Comparative Analysis of Statistical and Transformer-Based Models for Product Title Quality Estimation

Aarib Ahmed Vahidy
Dept. of Artificial Intelligence
FAST NUCES
Karachi, Pakistan
22K-4004

Partham Chawla
Dept. of Artificial Intelligence
FAST NUCES
Karachi, Pakistan
22K-4079

Arham Hussain Khan
Dept. of Artificial Intelligence
FAST NUCES
Karachi, Pakistan
22K-4080

***Abstract*—In the rapidly expanding domain of e-commerce, the quality of product listings—specifically the clarity and conciseness of titles—directly impacts user experience and conversion rates. This paper addresses the CIKM AnalytiCup 2017 challenge, aiming to automatically estimate clarity and conciseness scores for Lazada product titles. While modern Natural Language Processing (NLP) has shifted heavily towards Transformer-based architectures, this study investigates whether resource-intensive Deep Learning models offer tangible benefits over lightweight statistical approaches for short, noisy text in resource-constrained environments. We compare a bagged LightGBM model utilizing character-level n-grams against a DistilBERT-based feature extraction pipeline. Our experiments demonstrate that the statistical baseline (RMSE: 0.5293 for Clarity) outperforms the frozen Transformer approach (RMSE: 0.5303), suggesting that structural features are more predictive of "clarity" than semantic embeddings in this specific domain. We further explore an ensemble strategy that marginally improves performance, highlighting the trade-off between computational cost and predictive accuracy.**

***Index Terms*—Natural Language Processing, Text Regression, LightGBM, Transformers, DistilBERT, Product Quality Estimation**

I. Introduction

The exponential growth of e-commerce platforms has led to an influx of user-generated content, often resulting in product titles that are spammy, repetitive, or unintelligible. The CIKM AnalytiCup 2017 challenge identified "Clarity" (coherence) and "Conciseness" (lack of redundancy) as two critical metrics for assessing product title quality. Automating the evaluation of these metrics is essential for maintaining catalog hygiene and ensuring a positive customer experience.

Existing literature in text classification largely favors Transformer-based architectures (e.g., BERT, RoBERTa) due to their state-of-the-art performance on semantic tasks. However, a significant *research gap* exists in understanding the efficiency-accuracy trade-off of these models for *structural* quality estimation in resource-constrained environments (e.g., local machines without GPUs). Unlike sentiment analysis, where "meaning" is paramount, clarity estimation often relies on surface-level features like punctuation, capitalization, and formatting—patterns that traditional n-gram models capture effectively.

This paper proposes a comparative analysis between a traditional Gradient Boosting Machine (LightGBM) using character n-grams and a modern Transformer (DistilBERT) approach. We aim to determine if the computational cost of Transformers is justified for quality estimation tasks where surface structure may be more important than deep semantics.

II. Related Work

Early approaches to short-text quality estimation relied heavily on feature engineering, utilizing lexical features, part-of-speech (POS) tags, and readability indices (e.g., Flesch-Kincaid). In the CIKM 2017 competition, top-performing solutions frequently employed ensemble methods (XGBoost, Random Forests) combined with character-level n-grams to handle the noisy, multilingual nature of e-commerce data.

The advent of the Transformer architecture [1] revolutionized NLP by enabling models to capture long-range dependencies. BERT (Bidirectional Encoder Representations from Transformers) [2] and its lighter variant, DistilBERT [3], have set benchmarks across the GLUE leaderboard. However, recent studies suggest that for specific regression tasks involving noisy, short text, well-tuned statistical models can remain competitive with Deep Learning approaches, particularly when pre-trained models are applied without domain-specific fine-tuning [4].

III. Methodology

A. Dataset Description

We utilize the CIKM AnalytiCup 2017 dataset provided by Lazada. The data consists of product titles, descriptions, and category metadata. The target variables are 'Clarity' and 'Conciseness', provided as binary labels (0 or 1). For validation, we utilize the Root Mean Squared Error (RMSE) metric, treating the problem as a regression task to output continuous probability scores.

B. Data Preprocessing

We implemented two distinct preprocessing pipelines to suit the requirements of each model:

- **Statistical Pipeline:** Text was converted to lowercase. HTML tags and special characters were aggressively removed to isolate alphanumeric patterns.
- **Transformer Pipeline:** HTML tags were removed, but punctuation and numerical values were retained, as DistilBERT relies on these tokens for contextual understanding.

C. Feature Extraction

For the statistical model, we utilized `CountVectorizer` to generate character-level n-grams (range 2-6), limiting the vocabulary to the top 5,000 features. This captures morphological patterns (e.g., "50ml", "100%").

For the Deep Learning model, we utilized the `distilbert-base-uncased` tokenizer. Due to hardware constraints (CPU environment), we employed a *Feature Extraction* strategy rather than end-to-end fine-tuning. We extracted the 768-dimensional embedding vector corresponding to the `[CLS]` token for each input sequence.

D. Model Architecture

1) **Baseline (LightGBM):** We implemented a Light Gradient Boosting Machine (LightGBM) regressor. To reduce variance, we employed a bagging strategy with 4 iterations and 10-fold cross-validation.
2) **Transformer Approach:** The extracted DistilBERT embeddings were fed into a LightGBM regressor to capture non-linear relationships within the semantic vector space.
3) **Ensemble:** A weighted average ensemble was created, assigning a weight of $\alpha = 0.7$ to the Baseline and $(1 - \alpha) = 0.3$ to the Transformer predictions.

IV. Experiments and Results

We evaluated the models on the provided validation set. The performance is measured using Root Mean Squared Error (RMSE), where a lower score indicates better performance (closer to the ground truth).

TABLE I
PERFORMANCE COMPARISON (RMSE)

Model Type	Clarity	Conciseness
Baseline (Char N-Grams + LightGBM)	**0.52931**	**0.35273**
Transformer (DistilBERT + LightGBM)	0.53035	0.35992
Ensemble (Weighted Average)	0.52924	0.35274

As observed in Table I, the Baseline model outperformed the Transformer approach. The Ensemble method provided a marginal improvement of 0.00007 in Clarity but did not improve Conciseness.

V. Discussion

A. Analysis of Results

The superior performance of the character n-gram baseline suggests that "Clarity" in e-commerce titles is largely defined by structural coherence rather than deep semantic meaning. Character n-grams effectively capture "visual" noise—such as broken spelling, excessive capitalization, or strange formatting—which directly correlates with a human's perception of clarity. Conversely, pre-trained Transformers are optimized to understand *meaning* (e.g., relating "king" to "queen"), which is less relevant when judging if a title is simply "messy."

B. Limitations

The primary limitation of this study was the hardware constraint (lack of GPU acceleration). Consequently, we relied on **frozen embeddings** from DistilBERT. Had we been able to perform full **fine-tuning** (updating the weights of the Transformer layers), the Deep Learning model would likely have learned to identify the specific structural patterns of the dataset and potentially outperformed the baseline.

VI. Conclusion

This paper presented a comparative study of statistical and Deep Learning methods for product title quality estimation. We demonstrated that for specific tasks involving short, noisy text and structural assessment, a lightweight LightGBM model with character n-grams can outperform complex Transformer architectures, especially when computational resources prevent full fine-tuning. Future work will focus on utilizing GPU resources to fine-tune RoBERTa models and exploring stacking ensembles to better leverage the strengths of both approaches.

References

[1] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017.
[2] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[3] V. Sanh et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
[4] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Adv. Neural Inf. Process. Syst.*, 2017.