

Embedding Models for Addressing Linguistic Challenges

PRESENTED BY: TRUNG KIEN NGUYEN

INSTITUTION: A2I2 – DEAKIN UNIVERSITY

MENTORS: SCOTT BARNETT & SHENG WONG

DATE: TUESDAY, DECEMBER 17, 2024



AGENDA

1. Background and Problem Statement
2. Objectives
3. Datasets and Methodology
4. Experiments and Findings
5. Challenges and Recommendations
6. Conclusion and Future Work

BACKGROUND & PROBLEM STATEMENT

Context: Current SOTA text embedding models excel in general tasks but struggle with nuanced linguistic challenges.

Key Problems:

- Handling Nuanced Meanings
- Complex Relationships
- Contextual Awareness
- Evaluation Metrics.

Significance: Limitations affect applications needing deeper language understanding

OBJECTIVES

Primary Goal: Enhance embedding models to effectively process complex linguistic phenomena.

Specific Objectives:

- Evaluate Existing Models
- Dataset Exploration
- Introduce Better Metrics
- Provide Actionable Insights

DATASETS USED

NevIR Dataset

- a) Focus: Differentiates natural language sentences by their relationships: Entailment, Neutral, and Contradiction.
- b) Structure: Premises, Hypothesis, and categorized binoculars (Cause A, Cause B, Effect A, Effect B).
- c) Link: <https://github.com/EternityYW/TRAM-Benchmark/blob/main/datasets/causality.zip>
 - Queries (q1, q2) and Documents (doc1, doc2).
 - **Key Aspect:** doc2 introduces negation differences from doc1.
- c) Link: <https://huggingface.co/datasets/orionweller/NevIR>

Purpose of Datasets:

- Provide diverse benchmarks for testing linguistic understanding.
- Address specific challenges in negation, causality, spatial reasoning, and logical inference.

METHODOLOGY

1. Dataset Preparation

2. Embedding models

- Selected NV-Embed-v2 for generating embeddings (HuggingFace repository).
- Embedded text pairs (queries, documents, premises, and hypotheses).

3. Evaluation Metrics

- Cosine Similarity
- Density Distribution Analysis

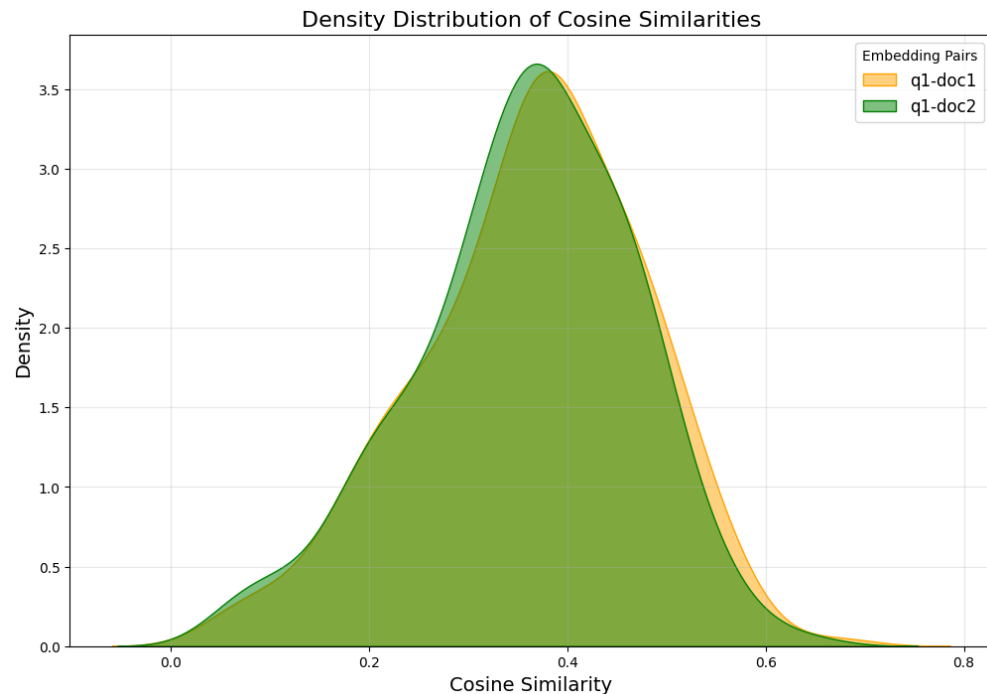
4. Experimentation

- Negation Detection
- Causal reasoning
- Logical inference (NLI)

EXPERIMENTS CONDUCTED

Negation Handling (NevIR Dataset)

- Compared similarity between queries and negated documents.
- Observation:

