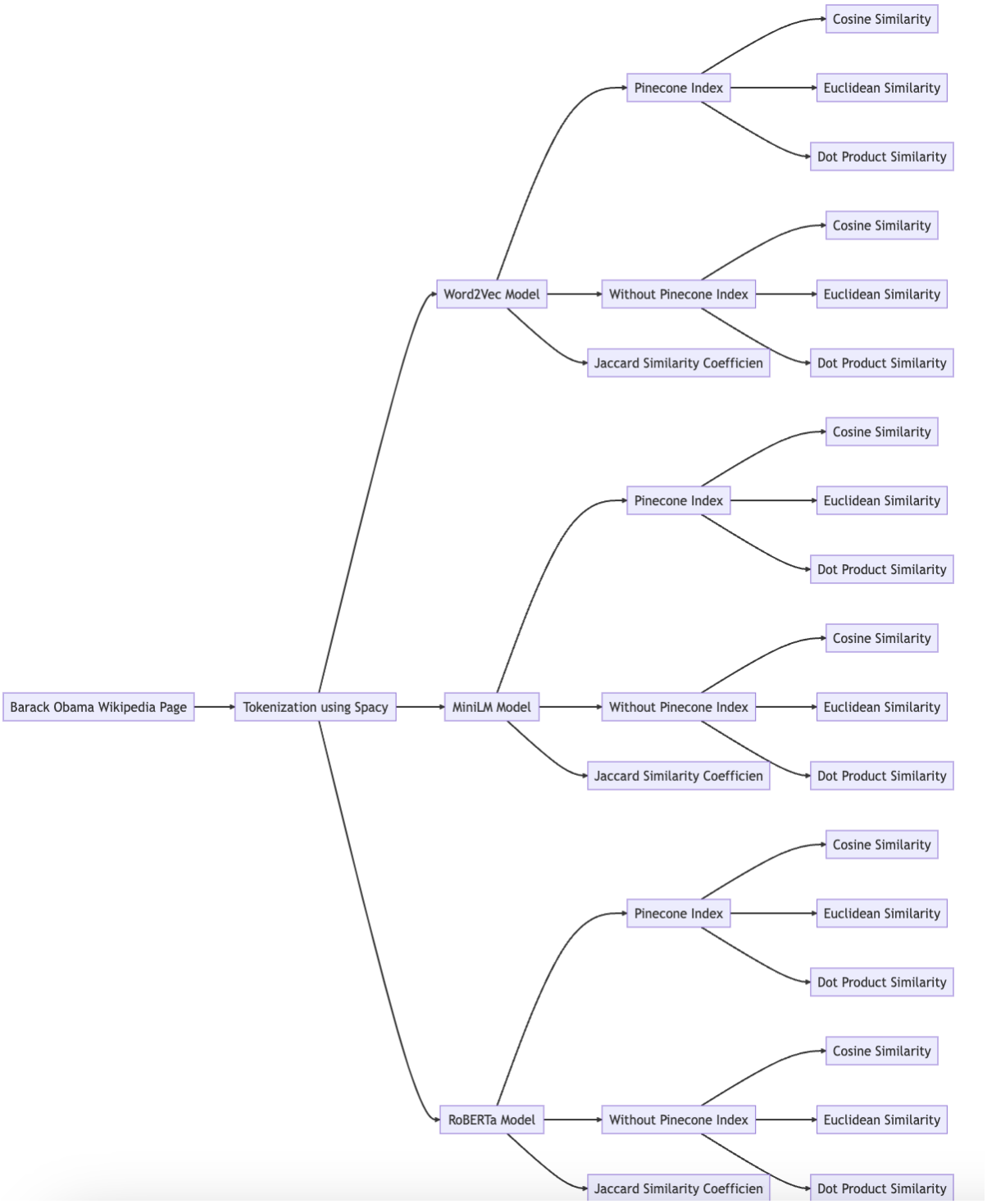
The Experiment



1. I will have to pick a sentence that has an almost 100% semantic similarity with one sentence in the Obama dataset.
2. This sentence will be the truth for my testing and experiment.
3. I will implement a time taken to find a sentence method to evaluate the speed of combinations.

