Report: Word Embeddings Analysis

Ayush Kumar Gupta
2023114001

February 12, 2026

1 Introduction

This report documents the implementation and analysis of word embeddings using pre-trained GloVe vectors (50-dimensional). The objective was to compute similarity metrics between word pairs and analyze semantic relationships through analogies.

Two primary metrics were implemented:

- **Cosine Similarity:** Measures the cosine of the angle between two vectors (captures orientation).
- **Euclidean Distance:** Measures the straight-line distance between two points in the vector space (captures magnitude).

2 Task B: Data Collection

The table below summarizes the computed metrics for the provided word pairs. The *Cosine Similarity* was verified against reference values, and the *Euclidean Distance* was calculated using the implemented function.

Table 1: Comparison of Cosine Similarity and Euclidean Distance

Word 1	**Word 2**	**Cosine Similarity**	**Euclidean Distance**
table	desk	0.5631	4.7041
football	baseball	0.7990	3.7186
water	fire	0.6160	4.9175
computer	calculator	0.5805	5.0053
number	math	0.3923	6.1206
boy	girl	**0.9327**	**2.0426**
sad	happy	0.6890	3.8399
good	bad	0.7964	3.3189
turkey	television	0.3478	6.4884
awesome	great	0.5445	4.5787
coffee	giraffe	0.0396	6.4400
cat	barcelona	0.0288	6.8339
school	disaster	0.2852	6.5456

3 Task C: Analysis

3.1 Highest Similarity

Question: Which 3 pairs have the highest cosine similarity? Does this align with human intuition?

Result:

1. **boy, girl** (0.9327)
2. **football, baseball** (0.7990)
3. **good, bad** (0.7964)

Analysis: Yes, these results align well with human intuition.

- "Boy" and "girl" denote humans of the same age group, sharing almost identical semantic contexts except for gender.
- "Football" and "baseball" are both sports involving balls, sharing high semantic overlap.
- "Good" and "bad" appear in very similar grammatical and contextual structures (adjectives describing quality), which explains their high similarity despite being antonyms.

3.2 Antonyms (Good vs. Bad)

Question: The pair (good, bad) are opposites, yet they have a high cosine similarity score (approx 0.8). Why do vector embeddings place them close together?

Analysis: Word embeddings like GloVe are trained on the principle of *distributional semantics*: words that appear in similar contexts have similar vector representations. While "good" and "bad" are semantic opposites, they are functionally interchangeable in most sentences (e.g., "It was a good movie" vs. "It was a bad movie"). Because they are surrounded by the same types of words, the model learns that they are related concepts (adjectives of quality), resulting in a high cosine similarity score.

3.3 Analogies

Question: Compute analogies using both Cosine Similarity and Euclidean Distance.

The standard analogy formula used was: $d = b - a + c$.

Table 2: Analogy Results Comparison

Query ($a : b :: c :?$)	**Cosine Result**	**Euclidean Result**
boy : girl :: man : ?	woman	woman
bat : baseball :: ball : ?	basketball	basketball
turkey : turkish :: colombia : ?	colombian	colombian
book : library :: coffee : ?	**heliospheric**	**warehouse**
orange : juice :: apple : ?	**juices**	**processor**

Comparison Note: For strong semantic relationships (like gender or country-nationality), both Cosine and Euclidean methods produced identical, correct results. However, for more ambiguous relationships (like book-library), the methods diverged.

- The Cosine method often prioritizes the *direction* of the relationship vector.

- The Euclidean method prioritizes the absolute *proximity* in the vector space.

The divergence in the "book" and "orange" examples highlights that the "nearest neighbor" logic can differ depending on whether we look at angle or distance, especially in lower-dimensional spaces (50d) where vectors might not be perfectly normalized.

3.4 Metric Comparison

Question: Does a higher Cosine Similarity always result in a lower Euclidean Distance for these pairs?

Analysis: Generally yes, but not strictly. In this dataset, the pair *boy-girl* had the highest cosine similarity (0.93) and the lowest Euclidean distance (2.04), showing a strong inverse correlation. However, this is not a mathematical rule unless all vectors are normalized to unit length.

- **Cosine Similarity** depends only on the angle (θ).
- **Euclidean Distance** depends on both the angle and the *magnitude* (length) of the vectors.

If two words have the same orientation but very different frequencies (magnitudes), they could have high cosine similarity (close to 1) but still have a large Euclidean distance.

4 Extension: Nearest Neighbors

Task: Find the Top 5 Nearest Neighbors for the word "computer".

Using Cosine Similarity, the closest words in the 50d GloVe vocabulary are:

1. **computers** (0.9165)
2. **software** (0.8815)
3. **technology** (0.8526)
4. **electronic** (0.8126)
5. **internet** (0.8060)