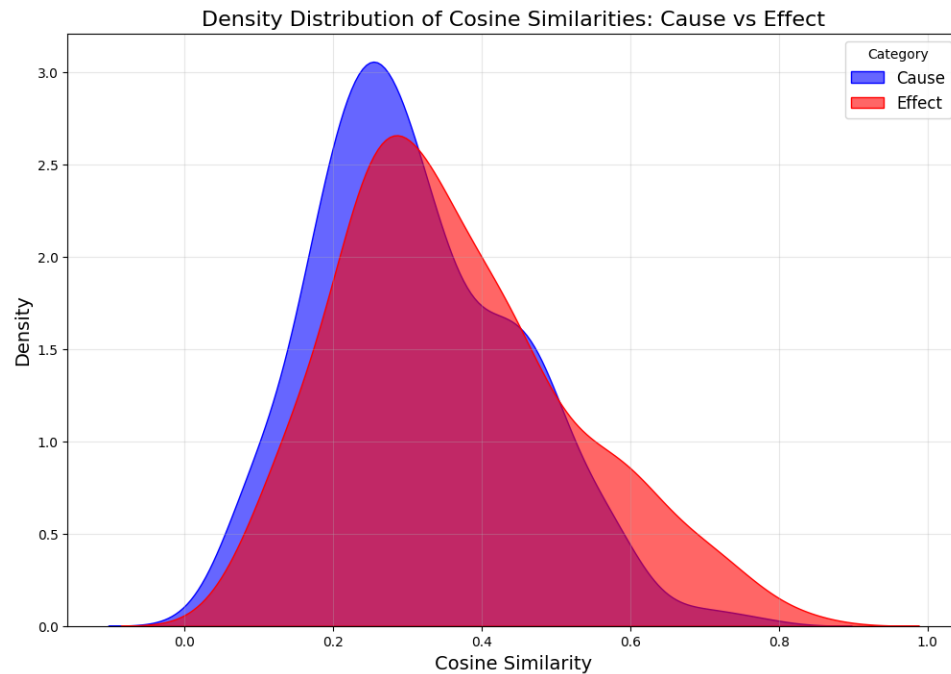


- Failed to capture the difference between Q1 & Q2
- Failed to capture the difference between the two documents (doc1 and doc2), which is expected to be significant

EXPERIMENTS CONDUCTED

Causal Reasoning (Causality Dataset)

- Evaluated semantic similarity between causes and effects.
- Observation:



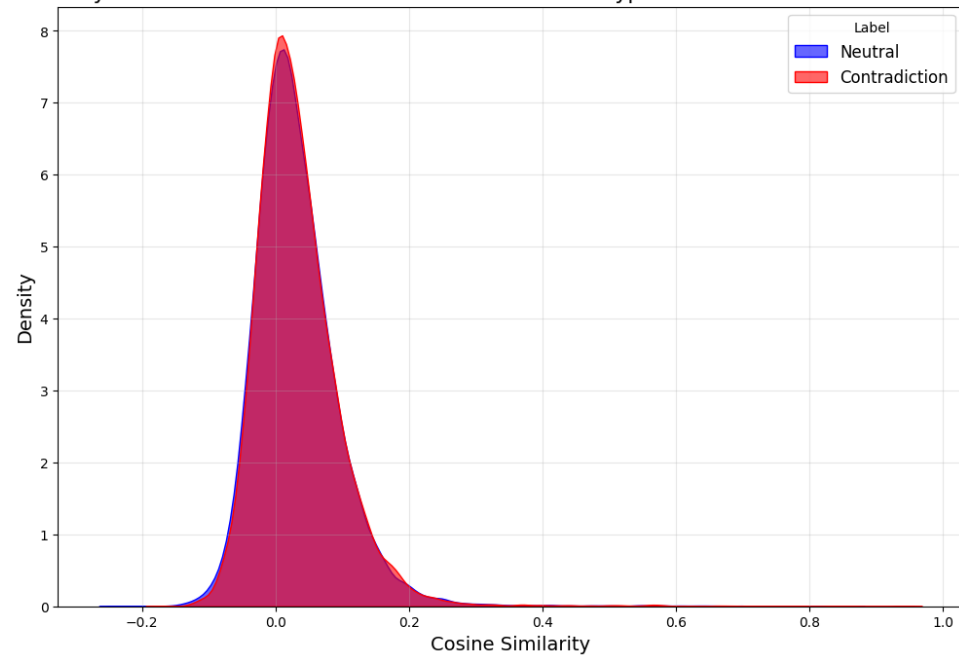
- Failed to capture the similarity between two causes, or two effects (0.35 to 0.4 at peak)
- Overlap in scores highlighted difficulty in differentiation.

EXPERIMENTS CONDUCTED

Logical Inference (NLI Dataset)

- Compared neutrality, and contradiction pairs.
- Observation:

Density Distribution of Cosine Similarities Premise & Hypothesis: Neutral vs Contradiction



- Overlap means it failed to distinguish the NLI relationship (Neutral or Contradiction)

CONCLUSION AND REFLECTION

Current SOTA Embedding models struggle with nuanced semantic tasks like negation, causality, and logical inference.

- Negation is Poorly Handled
- Cause and Effect Relationships Lack Distinction
- Significant overlap in Logical Relationships
- Cosine Similarity Alone is Inadequate

Advancing embedding models for linguistic challenges will pave the way for better real-world applications in NLP.

RECOMMENDATIONS AND FUTURE WORK

1. Task-Specific Models Development
2. Enhanced Metrics
3. Dataset Expansion
4. Fine-Tuning Pipelines

THANK YOU FOR LISTENING
