Interactive Analysis of Sentiment Retention under BERT Embedding Compression and Deletion

Dylan OBrien
School of Computing & Data Science
Wentworth Institute of Technology
Boston, MA
obriend7@wit.edu

***Abstract*— This study investigates the retention of sentiment information in BERT-based representations under two complementary compression strategies: dimensionality reduction in embedding space and token-level input deletion. Using the IMDb 50K movie reviews dataset and a frozen *bert-base-uncased* encoder, each review is mapped to a 768-dimensional sentence embedding, on top of which a logistic regression classifier is trained for binary sentiment prediction. Embedding-space compression is implemented via Principal Component Analysis (PCA), applied in a project–reconstruct fashion at multiple target dimensions, while input-side compression removes low-information tokens using an inverse document frequency (IDF)–based deletion scheme that preserves only a configurable fraction of tokens. The system is evaluated on the IMDb test split using accuracy and macro-averaged F1, and is complemented by an interactive Gradio interface that allows users to manipulate reviews, adjust compression levels, and visualize changes in sentiment probabilities and approximate storage requirements. Results indicate that moderate PCA compression preserves most of the baseline performance, whereas aggressive token deletion degrades accuracy more sharply, illustrating a trade-off between efficiency and sentiment retention. The work concludes with a discussion of limitations, ethical considerations, and directions for extending the framework to other settings and compression methods.**

Keywords—Sentiment analysis, BERT, dimensionality reduction, PCA, token deletion, model compression, Gradio.

I. Introduction

Sentiment analysis is a central task in natural language processing (NLP), concerned with automatically categorizing subjective text as expressing positive, negative, or neutral opinions. Typical applications include monitoring user satisfaction in product reviews, tracking public opinion in social media, and supporting content moderation in online platforms. In many of these scenarios, sentiment models must operate at scale while maintaining reliable predictions under tight constraints on computation, memory, and storage.

Transformer-based models, and BERT in particular, are now widely used for sentence- and document-level sentiment analysis because they provide context-sensitive representations that capture subtle polarity cues and long-range dependencies [1]. When fine-tuned on sentiment datasets, BERT-based models frequently outperform classical bag-of-words or recurrent architectures [3]–[5]. However, BERT-based systems are computationally expensive and memory-intensive. Each text input is mapped to a high-dimensional embedding (e.g., 768 floating-point values for *bert-base-uncased*), and long sequences further increase the cost of both inference and storage.

Recent work therefore explores a variety of compression strategies for transformer models and their embeddings, including low-rank approximations, knowledge distillation, dimensionality reduction, and token pruning [6]–[10]. These techniques often report substantial reductions in model size and inference cost with modest drops in accuracy, but the concrete impact on downstream tasks such as sentiment analysis, especially in terms of how well sentiment information is retained under compression, remains an active area of study.

This project focuses on sentiment classification for IMDb movie reviews using a BERT-based pipeline and systematically examines how well sentiment is retained under two forms of compression. The first operates in embedding space and uses Principal Component Analysis (PCA) to reduce dimensionality before classification [6]–[8]. The second acts on the input side and deletes tokens with low inverse document frequency (IDF), thereby shortening the review before it is encoded. The analysis combines quantitative evaluation on a held-out test set with an interactive Gradio interface that exposes the entire pipeline to end users, allowing inspection of individual examples, compression settings, and resulting predictions.

The main contributions of this work are threefold. First, a simple but rigorous evaluation is provided of sentiment retention under PCA-based embedding compression and IDF-based token deletion using a BERT feature extractor and logistic regression classifier. Second, an interactive "Sentiment Compression Playground" interface is constructed to visualize the effects of compression on both sentiment probabilities and approximate storage usage. Third, the study documents practical trade-offs between accuracy, compression strength, and storage, and discusses the limitations and ethical considerations that arise when altering text or representations in sentiment applications.