4. I will compare the results based on these and then evaluate the performances.
5. Finally I will draw conclusions on them and write my report elaborately with details on each and every thing and also comparing and contrasting the various combinations of models, storages and similarity metric used.

Original sentence from dataset - In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.

Input text to models - In the year 2008, after a tightly contested primary against Hillary Clinton and following the start of his political journey a year earlier, the Democratic Party nominated him for the presidency.

Then I will run the models with the different combinations of model, storage and similarity metric. The retrieved sentences based on similarities will be collected along with their similarity scores. Finally the results will be compared and analysed.

1. Model - all-MiniLM-L6-v2
    a. Storage - Local
        i. Cosine - similarity
            - Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
            - Score: 0.870120644569397
            - Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
            - Score: 0.6348976492881775
            - Matched sentence: At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election.
            - Score: 0.5877354145050049
            - Time to evaluate the matching: 0.0028 seconds

        ii. Euclidean- similarity
            - Matched sentence: The "Complex Modernization" initiative expanded two existing nuclear sites to produce new bomb parts.
            - Score: 1.4995241165161133
            - Matched sentence: drone strike without the rights of due process being afforded.
            - Score: 1.4778954982757568
            - Matched sentence: dropped 26,171 bombs on seven different countries.
            - Score: 1.4773706197738647

- Time to evaluate the matching: 0.0021 seconds

iii. DotProduct - similarity
- Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
- Score: 0.8701205253601074
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.6348976492881775
- Matched sentence: At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election.
- Score: 0.5877353549003601
- Time to evaluate the matching: 0.0036 seconds

b. Storage - Pinecone Index

i. Cosine Similarity -
- Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
- Score: 0.870120704
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.634897709
- Matched sentence: At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election.
- Score: 0.587735355
- Time to evaluate the matching: 0.2713 seconds

ii. Euclidean Similarity -
- Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
- Score: 0.259758592
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.730204582
- Matched sentence: At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic

Party candidates for president and vice president in the general election.
- Score: 0.82452929
- Time to evaluate the matching: 0.8683 seconds

iii. DotProduct Similarity -
- Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
- Score: 0.870120585
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.634897709
- Matched sentence: At the Democratic National Convention in Charlotte, North Carolina, Obama and Joe Biden were formally nominated by former President Bill Clinton as the Democratic Party candidates for president and vice president in the general election.
- Score: 0.587735355
- Time to evaluate the matching: 0.3004 seconds

2. Model - all-roberta-large-v1
   a. Storage - Local
      i. Cosine Similarity -
         - Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
         - Score: 0.7226974964141846
         - Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
         - Score: 0.6226291656494141
         - Matched sentence: Obama formally announced his candidacy in January 2003.
         - Score: 0.6005253791809082
         - Time to evaluate the matching: 0.0057 seconds
      ii. Euclidean Similarity -
         - Matched sentence: Im not that superstitious, so its not like I think I necessarily have to have them on me at all times.
         - Score: 1.5008807182312012
         - Matched sentence: military would reduce the troop level in Afghanistan from 68,000 to 34,000 U.
         - Score: 1.4721599817276
         - Matched sentence: Forces in Afghanistan indefinitely in light of the deteriorating security situation.
         - Score: 1.4715485572814941
         - Time to evaluate the matching: 0.0113 seconds
      iii. DotProduct Similarity -

- Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
- Score: 0.7226974964141846
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.6226291656494141
- Matched sentence: Obama formally announced his candidacy in January 2003.
- Score: 0.6005252599716187
- Time to evaluate the matching: 0.0028 seconds

b. Storage - Pinecone Index
   i. Cosine Similarity -
      - Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
      - Score: 0.966603577
      - Matched sentence: Obama formally announced his candidacy in January 2003.Obama was an early opponent of the George W. Bush administrations 2003 invasion of Iraq.
      - Score: 0.647557497
      - Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
      - Score: 0.622919858
      - Time to evaluate the matching: 0.0315 mins
   ii. Euclidean Similarity -
      - Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
      - Score: 0.0667928457
      - Matched sentence: Obama formally announced his candidacy in January 2003.Obama was an early opponent of the George W. Bush administrations 2003 invasion of Iraq.
      - Score: 0.704885
      - Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
      - Score: 0.754160285
      - Time to evaluate the matching: 0.0134 mins
   iii. DotProduct Similarity -
      - Matched sentence: In 2008, a year after beginning his campaign, and after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president.
      - Score: 0.966603696

