abstract The increasing need for transformer-based models in natural language processing (NLP) is rich in advantages in language generation and understanding tasks, but their expensive computational requirement, memory size, and power needs usually prevent deployment [1, 2]. We mitigate these drawbacks by proposing CE-BERT Lite, an optimized imple- mentation of the CE-BERT model [3] using the Hugging Face Transformers Library [4]. CE-BERT Lite utilizes a two-pipeline optimization process: knowledge distillation [5] and dynamic post-training quantization [6]. In comparison to its teacher model, CE-BERT Lite has a marginal 1.3% decrease in accuracy, while providing an inference speedup of 3.6×, a 75% model size reduction (from 420 MB to 25 MB), and considerable energy saving (0.08 kWh vs. 0.32 kWh) [3]. Utilizing a TinyBERT-like student model [7], we fine-tuned CE- BERT Lite on soft (KL-divergence) and hard (cross-entropy) targets, optimized on Google Colab with T4 GPU, with batch size of 8, 2000 training samples, and 3 epochs. Bench- marked on public datasets such as IMDb [8], Twitter Sentiment140 [9], TREC Question Classification [10], and Yelp Polarity [11], CE-BERT Lite shows strong generalization with more than 88% accuracy and a peak of 95.4% on Yelp Polarity. Building on existing work in model compression, e.g., DistilBERT [12], TinyBERT [7], and MobileBERT [13], CE-BERT Lite addresses domain-level issues in noisy, colloquial text, making it suitable for real-time, resource-scarce settings like mobile platforms and edge computing [15]. Ex- perimental findings, which were carried out on Google Colab using T4 GPU acceleration, place CE-BERT Lite as an eco-effective, high-performing solution to green NLP applica- tions [16]. The code and model will be published openly for reproducibility. Keywords: Natural Language Processing, Transformer Models, Knowledge Distillation, Dy- namic Quantization, CE-BERT Lite, Hugging Face 1