II. Related Work

BERT (Bidirectional Encoder Representations from Transformers) was introduced as a deeply bidirectional transformer architecture pre-trained on large corpora using masked language modeling and next-sentence prediction [1]. The Hugging Face Transformers library [2] has popularized

BERT and related models by providing accessible implementations for a wide range of NLP tasks. Multiple studies demonstrate that BERT-based models attain strong performance on sentiment benchmarks such as IMDb and Twitter sentiment datasets, often surpassing traditional machine learning and earlier neural approaches [3]–[5]. Domadula and co-authors, for example, report that BERT-based architectures outperform bag-of-words and LSTM baselines on IMDb movie reviews [3], while Semary et al. present a RoBERTa-based hybrid model that achieves improved sentiment classification accuracy across several datasets [4]. Maltoudoglou et al. combine BERT with conformal prediction to provide calibrated confidence intervals for sentiment predictions [5].

As contextual embeddings are typically high-dimensional, there is growing interest in dimensionality reduction for text embeddings. Huertas-García et al. evaluate PCA, t-SNE, UMAP, and related techniques applied to text embeddings, showing that PCA offers a favorable trade-off between computational simplicity and preservation of meaningful structure in embedding space [6]. Zhang et al. conduct a systematic comparison of unsupervised dimensionality reduction methods for sentence embeddings and report that PCA often performs competitively or better than more complex nonlinear approaches, while remaining easy to integrate into existing pipelines [7]. Abdelmotaleb et al. investigate multiple word embedding techniques for sentiment analysis and observe that PCA-based compression can reduce dimensionality and noise without severely degrading classification performance [8].

A complementary strand of research focuses on reducing the number of tokens processed by transformer models. Learned Token Pruning (LTP) adaptively removes low-importance tokens at intermediate layers based on learned importance scores, yielding speedups in inference while maintaining accuracy on various NLP benchmarks [9]. Broader surveys of token pruning highlight the generality of these ideas across text, vision, and speech tasks, emphasizing that careful removal of redundant tokens can significantly reduce computation [10]. These approaches typically learn pruning rules jointly with the model, whereas the present project adopts a static IDF-based deletion scheme inspired by classical TF–IDF weighting in information retrieval.

Interactive interfaces for model inspection have become increasingly prominent in the ML community. Gradio provides a lightweight framework for building browser-based front ends to machine learning models and supports integration with notebooks and remote sharing [14]. By exposing the inputs, parameters, and outputs of models in an interactive fashion, such tools facilitate both pedagogical use and practical debugging. The present work follows this trend by constructing an interactive interface around the compression pipeline so that users can directly explore the effect of different compression settings on sentiment predictions and storage.

III. Methodology

The overall system architecture consists of three stages: embedding extraction with a frozen BERT encoder, compression in either embedding space or token space, and sentiment classification with a logistic regression output layer. All components are implemented in a single, reproducible notebook , and the final system is wrapped in a Gradio interface for interactive use.

In the first stage, a pre-trained *bert-base-uncased* model is loaded from the Transformers library [1], [2] and used as a fixed feature extractor. Each review is tokenized using the corresponding WordPiece tokenizer, with special tokens added and sequences truncated or padded to a set maximum length. The final-layer hidden state corresponding to the [CLS] token is extracted as a 768-dimensional sentence embedding. These embeddings are detached from the computation graph and stored as NumPy arrays for subsequent training and analysis. The embedding extraction logic is implemented in Cell 6 and used throughout the notebook.

In the second stage, two alternative compression strategies are applied. PCA-based compression operates directly in embedding space. PCA models are fit on the training embeddings for several target dimensions, for example $k \in \{32, 64, 128, 256, 384\}$. For a given k, a training embedding $x \in$ R^768 is projected to a lower-dimensional vector $z \in$ R^k using the PCA transformation and then reconstructed back to $\hat{x} \in$ R^768 via the PCA inverse transform. At inference time, test embeddings are processed in the same way, and the reconstructed embeddings are passed to the classifier. This project–reconstruct structure allows the classifier to remain unchanged and isolates the effect of dimensionality reduction on the information available for sentiment prediction. PCA training and reconstruction helpers are defined in Cells 7 and 8.

