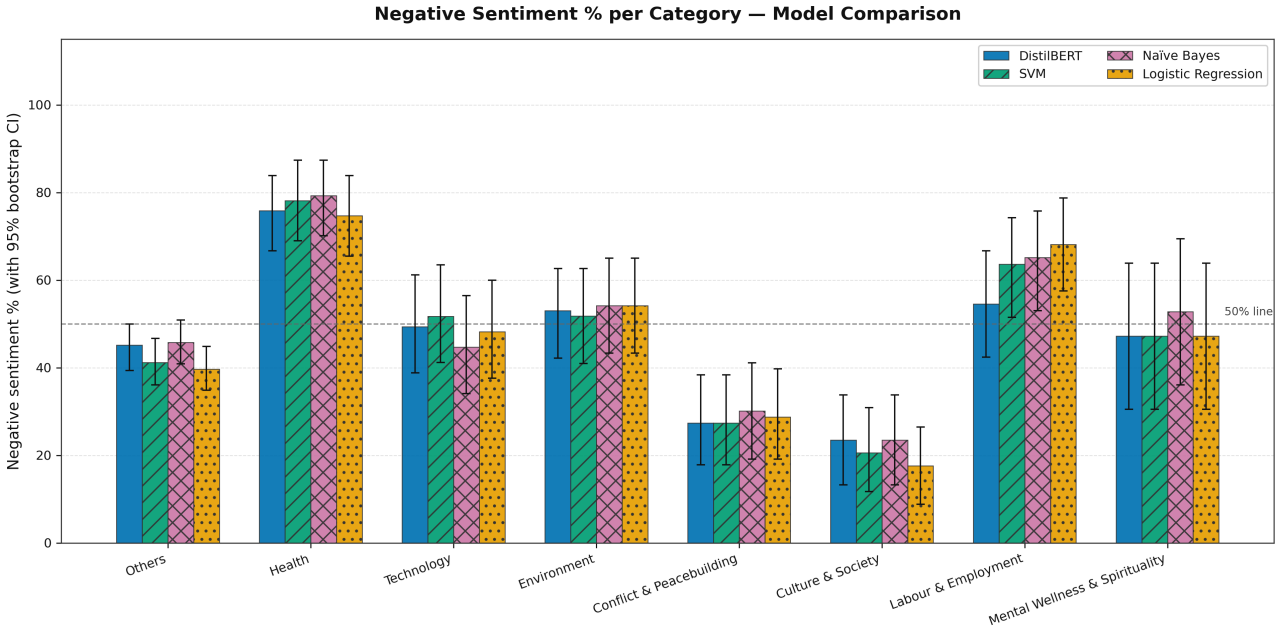

Figure 5: Topic-dependent sentiment analysis showing negative sentiment percentages across different zero-shot categories.

With Sentiment140 being a balanced dataset (800,000 negative / 800,000 positive tweets), and also with training / validation / testing split maintaining this balance through stratified sampling, macro-average and weighted average precision, recall and F1-score metrics become essentially identical for this particular classification problem; hence, the metrics presented in Table 2 could be interpreted equivalently.

Reported metrics in Table 2 are from a single training run per model (no repeated runs across random seeds); consequently no confidence intervals or standard deviations are reported for these point estimates. Future work will incorporate multi-seed evaluation to quantify variance.

Error bars show 95% bootstrap confidence intervals (1,000 resamples) on the per-category negative-sentiment proportion, reflecting sampling uncertainty from the finite number of tweets per category. This is distinct from the single-training-run limitation noted below Table 2, which concerns variance across repeated model training rather than variance in a fixed model's predictions across a resampled tweet population.

6 Discussion

6.1 Comparison with related work

In comparison to previous benchmarks using only sentiment analysis, our DistilBERT performance (F1 = 0.84) is comparable to, and slightly exceeds, the F1 achieved by BERT on the same Sentiment140 dataset by Elmitwalli and Mehegan (2024) [6] (F1 = 0.811, approximately similar in scale), which can be reasonably attributed to the disparity in preprocessing, training set sizes (200,000 vs. their entire corpus), and fine-tuning strategy. On the other hand, our SVM performance (81.52%) agrees with the >80% accuracy range that was previously achieved by Go et al. (2009) with the same distant supervision scheme [8], lending credibility to our preprocessing and stratification pipeline. As far as the zero-shot taxonomy-based classification task goes, the MSP method we employ through two confidence thresholding passes is most similar to that of Bongiovanni et al. (2023) [2], who propagate scores through the taxonomy from bottom up; however, unlike them, whose approach is tested on labeled benchmarks, we do not have any ground truth in our 1,000-tweet subset, hence making a quantitative comparison impossible - more on this limitation we discuss in Section 7.2.

