Large Language Models Are Zero-Shot Text Classifiers: A Reproduction Study on COVID-19 Sentiment Analysis

Smit Chavan, Swaraj Patil
K. J. Somaiya Institute of Technology, Mumbai

Introduction

Text classification is a core NLP task, powering sentiment analysis, topic categorization, and spam detection. Traditional ML models like Naive Bayes, SVM, and LSTM require labeled data, domain-specific tuning, and retraining for new tasks. With the advent of **Large Language Models (LLMs)** like GPT-4 and Llama2, text can now be classified using **Zero-Shot Learning (ZSL)**—through natural language prompts, without training or labeled data.
This study reproduces the work "LLMs are Zero-Shot Text Classifiers" to validate and compare GPT-based models on multi-domain text datasets.

Motivation and Research Questions

- Minimize manual labeling and model retraining.
- Evaluate if LLMs generalize across unseen domains.
- Compare GPT-4, GPT-3.5, and Llama2 to classical ML/DL.
- Investigate prompt phrasing effects on classification.

Background and Evolution

Rule-Based → Statistical → Deep → Zero-Shot:

- **Rule-Based:** Keyword rules, low flexibility.
- **Statistical:** Naive Bayes, Logistic Regression.
- **Deep Learning:** CNN, LSTM learn context but need training.
- **Transformers & LLMs:** Leverage pretraining for unseen text.

Applications of Zero-Shot Classification

- **Healthcare:** Identify misinformation or disease mentions in tweets.
- **Finance:** Gauge investor sentiment from news headlines.
- **E-Commerce:** Classify product reviews automatically.
- **Education:** Assess essay tone and student feedback.
- **Cybersecurity:** Detect phishing or fake messages.

Traditional vs. Zero-Shot Comparison

Traditional ML/DL Approach	LLM Zero-Shot Approach
Needs labeled datasets	Works without labels
Requires retraining	Adapts instantly via prompts
High training cost	One-time API call
Task-specific models	Universal text reasoning
Limited generalization	Understands unseen categories

Challenges and Limitations

- Accuracy highly depends on prompt structure.
- Risk of bias or verbose answers.
- Expensive for large-scale data due to token usage.
- Limited interpretability in reasoning process.

Methodology and Datasets

Traditional classification pipelines use preprocessing, TF-IDF, and model training. Zero-shot methods rely purely on natural-language inference using pre-trained LLMs.
Prompt Example:

```
"Classify the following tweet about COVID-19 as Positive, Negative, or Neutral."
```

Models Evaluated:

- **Machine Learning:** Naive Bayes, SVM, Random Forest
- **Deep Learning:** RNN, LSTM, GRU
- **LLMs:** GPT-3.5, GPT-4, Llama2

Datasets:

- **COVID-19 Tweets:** Sentiment (3-class)
- **E-Commerce Reviews:** 4 product categories
- **SMS Spam:** Spam vs Ham
- **Economic News:** Positive, Negative, Neutral tone

TABLE I
EXAMPLES OF COVID19 TWEETS WITH DIFFERENT SENTIMENT LABELS

Label	Tweet
Neural	Vistamalls says supermarket sales to 'balance' #COVID19 impact https://t.co/caE2rT6MhO
Negative	Just now on the telly, Woolies have stopped all online and click n collect orders. Due to overwhelming demand. #coronavirus #StopPanicBuying
Positive	Efforts 2 contain #COVID-19 are shifting demand & disrupting Ag supply chains. @raboresearch has collated our analysis of current & expected impacts in one place 2 help our @RabobankAU network keep informed. https://t.co/P41vjG4uD6

TABLE II
EXAMPLES OF E COMERCE TEXT WITH DIFFERENT LABELS

Label	E commerce Text
Clothing & Accessories	Cherokee by Unlimited Boys' Straight Regular Fit Trousers Cherokee kids beige trousers made of 100% cotton twill fabric.
Household	Nutella Hazelnut Spread with Cocoa, 290g Size:290g Because the taste is simply unique! The secret is its special recipe, the selected ingredients and the careful preparation. Here we want to tell you about Nutella and all the passion and care that we put in its production every day.
Books	NIACL Assistant Preliminary Online Exam Practice Work Book - 2280.
Electronics	Transcend 512 MB Compact Flash (TS512MCF300) Transcend's CF300 cards are high-speed industrial CF cards offering impressive 300X transfer rates. With matchless performance and durability, CF300 CF cards are perfect for POS and embedded systems that require both industrial-grade reliability and an ultra-high speed data transfer.