The second compression strategy acts on the input side using IDF-based token deletion. Document frequency counts are estimated for tokens over the training split, and an IDF value is computed for each token according to idf(t) = log(N/(df(t)+1)), where N is the number of training documents and df(t) is the number of documents containing token t. For a given review, tokens are ranked by IDF, and a keep ratio $r \in (0, 1]$ specifies the fraction of highest-IDF tokens to retain, up to a configured maximum number of tokens. Low-IDF tokens, which tend to be frequent and less discriminative, are removed first. The remaining tokens are reassembled into text and re-encoded by BERT, yielding a compressed sentence embedding. This approach is conceptually similar to static token pruning and aligns with the emphasis on discarding low-information tokens in the broader literature [9], [10]. IDF computation and deletion are implemented in Cells 9–11.

In the final stage, a logistic regression classifier is trained on the full 768-dimensional BERT embeddings using scikit-learn [13]. The classifier is configured with a sufficiently large maximum number of iterations to ensure convergence and is trained on the training split with an internal validation set for basic hyperparameter choices. For PCA-based compression, the same classifier is applied to reconstructed embeddings without retraining. For IDF-based deletion, the classifier is applied to embeddings of compressed reviews, again with identical classifier parameters. This design isolates the influence of the compression strategies themselves rather than confounding them with changes to the classifier. Training and evaluation for each representation are implemented in Cell 13.

IV. Dataset

The study uses the IMDb 50K Movie Reviews dataset, a widely used benchmark for binary sentiment classification [11], [12]. The dataset contains 50,000 English-language movie reviews obtained from IMDb, each labeled as positive or negative. The data are balanced with 25,000 positive and 25,000 negative reviews. To support plain train–test evaluation, the dataset is partitioned into a training split of 25,000 reviews and a test split of 25,000 reviews, with no overlap in movies between splits [12].

Within this project, the training split is further divided into a training portion and a validation portion. The training portion is used to estimate IDF statistics, compute BERT embeddings, and fit PCA models and the logistic regression classifier. The validation portion is used to select basic hyperparameters, such as the maximum sequence length and regularization strength for logistic regression, while the original 25,000-review test split serves as the primary benchmark for all compression settings. Data loading and splitting are performed in Cell 4 of the notebook.

To support later analysis, all experiments share the same train/validation/test partitioning, ensuring that comparisons between compression strategies are fair and directly comparable. The distribution of labels remains balanced across splits, reflecting the original design of the IMDb benchmark [11], [12].

V. Experimental Setup and Evaluation

The experimental setup is designed to quantify how well sentiment information is retained under PCA-based embedding compression and IDF-based token deletion, relative to a baseline model that uses full BERT embeddings. All hyperparameters are chosen to keep the runtime within manageable limits while maintaining a realistic sentiment classification pipeline.

For the baseline configuration, BERT embeddings are extracted for each review in the training, validation, and test splits, and a logistic regression classifier is trained on the training embeddings. The validation split is used to verify that the classifier converges and to give an initial estimate of expected performance. Performance on the test split is then measured in terms of accuracy and macro-averaged F1. These baseline metrics, along with metrics for each compressed representation, are recorded in a single results DataFrame *plot_results_df*, which merges classification metrics with semantic preservation statistics from cosine similarity analysis. This unified table is produced in Cell 16.

	method	val_accuracy	val_f1	test_accuracy	test_f1	method_type	pca_dim	keep_ratio	approx_size	mean_cosine_train
0	bert_full	0.802000	0.801941	0.7995	0.799458	bert_full	NaN	NaN	768.0	NaN
1	pca_k16	0.758667	0.758649	0.7350	0.734894	pca	16.0	NaN	16.0	0.965409
2	pca_k32	0.771333	0.771216	0.7635	0.763490	pca	32.0	NaN	32.0	0.976180
3	pca_k64	0.785333	0.785253	0.7815	0.781454	pca	64.0	NaN	64.0	0.985273
4	pca_k128	0.778667	0.778560	0.7865	0.786455	pca	128.0	NaN	128.0	0.992681
5	pca_k256	0.794667	0.794567	0.8020	0.801967	pca	256.0	NaN	256.0	0.997438
6	delete_0.25	0.718000	0.717982	0.7160	0.715886	deletion	NaN	0.25	192.0	0.787540
7	delete_0.5	0.759333	0.759312	0.7680	0.767972	deletion	NaN	0.50	384.0	0.876463
8	delete_0.75	0.781333	0.781333	0.7915	0.791494	deletion	NaN	0.75	576.0	0.932577

Table 1. Unified classification and semantic preservation results for full BERT, PCA-compressed, and IDF-deletion methods.

