Learned Token Pruning: Patterns and Performance

Background & Motivation

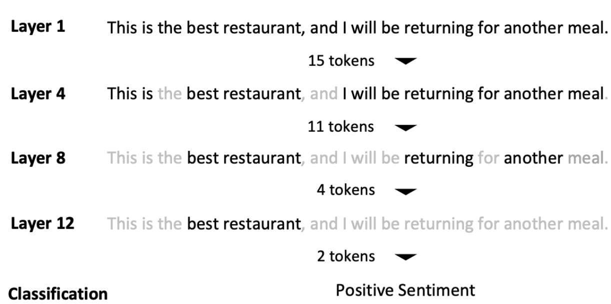
The paper Learned Token Pruning[1] reduces the number of tokens processed by Transformers to improve model efficiency (smaller size, faster inference).

LTP aims to increase inference speed by selectively pruning less impactful tokens, improving efficiency without significantly sacrificing performance.

However, the implications of pruning on interpretability, fairness, and domain-specific token retention remain unclear.

Problem Statement

Pruning Patterns: Analyze which types of tokens are retained (e.g., stopwords, named entities, adjectives, domain-specific terms) after pruning across different domains



Performance Impact: Assess how pruning affects text classification performance (e.g., accuracy).

Data

To conduct this analysis, we will utilize the following short text classification datasets:

Domain	Dataset Name	Description
General Text	**AG News**	News articles categorized into four classes (World, Sports, Business, Science), each with concise text (~120 words).
Sentiment	**IMDb Reviews**	Movie reviews labeled as positive or negative, generally around 200 words.

Methodology

Experiment 1: Model Training and Pruning

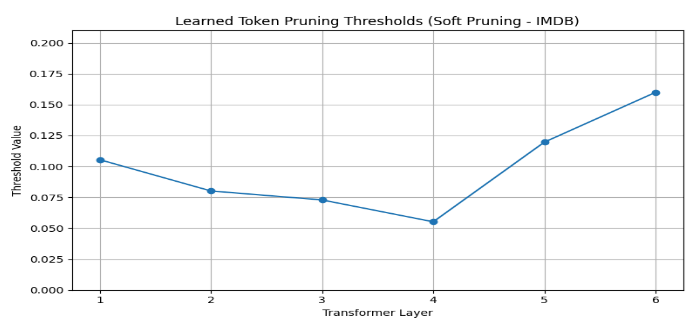
1. **Train a Baseline Model:** Fine-tune a transformer RoBERTa model on each dataset without pruning to establish baseline performance metrics (accuracy, F1 score).
2. **Implement Token Pruning:** Apply the learned token pruning (LTP) method to the model during fine-tuning to reduce the number of tokens processed.

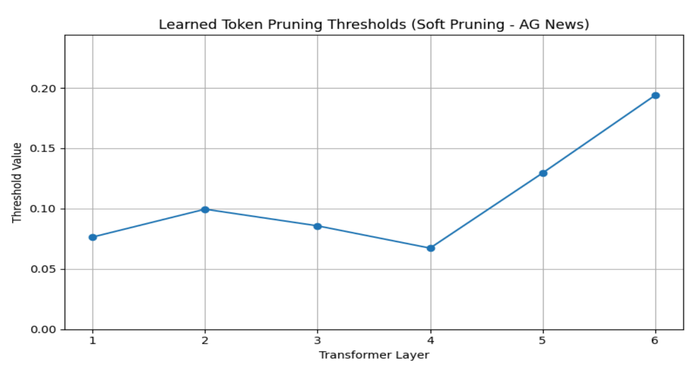
Algorithm 1 Three-step Training Procedure for Learnable Threshold Token Pruning

Input: $\mathcal{M}$: model finetuned on target downstream task

Step 1: Apply soft mask to $\mathcal{M}$ and train both the thresholds and model parameters ▷ Soft Pruning
Step 2: Binarize the mask and fix the thresholds
Step 3: Finetune the model parameters ▷ Hard Pruning

3. **Finetune Pruning Threshold:** Soft pruning training helps dynamically adjust the pruning threshold of each transformer layer based on the requirements of the downstream task. This adaptive behavior is evident in the figures, where each layer learns a distinct threshold, reflecting its relative importance in preserving task-relevant information.





