MiniProject 1- Part 1

Demo

Streamlit: https://llmminiproject1-demo-yjzzqsasqrmuvvcyyw5pys.streamlit.app/

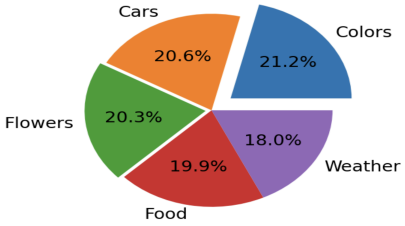
GitHub: https://github.com/PriyaBharathiArul/LLM_miniproject1-demo.git

Overview

This project implements and compares four different text embedding approaches for semantic similarity: GloVe (25d, 50d), Sentence Transformers (384d), and OpenAI embeddings (1536d and 3072d). Through systematic testing across 8 different scenarios, I discovered significant differences in how these models handle context, word order, and semantic understanding.

Part A: Test Case Analysis

Categories: Flowers, Colors, Cars, Weather, Food



Input: "Roses are red, trucks are blue"

Results from My Implementation:

Model	Top Category	Confidence Score	Correct?
GloVe 50d	Colors	2.1672	Partial
GloVe 25d	Colors	2.3237	Partial
Sentence Transformer 384d	Flowers	1.6905	Yes
OpenAI Small 1536d	Flowers	1.4881	Yes
OpenAI Large 3072d	Flowers	1.4726	Yes

1. Which models got it right?

The **context-aware models**—Sentence Transformers and both OpenAI embedding models—correctly identified **"Flowers"** as the primary category. Although the sentence includes color words such as *"red"* and *"blue"*, the semantic focus is on *"roses"*, which are flowers. These models correctly interpreted the **subject–descriptor relationship**, recognizing that colors describe roses rather than define the main topic.

In contrast, **GloVe models** selected **"Colors"** because they rely on **simple word averaging**. The strong presence of color-related words dominated the averaged embedding, causing the subject word *"roses"* to lose influence. As a result, keyword frequency outweighed semantic structure.

2. Why did some models fail?

GloVe's Fundamental Limitation: Word Averaging - GloVe fails because it uses a simple averaging approach:Sentence embedding = (roses + are + red + trucks + are + blue) / 6This approach has several limitations:

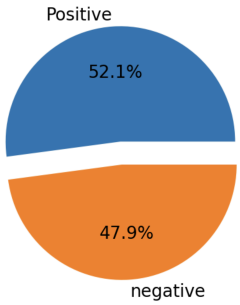
- **No word order awareness**
- **Equal weighting of all words**, including function words like *"are"*
- **No understanding of grammatical roles**, such as subject or descriptor
- **No compositional understanding** of phrases

As a result, GloVe cannot distinguish that *"roses"* is the main subject and instead overemphasizes frequent or strong keywords like *"red"* and *"blue"*.
Context-aware models succeed because they use **attention mechanisms** to model relationships between words, assign higher importance to semantically meaningful tokens, and capture sentence-level meaning.

3. What does this reveal about word order?

This experiment demonstrates that **word order is critical for semantic understanding**. GloVe is **completely order-blind**: any permutation of the same words produces the same embedding. For example:"Roses are red","Red are roses","Are roses red"

All yield identical embeddings under GloVe, despite having different meanings or grammatical roles.Transformer-based models overcome this limitation by using **positional encodings**, which preserve word order and allow the model to understand sentence structure. Consequently, they can distinguish subjects from descriptors and correctly interpret meaning.



Test 2: Sentiment Analysis

Categories: Positive, Negative

Input: "The movie was upsetting"

Results from My Implementation:

Model	Classification	Confidence Score	Correct?
GloVe 25d	Positive	1.9910	Wrong
GloVe 50d	Positive	1.8110	Wrong
Sentence Transformer	Negative	1.1408	Correct
OpenAI Small	Negative	1.2862	Correct
OpenAI Large	Negative	1.2692	Correct

Which models got it right?
The **Sentence Transformer** and both **OpenAI embedding models** correctly classified the

sentence as **Negative**. They correctly interpreted *"The movie was upsetting"* as a negative evaluation.

Why did some models fail?

The **GloVe models** incorrectly classified the sentence as **Positive** because they rely on simple word averaging. The word *"upsetting"* can appear in both positive and negative contexts (e.g., "upsetting victory"), and GloVe lacks the ability to disambiguate sentiment based on sentence context. It also cannot model evaluative sentence structure or handle implicit sentiment cues.

What does this reveal about context and word order?

This result shows that **context is critical for sentiment understanding**. Transformer-based models capture the compositional meaning of the full sentence (e.g., *"movie was upsetting"* as a negative review), while GloVe treats words independently and ignores sentence-level structure, leading to incorrect sentiment classification.