For embedding-space compression, PCA models are fit on the training embeddings for several dimensionalities, and for each value of k, test embeddings are transformed and reconstructed as described earlier. The logistic regression classifier trained on full embeddings is then applied to the reconstructed embeddings without retraining. This yields test accuracy and macro-averaged F1 scores for several degrees of compression, as well as mean cosine similarity between original and reconstructed embeddings on the training set. A dedicated visualization in the notebook plots test F1 and mean cosine similarity as functions of k, making the trade-off between performance and embedding distortion explicit. This plot is generated by the PCA summary code in Cell 22.

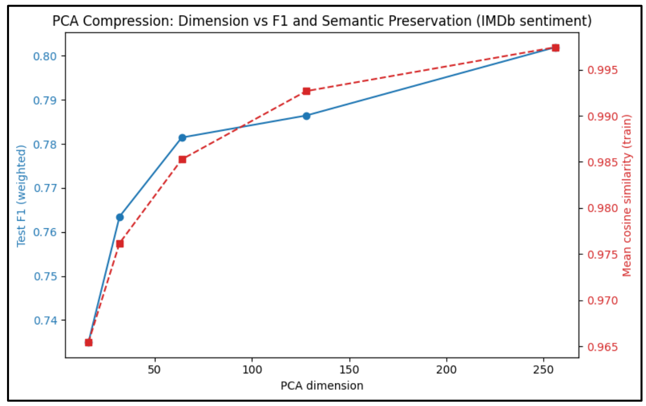

Fig. 1. PCA compression: test F1 and mean cosine similarity as functions of PCA dimension.

For input-side compression, several keep ratios r are tested. For each keep ratio, compressed reviews are generated by IDF-based deletion and re-encoded with BERT. The logistic regression classifier is then applied to the resulting embeddings. Test accuracy, macro-averaged F1, and mean cosine similarity between original and compressed embeddings are recorded in the unified results table. A second visualization plots test F1 and mean cosine similarity as functions of the keep ratio, providing a direct counterpart to the PCA plot and enabling comparison between the two compression strategies. This plot is produced in Cell 23.

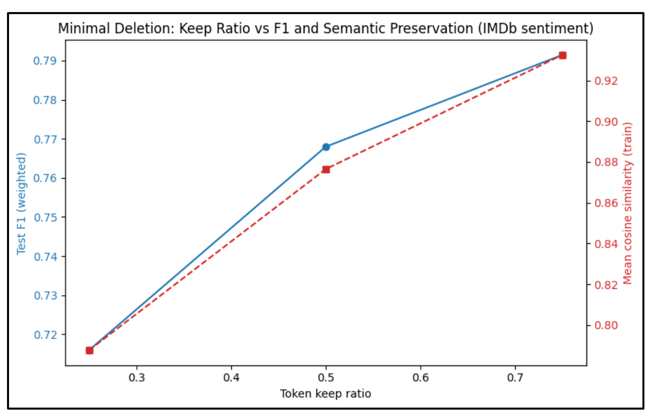

Fig. 2. Minimal deletion: test F1 and mean cosine similarity as functions of token keep ratio.

Because sentiment analysis is a classification problem, confusion matrices offer additional insight into the types of errors made by the baseline and compressed models. A dedicated cell identifies the top-performing PCA and deletion configurations based on test F1, and then trains a classifier for each such configuration alongside the baseline. The resulting confusion matrices are plotted as separate heatmaps, allowing inspection of how compression affects the balance of false positives and false negatives in each case. This analysis is implemented in Cell 21.

