Are modern text encoders more robust to textual noise?

Cristiana Cocheci, Apostol Ilie-Daniel
University of Bucharest, Bitdefender, Romania

cristiana.cocheci@gmail.com, apostoldaniel854@gmail.com

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

Introduction

Our study investigates whether ModernBERT maintains higher accuracy than BERT when tested on perturbed inputs, specifically examining:

1. Performance comparison on clean training data but **noisy test inputs**
2. Effects of **noise-aware training** on model robustness
3. **Generalization** capabilities across **different noise levels**

We evaluate both models on a classification task using perturbed text, providing both quantitative results and qualitative analyses of how these models process noisy text.

Dataset

Dataset Description: For our experiments, we utilized the IMDB movie review dataset, a sentiment analysis dataset consisting of user reviews with **binary sentiment labels** (positive/negative). The dataset was chosen for its real-world applicability and the natural presence of linguistic variations.

Data preprocessing: We preserved the original dataset in its entirety, while augmenting it using three targeted perturbation techniques to support our research objectives. Specifically, we applied typo generation, synonym replacement, and word dropout to assess model robustness. These methods were chosen to simulate common real-world noise and linguistic variation, thereby providing a more comprehensive evaluation of model performance. Each perturbed example has 20% of the words perturbed.

There are also two types of perturbed train datasets:

1. **All perturbations**: all reviews are perturbed with one of the three perturbations methods chosen randomly.
2. **Typo perturbations**: Five train datasets where 2% / 5% / 10% / 20% of reviews are perturbed by typos and the other samples are clean.

Models

Using the transformers library from Hugging Face we have taken the "**bert-base-uncased**" and "**answerdotai/ModernBERT-base**" pretrained models and finetuned them on our custom datasets.
We passed the original text through the matching pretrained tokenizers from the same library and then trained on the *train* - *validation* split (80%-20%) for 3 epochs.

We also used **Captum**, a model interpretability library. This tool allows us to visualize the influence of words on predictions, with red highlighting negative and green positive sentiment.

Results

Figure 1 compares performance across perturbation types with models trained on clean (left) versus perturbed data (right). ModernBERT performs better in all scenarios, with its biggest advantage seen with typos. It also degrades less on perturbed inputs. Moreover, noise-aware training (right side) closes performance gaps for both models on perturbed data.