Sample Datasets and Data Structure

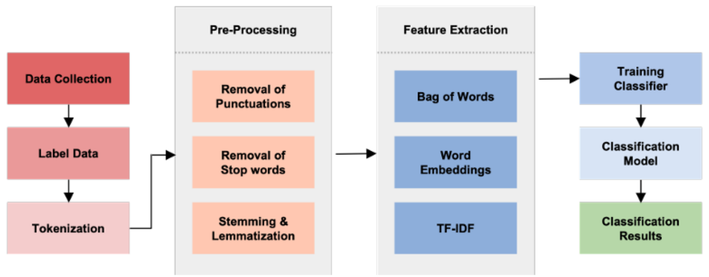

Fig. 1. Traditional text classification flow

Fig. 2. LLMs' zero shot text classification flow

Traditional vs. Zero-Shot Workflow

Evaluation Metrics: Accuracy, Precision, Recall, and F1-score. **Implementation:** GPT models queried via API (temperature = 0.01) for stable, deterministic output.

Results and Analysis

Overall Results:

- **GPT-4:** 97.33% (SMS), 90% (E-Commerce), 71.33% (Economic)
- **GRU:** 98.67% (SMS) – top trained model
- **GPT-3.5/Llama2:** Reliable across balanced data
- LLMs outperform ML/DL in 3 of 4 datasets

Key Observations:

- GPT-4 achieves the best zero-shot generalization.
- Prompt clarity impacts accuracy more than dataset size.
- GPT-3.5 and Llama2 remain competitive alternatives.
- Zero-shot models outperform traditional fine-tuned transformers.

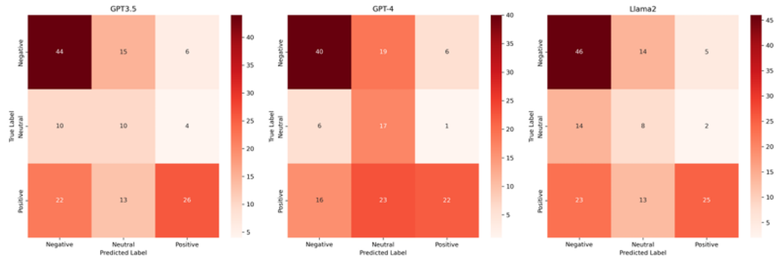

Fig. 3. The confusion matrices for LLMs' classification results in COVID19 tweets.

Confusion Matrices for GPT-3.5, GPT-4, and Llama2

Accuracy Comparison Table:

Model	COVID-19	E-Com	SMS	Economic
SVM	72.00	85.33	95.33	68.00
LSTM	76.00	87.33	97.33	69.33
GRU	78.67	88.00	**98.67**	70.67
GPT-3.5	81.33	89.33	96.00	70.67
GPT-4	**85.33**	**90.00**	97.33	**71.33**
Llama2	82.00	88.67	95.67	70.00

Interpretation: LLMs, especially GPT-4, demonstrate near-human zero-shot reasoning. They outperform baseline models without training, proving their capacity to generalize to unseen text tasks efficiently.

Conclusion and Future Work

Conclusion:

- LLMs provide robust zero-shot performance across domains.
- GPT-4 matches or exceeds trained deep learning models.
- The approach minimizes cost, data preparation, and training time.

Future Work:

- Extend analysis to multilingual datasets.
- Incorporate Chain-of-Thought reasoning for explainability.
- Develop prompt-optimization frameworks for adaptive inference.
- Combine LLM predictions with rule-based filters for trust assurance.