Experiment 2: Token Analysis

1. **Token Extraction:** After training, extract the retained tokens that were after pruned by the model.
2. **Token Classification**
   - Categorize these retained tokens using NLP tool spaCy for POS tagging.
   - Label tokens as stopwords, named entities, adjectives, domain-specific terminology, etc.
3. **Frequency Analysis**
   - Compute the frequency of each token category.
   - Identify patterns in which tokens are mostly retained.
4. **Cluster Analysis:** Generate t-SNE visualization to perform retained token cluster analysis.

Experiment 3: Performance Evaluation

Re-evaluate Classification Metrics: Measure the classification performance (accuracy, F1-score) of the pruned model on each dataset.

Experiment 4: Domain-Specific Retention

1. **Domain-Specific Term Identification:**
   - Identify key domain-specific terms in each dataset (e.g., movie-related terms in IMDb reviews).
2. **Impact on Domain-Specific Performance:**
   - Assess whether the pruning approach consistently retains or discards these terms, and how this retention (or removal) affects classification outcomes.

Performance Evaluation

IMDB:

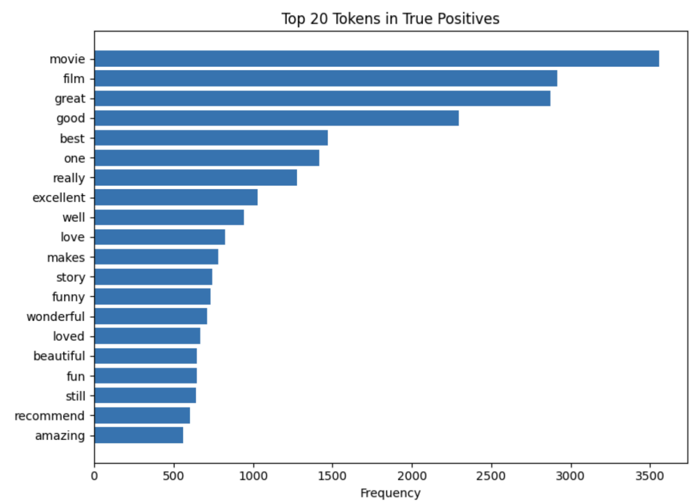
```
+-----------+---------+
| Metric    |  Value  |
+===========+=========+
| Accuracy  |  0.9017 |
+-----------+---------+
| Precision |  0.9114 |
+-----------+---------+
| Recall    |  0.89   |
+-----------+---------+
| F1-Score  |  0.9006 |
+-----------+---------+

Classification Report:
              precision    recall  f1-score   support

    Negative       0.89      0.91      0.90     12500
    Positive       0.91      0.89      0.90     12500

    accuracy                           0.90     25000
   macro avg       0.90      0.90      0.90     25000
weighted avg       0.90      0.90      0.90     25000
```

AG News:

```
+-------------------+----------+
| Metric            |    Value |
+===================+==========+
| Accuracy          |   0.9363 |
+-------------------+----------+
| Precision (Macro) |   0.9374 |
+-------------------+----------+
| Recall (Macro)    |   0.9363 |
+-------------------+----------+
| F1-Score (Macro)  |   0.9364 |
+-------------------+----------+

Classification Report:
              precision    recall  f1-score   support

       World       0.97      0.92      0.95      1900
      Sports       0.97      0.98      0.98      1900
    Business       0.92      0.89      0.91      1900
    Sci/Tech       0.88      0.94      0.91      1900

    accuracy                           0.94      7600
   macro avg       0.94      0.94      0.94      7600
weighted avg       0.94      0.94      0.94      7600
```

