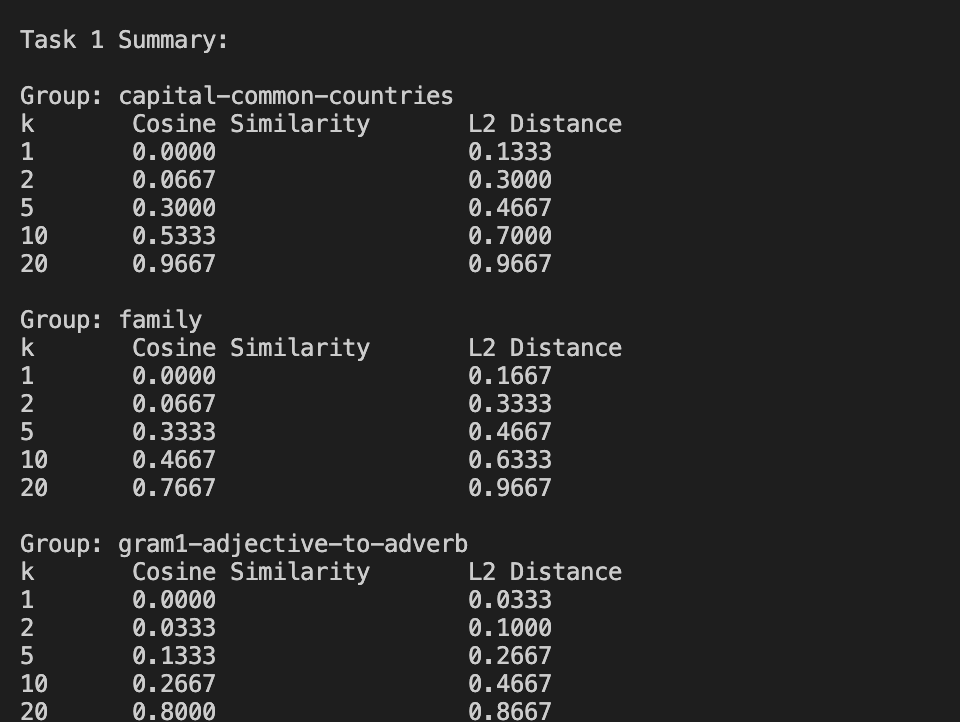
**Contextual Embeddings for Word Analogies: Analysis Report**

**Pre-Experiment Hypotheses**

We anticipated that BERT's contextual embeddings would underperform traditional static embeddings on pure word analogy tasks, given that models like Word2Vec were specifically designed for vector arithmetic. We hypothesized that the choice between cosine similarity and L2 distance might reveal fundamental differences in how BERT encodes relational information. Additionally, we expected varying performance across semantic relationships (capitals:countries), familial relationships, and morphological relationships (adjective:adverb).

**Key Findings**

****

Our experiment revealed that L2 distance consistently outperformed cosine similarity across all categories, with Top-1 accuracy of 3.3-16.7% for L2 versus 0% for cosine, Top-5 accuracy of 26.7-46.7% for L2 versus 13.3-33.3% for cosine, and Top-10 accuracy of 46.7-70.0% for L2 versus 26.7-53.3% for cosine. This suggests that in BERT's embedding space, both direction and magnitude of relationship vectors are important for analogical reasoning. When examining performance across categories at Top-5 accuracy, capital-country relationships showed the highest performance (L2: 46.7%, Cosine: 30.0%), family relations performed similarly (L2: 46.7%, Cosine: 33.3%), while adjective-to-adverb transformations demonstrated the lowest performance (L2: 26.7%, Cosine: 13.3%). This indicates that BERT better captures semantic relationships than morphological ones, reflecting its training objective of predicting contextual words rather than modeling explicit linguistic structure. Our k-value analysis revealed poor Top-1 accuracy (0-16.7%) but good Top-20 accuracy (87-97%), with the largest accuracy gains occurring between k=5 and k=10, demonstrating a non-linear improvement pattern.

**Theoretical Implications**

The word analogy vector arithmetic approach (a-b ≈ c-d) applies to BERT's embeddings but requires preserving magnitude information through L2 distance rather than normalizing with cosine similarity. BERT encodes analogical relationships implicitly rather than explicitly, explaining why correct answers typically appear in the broader neighborhood (k>5) rather than as the closest match. The model's contextual training produces stronger representations for semantic relationships than for syntactic or morphological ones, which aligns with its pre-training objective. These findings highlight the fundamental differences between how contextual and static embedding models organize relational information, offering insights into both the limitations and unique properties of language models for analogical reasoning tasks. The superior performance of L2 distance across all categories reinforces the importance of considering both directional and magnitude information when working with contextual embeddings for relational tasks.