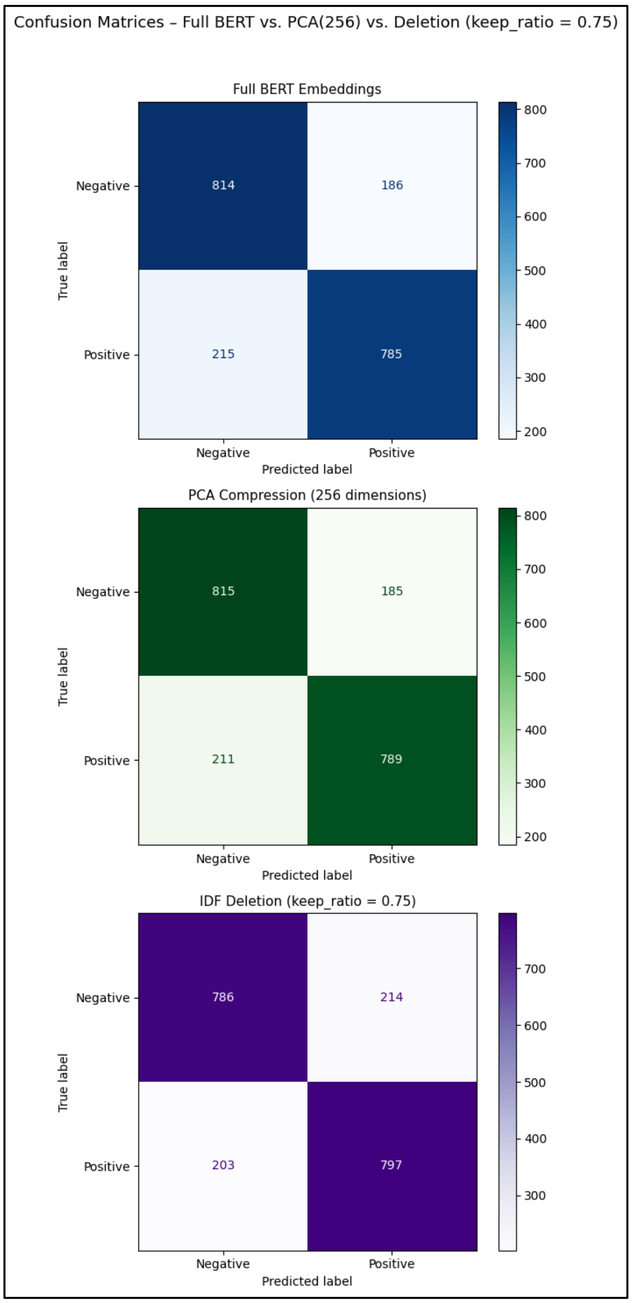

Fig. 3. Confusion matrices for full BERT embeddings and top PCA/deletion configurations on the IMDb test set.

Beyond scalar metrics and confusion matrices, the notebook estimates approximate storage requirements for each configuration. For a given review, text size is approximated by the number of bytes in its UTF-8 encoding, and embedding size is approximated as the number of float32 values multiplied by 4 bytes. In the baseline configuration and under token deletion, the embedding dimensionality remains at 768, whereas under PCA the effective representation uses only k float32 values. The Gradio interface uses a stacked bar chart to visualize the combined text-plus-embedding storage for a single review before and after compression, making the storage trade-off concrete for individual examples. The plotting function *make_space_plot* is defined in Cell 28 and invoked from the main GUI callback in Cell 29 when the user runs comparisons in the web interface.