6.2 Dominance of deep learning models over ML efficiency

As seen from the experimental results obtained using this data set, there is an overall trend which sets apart methods based on feature engineering from those based on deep context. In the context of this particular research, the highest accuracy result is obtained using DistilBERT with an accuracy of 84.44%. This implies that models based on deep context might perform better compared to traditional feature-engineering-based approaches for sentiment analysis on noisy Twitter text.

Table 2: Test accuracy and F1-score comparison across models

Model	Precision	Recall	F1-Score	Test Acc.
Logistic Regression	0.79	0.79	0.79	78.60%
Naïve Bayes	0.80	0.80	0.80	79.66%
SVM	0.82	0.82	0.82	81.52%
DistilBERT	0.84	0.84	0.84	84.44%

6.3 The optimal performance by SVM

As good as deep learning is, SVM proves to be the optimal choice from the traditional models, with very high accuracy of 81.52%. While Naïve Bayes assumes that the feature attributes are independent and logistic regression is applied as a probability mapping model, SVM has been designed mathematically to tackle high dimensional and sparse matrices that have been generated using N-gram and TF-IDF extraction. This is evident in the high correlation (85.8%) of the performance between SVM and BERT.

6.4 Zero-shot HTC and topic-dependent sentiment

One of the patterns observed from this analysis is that the ratio of negative sentiment was highly variable among different topic categories. From the observations made after categorizing the tweets using zero-shot methods, it can be observed that the 'Health' category of tweets mostly had negative sentiments, while the 'Culture & Society' category had mostly positive sentiments. This trend demonstrates how a general sentiment analysis without differentiating between the topics could hide differences in the data, which means that topic-wise sentiment analysis might provide an insight into the perception of the people that is greater than simple positive and negative sentiment analysis.

7 Conclusion and future scope

7.1 Conclusion

This study conducts an empirical study among the basic machine learning methods and the deep contextual models on a large sample selected from the 1.6 million tweets Sentiment140 dataset. In this experiment, although the traditional methods provided a good balance between time and computational efficiency, the deep contextual models showed better accuracy in classifying the data, which implies that deep learning models could be considered in sentiment analysis for noisy and unstructured social media text. For instance, Support Vector Machine (SVM), which is the best-performing method among all the traditional methods considered in this study, obtained an accuracy of 81.52% by finding the optimal hyperplane in the TF-IDF vector space. DistilBERT attained better accuracy and F1-Score at 84.44% and 0.84, respectively.

In addition, this work proves the possibility of applying an NLI-based zero-shot pipeline as an alternative to manual labeling in ZS-HTC. The data with respect to the sentiment polarity associated with topics was investigated via the dynamic classification of a subset of data into 30 main categories and 330 subcategories through two passes of hierarchical NLI evaluation with fallback categories based on confidence thresholding. As there is no ground truth label available for the subset of data, the classification results cannot be evaluated against the reference, and thus are just meant to serve as an illustration of feasibility.

For these reasons, this study highlights one of the limitations of the tri-class label system and provides early support, based on the category and data set under study, that topic-based sentiment analysis can uncover trends that may otherwise have been overlooked using only a positive-negative-neutral approach.

7.2 Directions for future research

Although the suggested framework is quite efficient, analyzing errors made while classifying data will help identify ways to improve its architecture and mathematics in the future. The following may be listed as two main objectives of further research:

1. **The Sarcasm Correction Layer**: Inclusion of sarcasm in messages sent via social networks constitutes a persistent trend and is characterized by adversarial noise due to the inversion of polarity of mathematical vectors. No initial experiments related to this module have been undertaken within the framework of this research project. This module is suggested as a topic for future research. In future instances of this pipeline, a development of a special layer of the neural network for detecting and correcting sarcasm will occur prior to the classification stage through DistilBERT. This is supposed to increase accuracy above the 88% mark.