IMDB Results

Token Frequency Analysis

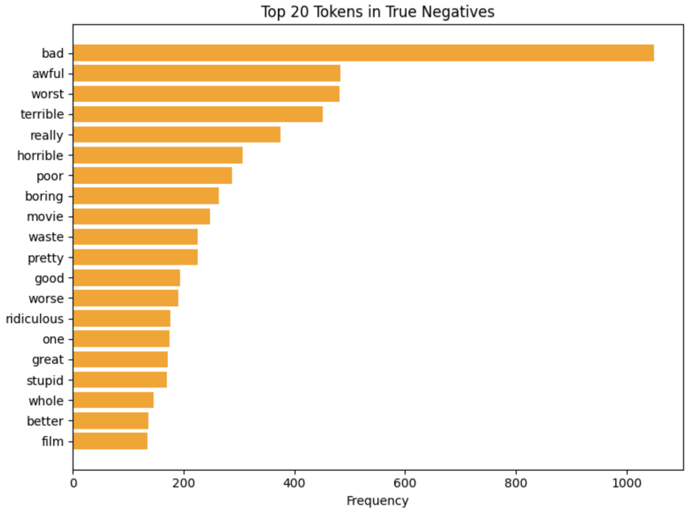


True Positives Inference:

Dominant words like “movie”, “film”, “great”, “good”, “best”, “excellent”, and “amazing” highlight a strong presence of positive sentiment vocabulary.

Words like “love”, “funny”, and “recommend” further reinforce user appreciation.

The frequency of these tokens confirms that the model has learned to associate these words with positive sentiment and leverages them effectively in prediction.



True Negatives Inference:

Negative sentiment is clearly represented by high-frequency words such as "bad", "awful", "worst", "terrible", "horrible", "boring", and "waste".

Some overlap with neutral or general words like "movie", "film", "good", "one", and "great" appears, but their frequency is lower here and likely used in negative context.

E.g., "good" or "great" might appear in a sarcastic or contrastive way (e.g., "It tries to be great but fails").

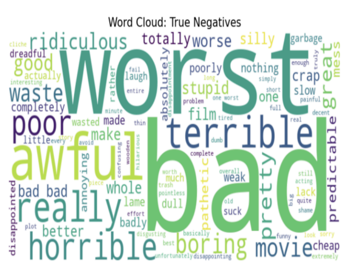
Overall Insight:

The model appears to have successfully learned polarity-specific cues from tokens.

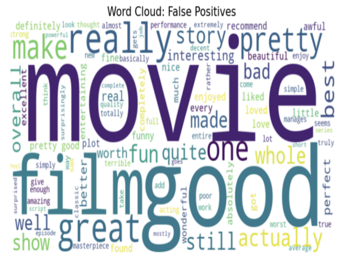
The use of context-dependent terms like "really", "movie", and "film" in both classes suggests it also utilizes contextual embedding, not just surface word frequencies

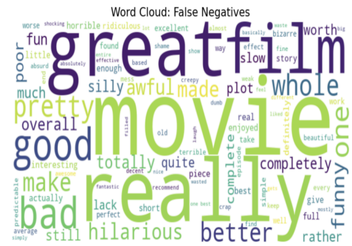
Word Cloud

Correct Classification

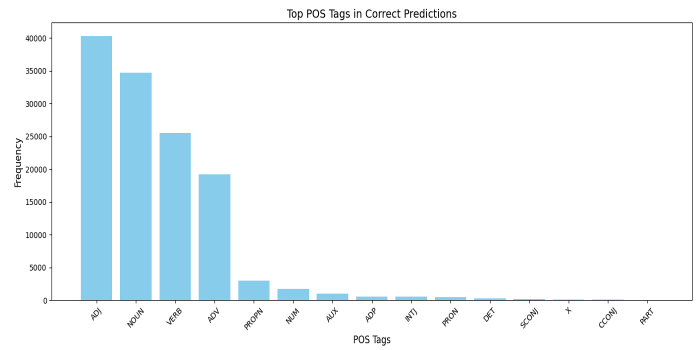


Incorrect Classification





POS Tag Frequency



This POS tag distribution for correct predictions shows that the most frequent parts of speech are:

Adjectives (ADJ) and Nouns (NOUN) - Detect polarity and subject matter.