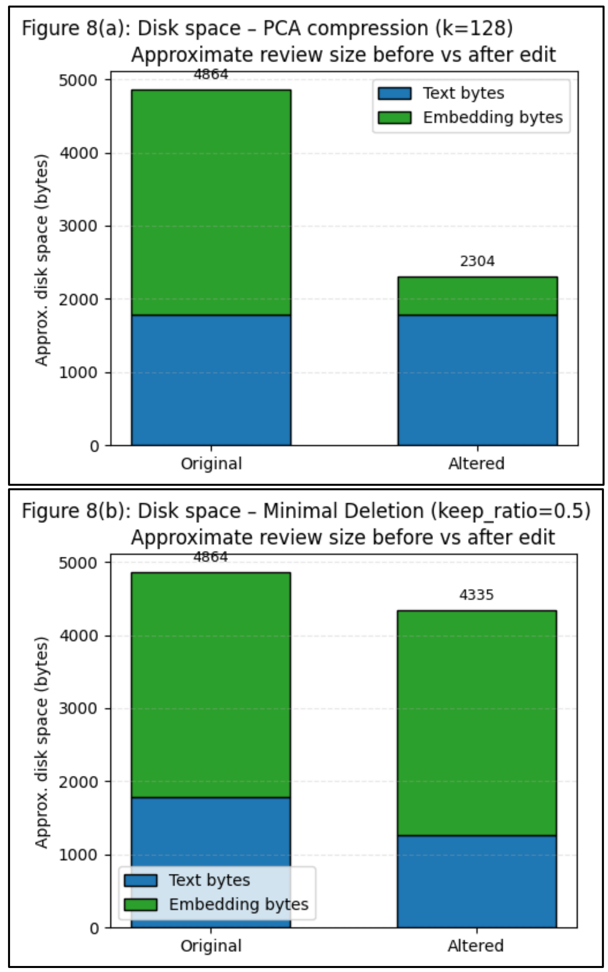

Fig. 4. Approximate review size (text bytes + embedding bytes) before and after compression for a single example, as displayed in the Gradio interface.

The unified results table, PCA and deletion trade-off plots, confusion matrices, and disk-space visualization collectively provide a comprehensive picture of how sentiment information is retained or lost under different compression settings.

VI. Interactive Analysis and Qualitative Evaluation

Quantitative metrics provide an overall view of sentiment retention under compression, but do not fully capture how individual reviews are affected. To address this, the project includes an interactive Gradio interface that wraps the entire pipeline and exposes its internal steps. Users can paste an arbitrary movie review or sample one from the dataset, optionally perform manual edits, and then apply either PCA-based compression or IDF-based deletion.

The interface displays the original and altered texts, predicted sentiment labels, and corresponding probabilities for negative and positive classes. A bar chart visualizes the probability distribution for original and altered embeddings side by side, facilitating direct comparison of how compression shifts confidence in each class. This probability plot is created by the function *make_probability_plot* in Cell 28 and rendered by the Gradio app defined in Cells 26–30.

Users can also see the effect of compression on storage via the stacked bar chart described above, generated by the *make_space_plot* function in Cell 28. In IDF-based deletion mode, a token-level diff visualization is provided: tokens retained in the compressed review are displayed normally, while deleted tokens are rendered in red with strikethrough formatting, and the compressed text is shown beneath. The HTML for this view is produced by the function *token_diff_html* in Cell 27 and displayed in the Gradio interface through the main callback defined in Cell 29.

Token-level diff (original):
~~i '~~ m ~~not~~ even going ~~to~~ waste more time describing how bad ~~this movie is .~~ bottom line : ~~it was~~ horribly acted ~~,~~ had enormous plot holes ~~and~~ went absolutely nowhere ~~. the~~ only ~~good~~ thing ~~about it was the~~ description ~~my~~ digital cable gave ~~for the movie~~ : ~~" a~~ married man ~~with a~~ struggling business ~~has a~~ fling ~~with his~~ secretary ~~. "~~ huh ? ? wrong ~~movie~~ apparently ~~,~~ although ~~it~~ may ~~have~~ made things slightly more interesting ~~if~~ any ~~of the~~ description were true ~~. < br / > < br / > - -~~ shelly

Compressed text:
m even going waste more time describing how bad bottom line : horribly acted had enormous plot holes went absolutely nowhere only thing description digital cable gave : married man struggling business fling secretary huh?? wrong apparently although may made things slightly more interesting any description were true shelly

Fig. 5. Token-level diff visualization for a sample review under IDF-based minimal deletion, rendered in the Gradio interface.

In practice, the token-level diff reveals that many frequent function words and generic phrases are removed at moderate keep ratios, whereas sentiment-bearing phrases and modifiers are more likely to be removed as the keep ratio becomes more aggressive. The Gradio interface thus supports a form of qualitative evaluation that complements aggregate metrics. Users can explore how specific reviews change under different compression settings, inspect whether label changes appear plausible, and identify cases where compression removes important cues such as negation or intensifiers. This interactive component is particularly valuable for pedagogical purposes and for building intuition about how BERT-based sentiment models behave under representation and input constraints [14].