2. **Taxonomies Development and Enhancement**: While the taxonomies built using zero-shot learning showed promise, problems of classification variance arose while working with similar classes, when tweets classified into closely related child nodes tended to get misclassified into an adjacent class. Without any ground truth labels available for the subset of 1,000 tweets, no metrics about accuracy of the zero-shot classification stage are provided in this paper; this conclusion was made after manual examination of the categories assigned. Therefore, the next step of development would require the enhancement of

the MSP algorithm and prototypes augmented by knowledge of large language models, along with creation of the manually labeled ground-truth subset that will make the quantitative assessment possible.

Statements and declarations

- **Funding:** Funding from any sources was not provided for assisting in the preparation of this manuscript.

- **Competing Interests:** The author declares no competing interest relevant to this article.
- **Ethical Approval:** Not applicable.
- **Informed Consent:** Not applicable.
- **Data Anonymization:** Twitter handles, URLs, and hashtags were automatically stripped off all tweets before any analysis or modeling was done, and no identifying information (account usernames, IDs, timestamp) was kept beyond the tweet text and its associated sentiment label.
- **Data Availability Statement:** The data set for this study is Sentiment140, which can be accessed online `https://www.kaggle.com/datasets/kazanova/sentiment140`. The code that was used to prepare the data and conduct the classification process is obtainable on request from the corresponding author.
- **Use of AI-Assisted Tools:** The use of AI generation technology was applied to aid in the review of the existing literature. The methodology, results, and conclusions of this paper were not generated using AI technology; they are the original works of the authors.

References

[1] Hero O. Ahmad and Shahla U. Umar. “Sentiment Analysis of Financial Textual data Using Machine Learning and Deep Learning Models”. In: *Informatica* 47 (May 2023). DOI: 10.31449/inf.v47i5.4673.

[2] Lorenzo Bongiovanni et al. “Zero-Shot Taxonomy Mapping for Document Classification”. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. Tallinn Estonia: ACM, Mar. 2023, pp. 911–918. ISBN: 978-1-4503-9517-5. DOI: 10.1145/3555776.3577653.

[3] Amit Chauhan, Aman Sharma, and Rajni Mohana. “An Enhanced Aspect-Based Sentiment Analysis Model Based on RoBERTa For Text Sentiment Analysis”. In: *Informatica* 49 (Mar. 2025), pp. 193–202. DOI: 10.31449/inf.v49i14.5423.

[4] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20 (Sept. 1995), pp. 273–297. DOI: 10.1007/BF00994018.

[5] Jacob Devlin et al. “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, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[6] Sherif Elmitwalli and John Mehegan. “Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques”. English. In: *Frontiers in Big Data* 7 (Mar. 2024). DOI: 10.3389/fdata.2024.1357926.

[7] Alec Go, Richa Bhayani, and Lei Huang. *Twitter Sentiment Classification using Distant Supervision*. Sentiment140 Dataset. 2009.

[8] Alec Go, Richa Bhayani, and Lei Huang. *Twitter Sentiment Classification using Distant Supervision*. Tech. rep. Stanford University, Jan. 2009.

[9] Moritz Laurer. *deberta-v3-xsmall-zeroshot-v1.1-all-33*. Hugging Face. 2023.

[10] Min Li, Shujuan Guo, and Runchen Liu. “BERT-based Consumer Sentiment Analysis for Personalized Marketing Strategies”. In: *Informatica* 49 (Aug. 2025). DOI: 10.31449/inf.v49i28.8232.

[11] Tony Mullen and Nigel Collier. “Sentiment Analysis using Support Vector Machines with Diverse Information Sources”. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 412–418.

[12] Shaha T. Al-Otaibi and Amal A. Al-Rasheed. “A Review and Comparative Analysis of Sentiment Analysis Techniques”. In: *Informatica* 46 (July 2022). DOI: 10.31449/inf.v46i6.3991.

[13] Lorenzo Paletto, Valerio Basile, and Roberto Esposito. “Label Augmentation for Zero-Shot Hierarchical Text Classification”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre

Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7697–7706. DOI: 10.18653/v1/2024.acl-long.416.

[14] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, July 2002, pp. 79–86. DOI: 10.3115/1118693.1118704.