Verbs (VERB) and Adverbs (ADV) – contribute action and intensity

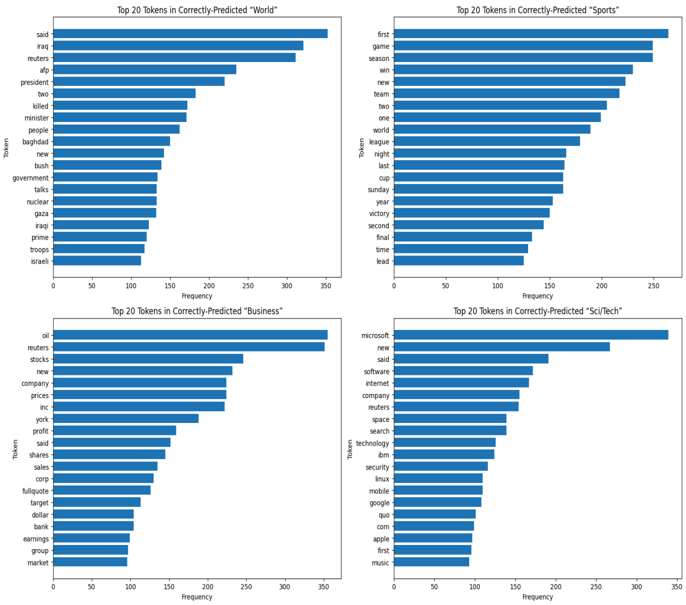
Proper nouns (PROPN) appear less frequently, indicating that specific named entities may not be as crucial for sentiment classification.

Inference:

Correct predictions are strongly driven by descriptive and sentiment-bearing POS tags (adjectives, nouns, verbs, adverbs), highlighting that the model effectively uses core linguistic features tied to sentiment and meaning. Less frequent reliance on functional or grammatical tags (e.g., DET, AUX, ADP) suggests the model focuses more on content words for decision-making.

AG News Results

Token Frequency Analysis



Inference:

World: The model identifies this class primarily through geopolitical and government-related terms such as iraq, president, minister, troops, baghdad. Frequent mentions of international news agencies like Reuters and AFP suggest reliance on named entities and country-specific vocabulary. This indicates strong model performance in recognizing global and political discourse.

Sports: Characterized by event-driven and temporal language, with frequent use of words like season, win, final, team, league. The presence of ordinal indicators such as first, second, last points to a focus on match sequences, rankings, and outcomes. The model effectively leverages competitive and time-based cues to classify sports-related content.

Business: Dominated by market and finance-related vocabulary including oil, stocks, prices, profit, bank. The use of corporate references like inc, corp, shares, earnings suggests the model captures this class through economic indicators and company-specific terms. It demonstrates a strong grasp of financial reporting language.

Sci/Tech: Marked by a high density of tech-related entities and companies such as microsoft, google, apple, linux, along with broader scientific and technical terms like internet, space, security, technology. The model appears adept at recognizing specialized jargon and named entities that define this domain.

Overall Insight:
The model effectively distinguishes classes using domain-specific keywords and named entities.

Word Cloud

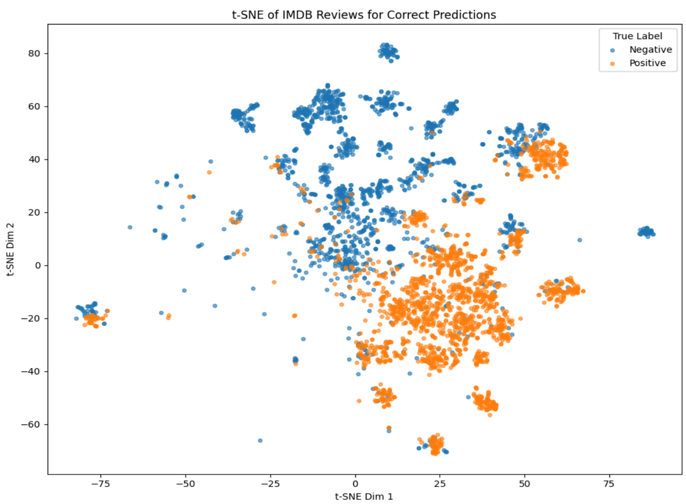
POS Tag Frequency

Inference:

- Nouns (NOUN) dominate all classes: Reflecting the factual, entity-driven nature of news articles (e.g., people, places, organizations, technologies).
- PROPN are prominent in World, Sports, and Business, reflecting frequent mentions of countries, teams, companies, and people. Sci/Tech shows fewer proper nouns, likely due to a focus on general tech terms over named entities.
- Verbs (VERB) are more frequent in Sports and Business, reflecting action-oriented language in match or market reporting.
- Adjectives (ADJ) are fairly balanced across classes, with a slight increase in Sci/Tech, likely due to descriptive coverage of features and innovations.
- Adverbs (ADV) and Numerals (NUM):
  - Sports and Business have higher counts of numerals and adverbs — consistent with statistics and performance summaries.

- Sci/Tech has more adverbs than numerals, perhaps indicating a focus on method/process description over quantitative summaries.

t-SNE Visualization on IMDB



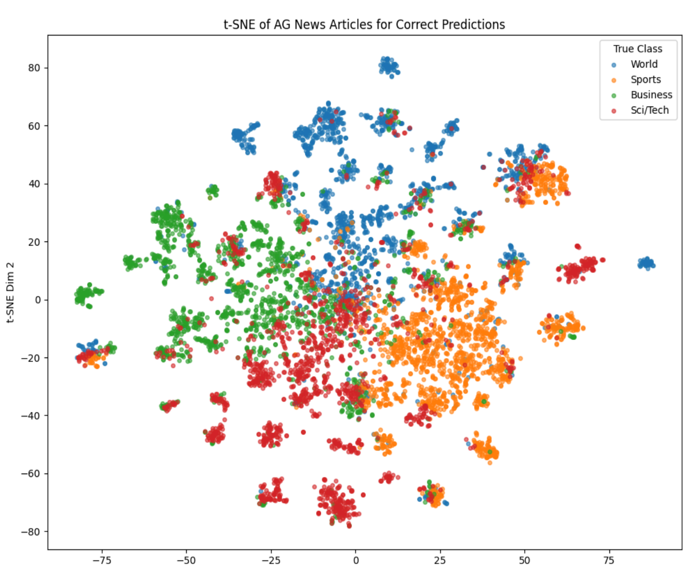
Inference:

Distinct separation between positive and negative clusters demonstrates the model's strong ability to capture sentiment-specific semantics.

The positive cluster (orange) is denser, reflecting consistent language patterns used to express positive sentiments, while the negative cluster (blue) is more spread out, suggesting greater variability and complexity in negative expressions.

Some minor overlap or boundary regions reflect nuanced or ambiguous sentiment, highlighting cases where the sentiment might be less explicitly expressed.

t-SNE Visualization on AG News



Inference:

- Each class forms visibly coherent clusters, indicating that the model has learned discriminative features that separate the categories well in latent space.
- Retained tokens in AG News capture strong domain-specific vocabulary
- Topics are inherently easier to separate because they rely heavily on specific nouns and proper terms — words that remain even after aggressive token pruning.

Expected Outcomes

1. **Token Pruning Patterns :** Quantitative insights into the types of tokens (stopwords, named entities, adjectives, etc.) most frequently removed by the LTP mechanism across different domains.
2. **Performance Implications :** An evidence-based understanding of how pruning affects classification metrics, with a focus on potential accuracy improvements.

Conclusion

Overall, the retained token analysis shows that the model is not relying on superficial cues. Instead, it's focusing on meaningful, domain-specific keywords that genuinely differentiate each category.

This suggests that the model's behavior is aligned with what we would hope to see: using truly informative vocabulary, not just spurious patterns.

References

1. Kim, Y., Chang, J., & Smith, A. (2021). *Learned Token Pruning for Transformers.* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 432–445.
2. Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.