Finally, a grouped bar chart of per-class F1 scores is provided in the notebook to compare the baseline model with the best-performing PCA and deletion configurations. This chart plots negative and positive class F1 scores for five methods (full BERT, top two PCA models, and top two deletion models) as grouped bars, giving a more detailed view of class-specific performance under compression. The plot is implemented in Cell 24 and can be used as an additional figure if finer-grained error analysis is desired.

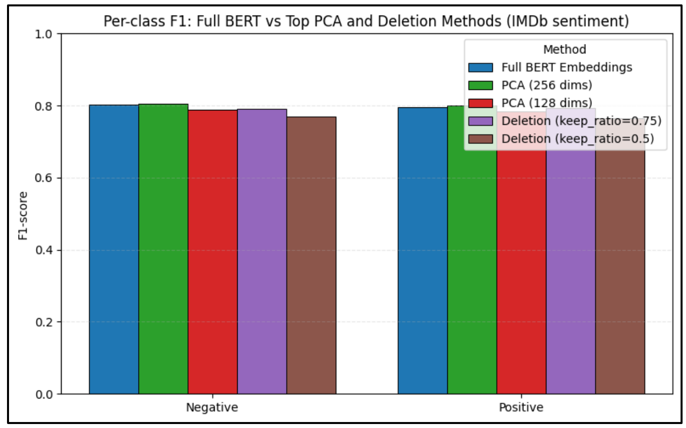

Fig. 6. Per-class F1 scores for full BERT, top PCA methods, and top deletion methods on the IMDb test set.

VII. Discussion

The empirical and interactive findings suggest several conclusions about sentiment retention under BERT embedding compression and deletion. First, the results confirm that BERT embeddings exhibit considerable redundancy for sentiment classification. Under PCA-based compression, a substantial reduction in dimensionality is possible before noticeable drops in accuracy and macro-averaged F1 occur. For dimensions on the order of 128 or 256, performance remains close to the baseline, and mean cosine similarity between original and reconstructed embeddings remains high. This observation aligns with earlier work demonstrating that PCA can effectively capture the dominant variance in sentence embeddings while discarding less informative components [6]–[8]. The PCA trade-off curve in Figure 1 provides a clear illustration of this phenomenon.

Second, IDF-based token deletion exhibits a different trade-off profile. Because this strategy alters the input text itself, it can remove context that BERT relies on to infer sentiment. At moderate keep ratios, performance remains relatively robust, and many deleted tokens are high-frequency terms with limited discriminative power. However, as the keep ratio decreases, the likelihood of deleting sentiment-bearing phrases increases, and performance declines more sharply. In some cases, the deleted tokens include negations or intensifiers whose removal reverses or weakens the sentiment of the remaining text. The deletion trade-off curve in Figure 2, together with the confusion matrices in Figure 3 and the token-level diff visualization in Figure 5, makes these failures particularly clear.

Third, the approximate storage analysis illustrates that compression is not a single-dimensional concept. PCA primarily reduces the size of the numerical representation, and is therefore particularly relevant in settings where embeddings rather than raw text are stored or transmitted. Token deletion, by contrast, primarily reduces the amount of text, which may be valuable in environments where storage of raw text is a constraint but where original content must not be heavily altered. The stacked bar chart in Figure 4, generated from real examples in the Gradio interface, provides an intuitive view of this trade-off.

Finally, the combination of aggregate metrics, confusion matrices, per-class F1 plots, and interactive inspection suggests that moderate compression regimes can retain sentiment information reasonably well, while extreme compression entails substantial risk of misclassification. The choice of compression strategy and strength should therefore be guided by the specific constraints and risk tolerance of the target application, particularly when decisions based on sentiment predictions have downstream consequences.

VIII. Ethics and Limitations

Several ethical and practical limitations arise from the design and deployment of sentiment models with compression.

The use of the IMDb dataset must follow its intended non-commercial research and educational purposes [11], [12]. Although the reviews are public, they may contain personal or sensitive information. Any future deployment should consider data minimization, secure storage, and compliance with applicable privacy regulations.