[15] Vrinda Tandon and Ritika Mehra. "An Integrated Approach For Analysing Sentiments On Social Media". In: *Informatica* 47 (June 2023). DOI: 10.31449/inf.v47i2.4390.

[16] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[17] Mingxuan Xia et al. "Ensembling Prompting Strategies for Zero-Shot Hierarchical Text Classification with Large Language Models". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos et al. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 18189–18208. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.918.

[18] Wenpeng Yin, Jamaal Hay, and Dan Roth. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3914–3923. DOI: 10.18653/v1/D19-1404.

[19] Qianbo Zang et al. "KG-HTC: Integrating Knowledge Graphs into LLMs for Effective Zero-shot Hierarchical Text Classification". In: *arXiv e-prints*, arXiv:2505.05583 (May 2025), arXiv:2505.05583. DOI: 10.48550/arXiv.2505.05583. arXiv: 2505.05583 [cs.CL].

Appendix: Full taxonomy

The following table presents the entire two-tier taxonomy consisting of 30 categories and 300 subcategories (10 subcategories for each parent) which are utilized by the zero-shot NLI classification method discussed in Sections 3.4 and 4.4. As pointed out earlier, an additional subcategory of "Others" type is added under each parent category if none of the subcategories' scores is greater than the confidence threshold of θ_{sub}, resulting in a total of 330 subcategories used in this paper.

- **Health:** Healthcare Access, Mental Health, Public Health, Disability, Elderly Care, Maternal Health, Substance Abuse, Nutrition, Epidemic & Disease, Health Insurance, Others
- **Education:** Primary Education, Higher Education, Vocational Training, Digital Literacy, Teacher Quality, School Infrastructure, Student Welfare, Scholarships, Adult Literacy, Online Education, Others
- **Infrastructure:** Roads & Highways, Public Transport, Water Supply, Power & Electricity, Waste Management, Internet & Telecom, Urban Planning, Rural Development, Housing, Public Buildings, Others
- **Environment:** Climate Change, Air Pollution, Water Pollution, Natural Disasters, Wildlife & Forests, Renewable Energy, Plastic Waste, Soil Degradation, Noise Pollution, Carbon Emissions, Others
- **Economy:** Employment & Jobs, Wages & Labour, Small Business, Agriculture, Food Security, Inflation & Prices, Banking & Finance, Taxation, Trade & Exports, Economic Inequality, Others
- **Governance:** Government Policy, Corruption, Elections, Judiciary, Police & Law, Human Rights, Military & Defence, Public Administration, Local Governance, International Affairs, Others
- **Social Issues:** Poverty & Inequality, Women & Gender, Child Protection, Youth Issues, Minority Rights, Immigration, Religious Freedom, Homelessness, Discrimination, LGBTQ+ Rights, Others
- **Technology:** Cybersecurity, Artificial Intelligence, Social Media, E-Governance, Fintech, Space & Science, Data Privacy, Automation & Jobs, Digital Infrastructure, Tech Startups, Others
- **Public Safety:** Crime & Violence, Terrorism, Road Safety, Fire Safety, Disaster Response, Food Safety, Drug Trafficking, Cybercrime, Child Safety, Workplace Safety, Others
- **Culture & Society:** Sports & Recreation, Arts & Culture, Media & Press, Religion, Tourism, Heritage, Language Rights, Community Welfare, Festivals & Events, Entertainment, Others
- **Agriculture & Food:** Crop Production, Irrigation, Fertilisers & Pesticides, Farmer Welfare, Agricultural Policy, Food Processing, Livestock & Dairy, Organic Farming, Food Markets, Rural Livelihoods, Others
- **Water & Sanitation:** Drinking Water, Sewage & Drainage, Open Defecation, Groundwater, Flood Drainage, Water Scarcity, River & Lake Health, Industrial Discharge, Handwashing & Hygiene, Water Pricing, Others
- **Energy:** Power Outages, Solar Energy, Wind Energy, Fuel Prices, Energy Access, Grid Infrastructure, Nuclear Energy, Energy Conservation, Coal & Mining, EV & Clean Transport, Others
- **Housing & Urban:** Affordable Housing, Slums & Informal Settlements, Real Estate Prices, Eviction & Displacement, Building Safety, Smart Cities, Public Spaces, Parking & Traffic, Street Lighting, Solid Waste in Cities, Others