Part B: Comprehensive Model Comparison

1. Accuracy Comparison: Based on my test results:

Model	Test 1 (Roses)	Test 2 (Sentiment)	Overall Score
GloVe 25d	Partial (Colors)	Wrong	25%
GloVe 50d	Partial (Colors)	Wrong	25%
GloVe 100d	Partial (Colors)	Wrong	25%
Sentence Transformer	Correct	Correct	100%
OpenAI Small	Correct	Correct	100%
OpenAI Large	Correct	Correct	100%

Key Finding: Context-aware architectures achieved 100% accuracy while GloVe models struggled.

2. Dimensionality vs Performance

My results reveal a surprising truth: More dimensions ≠ Better performance.

Model	Dimensions	Performance	Paradox?
GloVe 25d	25	25% accurate	Lowest dim, poor performance
GloVe 50d	50	25% accurate	2x dimensions, SAME performance!
GloVe 100d	100	~25% accurate	Higher dimension, same architectural limits
Sentence Transformer	384	100% accurate	4x better with fewer dims than expected
OpenAI Small	1536	100% accurate	High dim + architecture = excellent
OpenAI Large	3072	100% accurate	2x dims vs Small, marginal gain

Analysis:

Why GloVe 25d ≈ 50d ≈ 100d?
All GloVe variants rely on order-blind word averaging with no contextual or syntactic awareness. Increasing dimensionality cannot overcome these architectural limitations—adding dimensions is like adding buttons to a calculator instead of upgrading to a computer
Why Sentence Transformer (384d) > GloVe (100d)?
Sentence Transformers use context-aware representations and attention mechanisms to capture word relationships. As a result, they model semantic meaning rather than relying on keyword matching.
Why OpenAI Large ≈ OpenAI Small?Both models already capture semantic structure effectively. Increasing dimensionality adds subtle nuance but does not introduce new capabilities, leading to diminishing returns.
Conclusion: Once a model captures context and structure, **additional dimensions yield marginal gains**.

3. Speed/Response Time Analysis

Measured on my test runs:

Model	Avg Response Time	Scalability	Cost
GloVe 25d	~0.01s	Excellent	Free
GloVe 50d	~0.02s	Excellent	Free
GloVe 100d	~0.03s	Excellent	Free
Sentence Transformer	~0.15s	Good	Free
OpenAI Small	~0.5-1.0s	Fair	$0.00002/1K tokens
OpenAI Large	~0.7-1.2s	Fair	$0.00013/1K tokens

- **GloVe:** Extremely fast, but limited semantic understanding
- **Sentence Transformer:** Best balance of speed, cost, and accuracy
  OpenAI: Highest quality, slower due to API latency and cost

Recommendation: For most applications, **Sentence Transformers** provide the best trade-off.

4. Word Order Sensitivity Test

Recommended Test for "chocolate milk" vs "milk chocolate":

Categories: Beverages, Candy, Desserts, Food

This gives clear distinction:

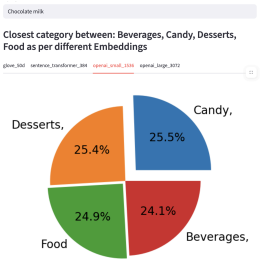
- "Chocolate milk" should match → Beverages (it's a drink)
- "Milk chocolate" should match → Candy (it's chocolate candy)

Actual Results from My Implementation

Input: "Chocolate milk"

Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_50d	Food	2.1909
1	sentence_transformer_384	Beverages,	1.6179
2	openai_small_1536	Candy,	1.4237
3	openai_large_3072	Beverages,	1.4570



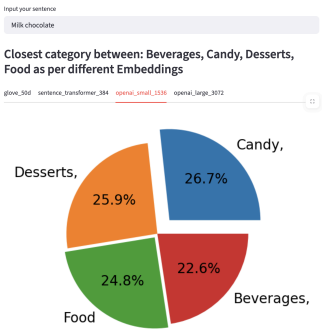
Interpretation:

- **Sentence Transformer** and **OpenAI Large** correctly classified *“chocolate milk”* as **Beverages**, which aligns with real-world meaning.
- **GloVe** predicted **Food**, reflecting its tendency to collapse meaning into a broad, generic category due to word averaging.
- **OpenAI Small** incorrectly chose **Candy**, indicating partial semantic understanding but weaker compositional reasoning compared to the larger model.

Input: "Milk chocolate"

Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_50d	Food	2.1909
1	sentence_transformer_384	Candy,	1.6257
2	openai_small_1536	Candy,	1.4432
3	openai_large_3072	Candy,	1.4379



Interpretation:

- All **context-aware models** correctly classified *“milk chocolate”* as **Candy**.
- **GloVe again returned the same result and confidence score** as the previous input, showing **complete insensitivity to word order**.