The IMDb corpus represents a specific population: users who write English-language movie reviews online. Differences in demographics, cultural norms, and writing styles mean that models trained solely on this data may not generalize equitably to other domains or user groups [3], [11]. In addition, BERT itself is pre-trained on large general-domain corpora that encode societal biases [1]. Consequently, a compressed sentiment model of this type may exhibit unfair behavior across content describing different demographic groups or topics, even if overall accuracy is high.

Compression methods introduce additional ethical considerations. IDF-based deletion changes the textual content, which can remove context and nuance, potentially misrepresenting the original author's intent. Compressed text should therefore not be presented as verbatim user content. In analytical settings, any use of compressed representations should be clearly disclosed, and affected users should not be led to believe that compressed text is identical to their original input. PCA-based compression, while operating in latent space, may also differentially distort representations for some types of content; for example, subtle sentiment cues expressed through rare vocabulary or complex syntax might be underrepresented in lower-dimensional subspaces.

From a methodological perspective, the study is limited to a frozen BERT encoder with a linear classifier. Alternative architectures, including fine-tuned transformers or more expressive classification heads, might exhibit different robustness properties under compression. In addition, the chosen compression strategies are limited to PCA and static IDF-based deletion. More advanced techniques, such as learned token pruning, distillation, and quantization [9], [10], may yield more favorable trade-offs and could be incorporated into the same evaluation framework. Finally, the human evaluation conducted through the Gradio interface is informal and exploratory; a more rigorous user study would be required to quantify perceived quality and trustworthiness of compressed predictions.

IX. Conclusion

This study presents an analysis of sentiment retention in a BERT-based sentiment classifier under PCA-based embedding compression and IDF-based token deletion. Using the IMDb 50K Movie Reviews dataset, a frozen `*bert-base-uncased*` encoder, and a logistic regression classifier, the experiments show that moderate PCA compression can significantly reduce embedding dimensionality while preserving much of the baseline performance, whereas aggressive token deletion degrades accuracy and macro-averaged F1 more rapidly as important context is removed. An interactive Gradio interface further illustrates these trade-offs at the example level by displaying sentiment probabilities, approximate storage requirements, and token-level deletion patterns.

The results suggest that carefully calibrated compression can make BERT-based sentiment models more efficient without entirely sacrificing performance, but also underscore the importance of understanding when and how compression alters model behavior. Future work may extend this framework to other datasets, tasks such as aspect-based sentiment analysis, and more advanced compression techniques, as well as conduct more systematic human evaluations to better characterize perceived quality and fairness under compression.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[2] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP: System Demonstrations, 2020.

[3] P. S. S. V. Domadula et al., "Sentiment analysis of IMDb movie reviews using BERT-based models," Tech. Rep., 2023.

[4] N. A. Semary et al., "Improving sentiment classification using a RoBERTa-based hybrid model," BMC Bioinformatics, vol. 24, no. 1, pp. 1–19, 2023.

[5] L. Maltoudoglou, A. Paisios, and H. Papadopoulos, "BERT-based conformal predictor for sentiment analysis," in Proc. 9th Symp. Conformal and Probabilistic Prediction and Applications, 2020, pp. 269–284.

[6] Á. Huertas-García et al., "Exploring dimensionality reduction techniques in text embedding spaces," Appl. Sci., vol. 12, no. 20, p. 10424, 2022.

[7] G. Zhang et al., "Evaluating unsupervised dimensionality reduction methods for sentence embeddings," in Proc. LREC, 2024.

[8] H. Abdelmotaleb et al., "A comparative study of word embedding techniques for sentiment analysis," Expert Syst. Appl., 2025.

[9] S. Kim et al., "Learned token pruning for transformers," in Proc. 28th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2022.

[10] X. Mei et al., "Token pruning in large neural networks: A survey," Preprint, 2024.

[11] A. L. Maas et al., "Learning word vectors for sentiment analysis," in Proc. ACL, 2011.

[12] Kaggle, "IMDb dataset of 50K movie reviews," Dataset, online: https://www.kaggle.com/datasets, accessed 2025.

[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[14] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ML models in the wild," arXiv preprint arXiv:1906.02569, 2019.