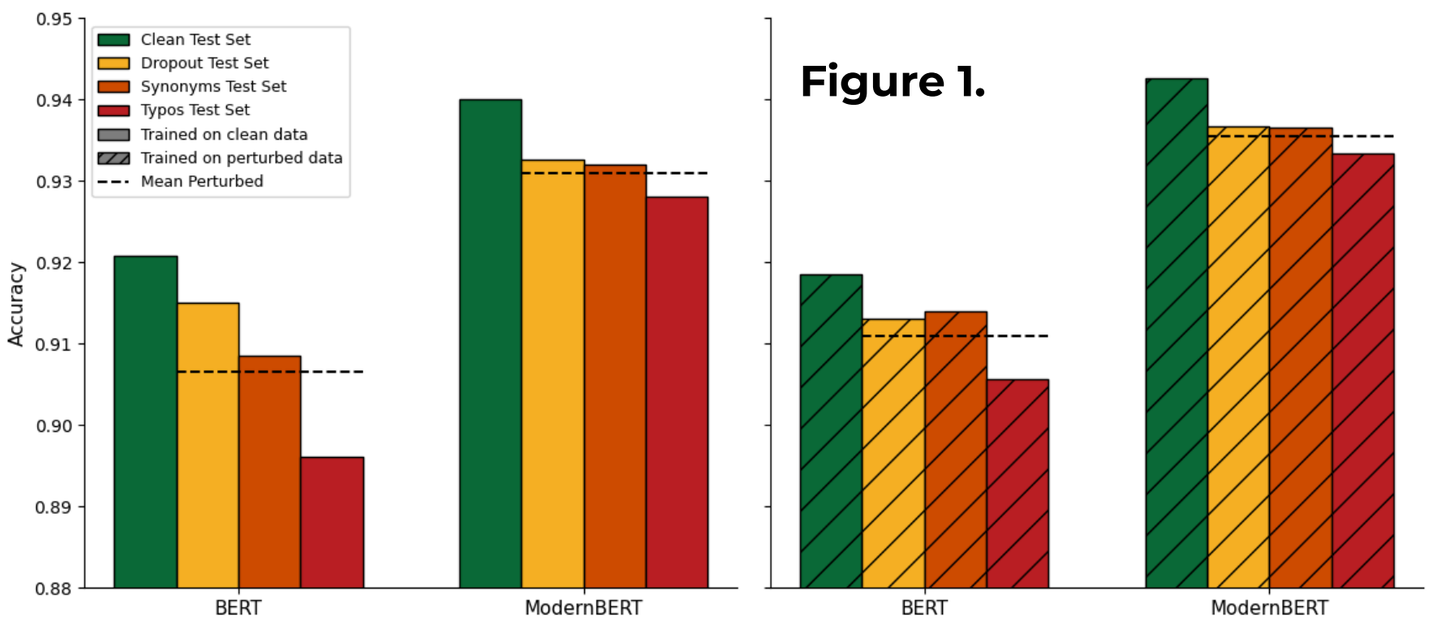

Figure 1.

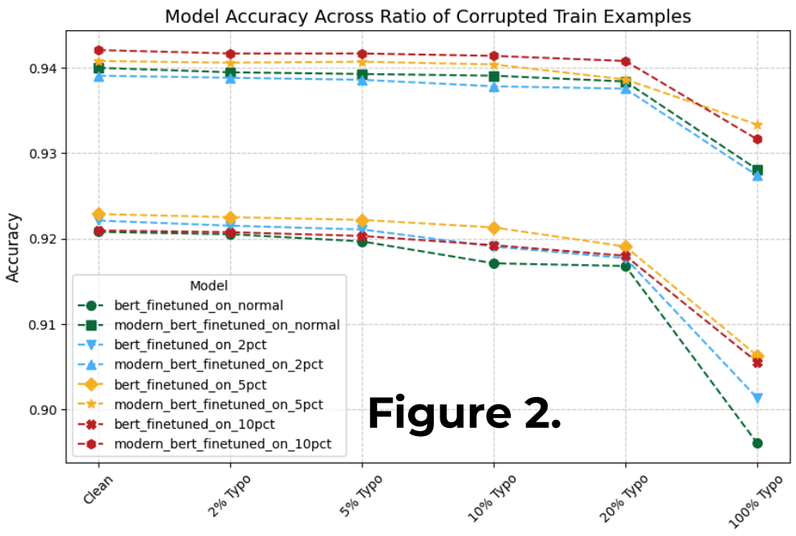

Figure 2.

Figure 2 shows ModernBERT consistently outperforming BERT across all typo corruption levels. Training with just 2-5% noise significantly improves both models' resilience to corruption, with ModernBERT maintaining its advantage even at 100% corruption.

Qualitative analysis

In **Figure 3**, a typo review, ModernBERT identifies terms such as *non-existent* and *hodrorr* (typo for horror) as conveying negative sentiment, whereas BERT fails.

Even though some negative words such as *nauseating* have a positive attribution score for ModernBERT, they have a small overall impact on the sentence attribution score, leading to a correct classification of the review.

MODEL	Legend: Negative / Neutral / Positive — True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
BERT-FP	0	1 (0.99)	1	4.80	is prc ##tty much non - existent nausea ##ting ho ##dr ##ror
ModernBERT-TN	0	0 (0.76)	1	-0.93	is prctty much non-existent nauseating hodrror![SEP]

Figure 3.

In **Figure 4** we observe a FN from Bert (that ModernBERT classified correctly). The typos of *family* and *enjoy* have strongly impacted the prediction.
However, when we replace the typos with the original words, Bert is able to predict correctly.

MODEL	Legend: Negative / Neutral / Positive — True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
BERT-FN	1	0 (0.88)	1	-2.98	[CLS] my fa ##jm ##ily and i en ##jo ##hy this show
BERT-TP	1	1 (1.00)	1	5.18	[CLS] my family and i enjoy this show

Figure 4.

Conclusion

ModernBERT demonstrates **significantly greater robustness** to various types of input perturbations compared BERT. Its performance degradation remains minimal when evaluated on datasets containing a moderate proportion of altered samples <20%. Both models **benefit from noise-aware training**, with ModerBERT consistently outperforming BERT.