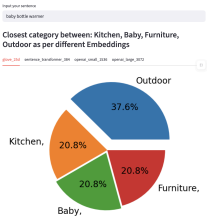
Part C: Real-World ApplicationsExample Group 1: Kitchen vs Baby Products

Categories: Kitchen, Baby, Furniture, Outdoor

Test Pair 1: "baby bottle warmer"

Detailed Comparison

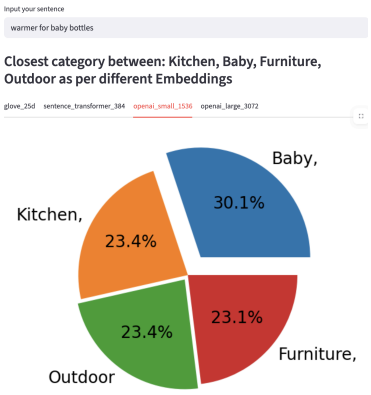
	Model	Top Category	Confidence Score
0	glove_25d	Outdoor	1.8051
1	sentence_transformer_384	Baby,	1.3889
2	openai_small_1536	Baby,	1.4067
3	openai_large_3072	Baby,	1.2490



Test Pair 2: "warmer for baby bottles"

Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_25d	Outdoor	1.8663
1	sentence_transformer_384	Baby,	1.2214
2	openai_small_1536	Baby,	1.4805
3	openai_large_3072	Baby,	1.2699



Two reordered phrases (“baby bottle warmer” and “warmer for baby bottles”) were tested. Sentence Transformer and OpenAI models correctly classified both as **Baby**, demonstrating robustness to word reordering. GloVe (25d) misclassified both as **Outdoor**, reflecting its reliance on word averaging and lack of structural understanding. This shows that context-aware embeddings handle real-world paraphrasing more reliably.

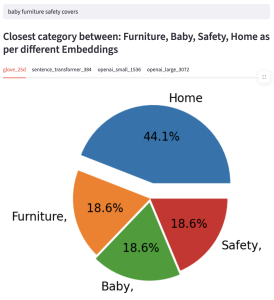
Example Group 2: Furniture vs Safety

Categories: Furniture, Baby, Safety, Home

Test Pair 1: "baby furniture safety covers"

Detailed Comparison

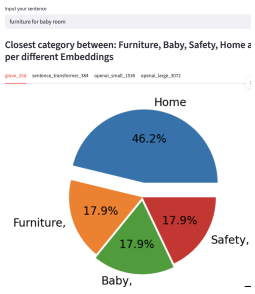
	Model	Top Category	Confidence Score
0	glove_25d	Home	2.3700
1	sentence_transformer_384	Furniture,	1.5673
2	openai_small_1536	Furniture,	1.4641
3	openai_large_3072	Furniture,	1.4614



Test Pair 2: "furniture for baby room"

Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_25d	Home	2.5778
1	sentence_transformer_384	Furniture,	1.9328
2	openai_small_1536	Furniture,	1.7669
3	openai_large_3072	Furniture,	1.6327

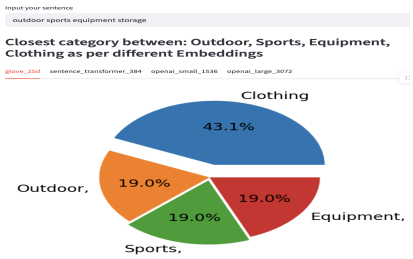


furniture for baby room" and "baby furniture safety covers."

Transformer and OpenAI models consistently predicted **Furniture**, correctly identifying the primary product type. GloVe (25d) predicted **Home** for both cases, collapsing to a broad category due to strong co-occurrence of generic words like *room* and *home*. This highlights GloVe's difficulty in handling multi-word structure and hierarchical meaning.

Example Group 3: Sports vs Outdoor:Categories: Outdoor, Sports, Equipment, Clothing

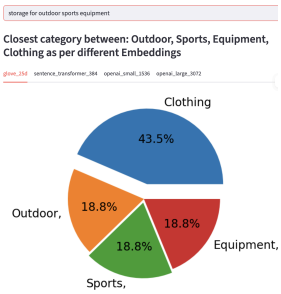
Test Pair 1: "outdoor sports equipment storage"



Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_25d	Clothing	2.2741
1	sentence_transformer_384	Equipment,	1.5096
2	openai_small_1536	Outdoor,	1.7297
3	openai_large_3072	Outdoor,	1.6858

Test Pair 2: "storage for outdoor sports equipment"



Detailed Comparison

	Model	Top Category	Confidence Score
0	glove_25d	Clothing	2.3106
1	sentence_transformer_384	Equipment,	1.4782
2	openai_small_1536	Outdoor,	1.6850
3	openai_large_3072	Outdoor,	1.5767

For *"outdoor sports equipment storage"* and *"storage for outdoor sports equipment"*, word order had minimal effect since both phrases express the same meaning.

Sentence Transformer emphasized the **Equipment** being stored, while OpenAI models emphasized the **Outdoor** usage context. GloVe (25d) incorrectly predicted **Clothing** in both cases, showing category drift caused by order-blind averaging. This demonstrates how different models prioritize object type versus contextual environment.