Final Project Proposal

1. Team Members and Roles

1. **Yuyang Ding -** designs the testing strategy, looks for reliable datasets, computes performance and robustness metrics, and helps to visualize results.
2. **Ziqi Gu** - is responsible for experimenting and evaluating MiniMind, a decoder-only Transformer-based model, and conducting comparative robustness experiments.
3. **Shengyang Tao** - sets up data pipeline, trains the rest models (TF-IDF, BERT, and its distilled version), and maintains the experiment codebase.
4. **Alissa Hsu** - manages paper writing, ensures readability and style, organizes related work, diagrams, tables, and data visualization, as well as references.

2. Problem Statement

Short text topic classification is one of the most fundamental tasks in NLP, and is of great importance in downstream applications such as recommendation systems.

Traditional methods, like TF-IDF, are unable to capture context due to their bag-of-words architecture. Modern models, such as GPT and BERT, utilize self-attention mechanisms, leading to better performance thanks to their capability to understand context, but also resulting in heavy computational burdens and vulnerability to noise such as typos and grammatical issues.

Hence, this paper compares classical and modern architectures, focusing on their correctness, efficiency, and robustness. We aim to explore the differences between different architectures, whether we can achieve comparable performance with fewer resources used, and how robust these models are in the face of noisy input.

3. Related Work

Zhang et al. (2015) – *Character-Level Convolutional Networks for Text Classification*.
– This paper proposed the idea of learning directly from characters rather than words, showing that character-level models are more robust to misspellings and noise.
Therefore, this paper inspires us with the idea of robustness testing with intentionally noisy text.

Vaswani et al. (2017) – *Attention Is All You Need*.
– This work introduced the Transformer architecture, which replaced recurrence with self-attention and became the foundation of modern contextual models.
Therefore, this paper provides the structural backbone for both BERT and MiniMind.

Devlin et al. (2019) – *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

– BERT defined the standard encoder architecture for contextual embeddings and pretraining objectives (Masked LM, Next Sentence Prediction).
Therefore, this paper serves as our main contextual baseline model.

Sanh et al. (2019) – *DistilBERT: A Distilled Version of BERT*
– DistilBERT demonstrated that model distillation can preserve most of BERT's accuracy with even fewer parameters.
Therefore, this paper motivates our efficiency experiments with MiniMind, which applies similar principles at a larger scale.

Mamata Das et al. (2023) – *A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset*

– This paper discusses TF-IDF's strengths and limitations for text classification, showing that even though TF-IDF is simple and not advanced, it remains a competitive baseline.
Therefore, this paper supports our use of TF-IDF and Logistic Regression as the non-contextual benchmark.

4. Evaluation Plan

Task

We use the AG News as our dataset, which contains 4 balanced categories (*World, Sports, Business, Sci/Tech*).
Each model classifies a given news headline and short description into one of these topics.

Data Splits

- **Training set:** 120,000 samples
- **Validation set:** 7,600 samples
- **Test set:** 7,600 samples
  A **noisy test version** is created by adding misspellings, mixed capitalization, and emojis to test robustness.

Metrics

To evaluate our systems comprehensively, we focus on three dimensions: performance, robustness, and efficiency.

For performance, we use Accuracy and Macro-F1, which measure the overall classification quality and balance across all four topic categories.

To evaluate robustness, we compute ΔAccuracy (Clean – Noisy), representing how much each model's performance drops when facing noisy or perturbed input sentences.

Finally, for efficiency, we record the inference time per sample and the total parameter count of each model to quantify both computational speed and model size.

We ensure the validity of our experiments by fixing random seeds, repeating runs multiple times, and keeping the test data strictly separate from training and development sets.

5. Strategy for Solving the Problem

We will first implement a non-contextual TF-IDF model as a baseline. Then, we will finetune the modern models of transformer architectures, namely BERT and Minimind. Due to limited resources, we will test on a small scale of around 100M parameters.

After training, we will test their robustness on noisy inputs, such as the ones with spelling errors or grammatical issues.

In addition to that, for efficiency, we will explore whether distilling BERT is able to achieve comparative performance using less resources.

The evaluation metric includes accuracy, F1 score, and time used per reference.

Finally, we will conduct a brief manual error analysis to look for common failure patterns to reveal the sensitivity characteristics of each model.