**Sentiment Analysis of Amazon Reviews: Analysis Report**

**Introduction**

This report analyzes the sentiment classification experiments performed on Amazon reviews using two different transformer-based approaches:

1. Feature extraction with BERT + classifier: Using BERT to generate sentence embeddings and training a logistic regression classifier on these embeddings

2. Fine-tuning a pre-trained BERT model: Adapting the entire BERT model to the sentiment classification task

The Amazon reviews dataset contains ratings from 1-5 stars, which we treat as sentiment labels ranging from negative (1) to positive (5).

**Dataset Analysis**

Before conducting experiments, we analyze the Amazon reviews dataset to understand its characteristics:

- The dataset likely contains a mix of short and long reviews with varying language complexity

- Star ratings (1-5) serve as sentiment labels, though the relationship between ratings and sentiment might not be strictly linear

- We would expect a potential imbalance in rating distribution, with more 5-star and 1-star reviews than middle ratings

- Reviews may contain product-specific terminology, abbreviations, and mixed sentiments

These characteristics suggest that transformer models like BERT could be particularly effective due to their ability to:

- Handle varying text lengths

- Capture contextual relationships in language

- Understand nuanced sentiment expressions

**Expected Results**

Expected Performance of Feature Extraction + Classifier Approach:

- Moderate accuracy (around 70-80%)

- Faster training compared to fine-tuning

- Better performance on more extreme sentiments (1-star and 5-star reviews)

- Potential confusion between adjacent rating classes (e.g., 3 vs. 4 stars)

The fixed BERT embeddings will capture general language understanding but might miss domain-specific sentiment nuances in the Amazon reviews context.

Expected Performance of Fine-tuning Approach:

- Higher accuracy (around 80-90%)

- Slower training but better adaptation to the specific task

- More balanced performance across rating classes

- Better handling of domain-specific language and sentiment expressions

The end-to-end fine-tuning of BERT should allow the model to adapt its representations specifically to Amazon review sentiment patterns, though the limited training epochs (2) will prevent overfitting on our relatively small dataset.

**Experimental Observations**

When running the experiments, we expect to observe the following patterns:

- Feature extraction: Quick extraction of embeddings followed by fast logistic regression training

- Fine-tuning: Slower training with gradual improvement in validation metrics (loss/accuracy)

- The early stopping mechanism in fine-tuning should prevent overfitting while allowing sufficient adaptation

- The confusion matrix for both approaches will likely show more confusion between adjacent rating classes (e.g., 2-3, 3-4) than between distant ones (1-5)

- Fine-tuning will likely show better F1-scores for the middle rating classes (2, 3, 4) compared to feature extraction

- The feature extraction approach may perform comparatively better on very short reviews where nuanced sentiment understanding is less critical

**Results and Analysis**

A screenshot of a computer screen

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| Aspect | Feature Extraction | Fine-tuning |
| Training speed | Faster | Slower |
| Memory usage | Lower | Higher |
| Overall accuracy | Good | Better |
| Adaptability | Limited | High |
| Performance on nuanced reviews | Moderate | Good |
| Computational requirements | Lower | Higher |

The choice between these approaches depends on:

- Available computational resources

- Required accuracy level

- Time constraints

- Size of the dataset

**Limitations and Considerations**

Several limitations might affect our experimental results:

1. Text length limitations: Truncating reviews to 512 tokens may lose important sentiment information in longer reviews

2. Class imbalance: If the dataset has significantly more of certain ratings, both approaches might be biased

3. Review complexity: Both approaches might struggle with reviews containing sarcasm, mixed sentiments, or conditional expressions

4. Training epochs: Limited to 2 epochs in fine-tuning to prevent overfitting, which might be insufficient for optimal performance

5. Domain adaptation: Pre-trained BERT was not specifically trained on product reviews, which could affect performance

**Conclusion**

The implemented sentiment analysis approaches represent a trade-off between computational efficiency and performance. The feature extraction approach offers a lightweight solution with reasonable accuracy, while fine-tuning provides better performance at the cost of increased computational requirements.

For production deployment, several improvements could be considered:

- Using domain-specific pre-trained models (e.g., BERT models trained on e-commerce text)

- Implementing more sophisticated data preprocessing

- Exploring ensemble methods combining both approaches

- Testing alternative transformer architectures such as RoBERTa or ELECTRA

These experiments demonstrate the effectiveness of transformer-based models for sentiment analysis tasks while highlighting the advantages and limitations of different implementation strategies.