- **Labour & Employment:** Job Creation, Unemployment, Child Labour, Worker Rights, Minimum Wage, Gig Economy, Migrant Workers, Workplace Harassment, Skill Development, Informal Sector, Others
- **Finance & Banking:** Loan Access, Loan Defaults & NPA, Insurance, Pension & Retirement, Digital Payments, Financial Fraud, Stock Market, Cryptocurrency, Microfinance, Financial Inclusion, Others
- **Transport & Mobility:** Railways, Air Travel, Metro & Rapid Transit, Road Freight, Last Mile Connectivity, Traffic Congestion, Cab & Ride Sharing, Cycling & Walkways, Toll & Taxes, Accident Black Spots, Others
- **Law & Justice:** Court Delays, Police Misconduct, Prison Conditions, Legal Aid, Domestic Violence, Property Disputes, Consumer Rights, Cyber Laws, Anti-Corruption Laws, Tribal & Indigenous Rights, Others
- **International & Geopolitical:** Border Disputes, Trade Relations, Diplomatic Relations, War & Conflict, Refugee Crisis, Sanctions & Embargoes, Climate Diplomacy, United Nations, Terrorism & Extremism, Foreign Aid, Others
- **Science & Innovation:** Medical Research, Space Research, Defence Technology, Agricultural Science, Clean Technology, AI & Robotics, Biotech & Genetics, Research Funding, Innovation Ecosystem, Open Source & Patents, Others
- **Mental Wellness & Spirituality:** Mindfulness & Meditation, Grief & Loss, Stress & Burnout, Therapy & Counselling, Spiritual Practices, Community Support Groups, Loneliness & Isolation, Addiction Recovery, Trauma & PTSD, Sleep & Rest, Others
- **Citizenship & Civic Participation:** Voter Registration, Civil Society & NGOs, Petitions & Protests, Community Volunteering, Political Participation, Public Consultation, Local Democracy, Transparency & RTI, Youth Civic Education, Electoral Reform, Others
- **Disability & Accessibility:** Physical Accessibility, Assistive Technology, Inclusive Education, Employment for PwD, Disability Benefits, Cognitive Disabilities, Visual Impairment, Hearing Impairment, Mental Health Disability, Disability Legislation, Others
- **Migration & Diaspora:** Internal Migration, Overseas Workers, Visa & Immigration Policy, Diaspora Engagement, Statelessness, Human Trafficking, Refugee Resettlement, Cultural Integration, Brain Drain, Remittances, Others
- **Climate Adaptation & Resilience:** Heatwave Preparedness, Coastal Erosion, Drought Management, Flood Mitigation, Urban Heat Islands, Climate Migration, Agricultural Adaptation, Early Warning Systems, Ecosystem Restoration, Community Resilience, Others
- **Digital Rights & Privacy:** Surveillance & Tracking, Freedom of Expression Online, Right to Be Forgotten, Biometric Data, Internet Shutdowns, Platform Accountability, Algorithmic Bias, Children Online Safety, Open Internet, Digital Identity, Others
- **Gender & Family:** Marriage & Divorce, Parenting & Child Custody, Domestic Work & Care, Reproductive Rights, Menstrual Health, Paternity & Parental Leave, Gender Pay Gap, Single Parenthood, Ageing Parents & Elder Care, Gender-Based Violence, Others
- **Food Systems & Nutrition Policy:** Malnutrition, Obesity & Diet Policy, Food Labelling, School Meals, Food Adulteration, Traditional & Indigenous Foods, Urban Food Deserts, Hunger & Famine Relief, Vegetarianism & Food Choices, Supplement & Health Claim Regulation, Others
- **Urban Liveability:** Air Quality in Cities, Green Spaces & Parks, Noise & Nuisance, Pet & Animal Welfare, Night-time Economy, Public Toilets & Sanitation, Community Spaces, Urban Food Vendors, Pedestrian Infrastructure, City Aesthetics, Others
- **Conflict & Peacebuilding:** Communal Violence, Caste Conflict, Land Conflict, Armed Insurgency, Reconciliation & Truth Commissions, Peacekeeping Operations, Post-Conflict Rehabilitation, Hate Speech & Incitement, Mediation & Dialogue, Extremism Prevention, Others