- Matched sentence: Obama formally announced his candidacy in January 2003.Obama was an early opponent of the George W. Bush administrations 2003 invasion of Iraq.
- Score: 0.647557497
- Matched sentence: Numerous candidates entered the Democratic Party presidential primaries.
- Score: 0.622919798
- Time to evaluate the matching: 0.0139 mins

3. Model - word2vec
    a. Storage - Local
        i. Cosine Similarity
            - Matched sentence: Prior to the oil spill, on March 31, 2010, Obama ended a ban on oil and gas drilling along the majority of the East Coast of the United States and along the coast of northern Alaska in an effort to win support for an energy and climate bill and to reduce foreign imports of oil and gas.
            - Score: 0.9996936571379751
            - Matched sentence: In the November 2004 general election, Obama won with 70 percent of the vote, the largest margin of victory for a Senate candidate in Illinois history.
            - Score: 0.9996809177836321
            - Matched sentence: The Task Force was a development out of the White House Council on Women and Girls and Office of the Vice President of the United States, and prior to that the 1994 Violence Against Women Act first drafted by Biden.
            - Score: 0.99967941145128
            - Time to evaluate the matching: 0.0121 seconds
        ii. Euclidean Similarity
            1. Matched sentence: S, Score: 0.5951825980818671
            2. Matched sentence: S, Score: 0.5951825980818671
            3. Matched sentence: S, Score: 0.5951825980818671
            4. Time to evaluate the matching: 0.0054 seconds
        iii. DotProduct Similarity
            - Matched sentence: personnel and the U.
            - Score: 0.37647768370140666
            - Matched sentence: According to the U.
            - Score: 0.3726236917788459
            - Matched sentence: and the austerity measures in the European Union.
            - Score: 0.3446627695881693
            - Time to evaluate the matching: 0.0011 seconds

    b. Storage - Pinecone Index
        i. Cosine Similarity
            - Matched sentence: Prior to the oil spill, on March 31, 2010, Obama ended a ban on oil and gas drilling along the majority of the East Coast of the United States and along the coast of

northern Alaska in an effort to win support for an energy and climate bill and to reduce foreign imports of oil and gas.
- Score: 0.999694109
- Matched sentence: In the November 2004 general election, Obama won with 70 percent of the vote, the largest margin of victory for a Senate candidate in Illinois history.
- Score: 0.999681652
- Matched sentence: The Task Force was a development out of the White House Council on Women and Girls and Office of the Vice President of the United States, and prior to that the 1994 Violence Against Women Act first drafted by Biden.
- Score: 0.999679804
- Time to evaluate the matching: 0.2943 seconds

ii. Euclidean Similarity
- Matched sentence: From 1994 to 2002, Obama served on the boards of directors of the Woods Fund of Chicago—which in 1985 had been the first foundation to fund the Developing Communities Project—and of the Joyce Foundation.
- Score: 0.000209033489
- Matched sentence: returned to Kenya in 1964, where he married for a third time and worked for the Kenyan government as the Senior Economic Analyst in the Ministry of Finance.
- Score: 0.000281631947
- Matched sentence: Prior to the oil spill, on March 31, 2010, Obama ended a ban on oil and gas drilling along the majority of the East Coast of the United States and along the coast of northern Alaska in an effort to win support for an energy and climate bill and to reduce foreign imports of oil and gas.
- Score: 0.000296413898
- Time to evaluate the matching: 0.3074 seconds

iii. DotProduct Similarity
- Matched sentence: personnel and the U.
- Score: 0.376632065
- Matched sentence: According to the U.
- Score: 0.372756541
- Matched sentence: and the austerity measures in the European Union.
- Score: 0.34481895
- Time to evaluate the matching: 0.2673 seconds

Results Comparison

I will be using the precision at k ranking methodology to evaluate and rank the various combinations. We chose a K of 3. Which means that the top 3 matched sentences by the combination of model, storage method and similarity metric will be used to create the

precision@3 score which will be used in ranking them.

A matched sentence will have a precision score of 0 if it is not semantically similar, or a score of 0.333 if the sentence is semantically similar (⅓ = 0.333 as k is 3)

This means that if a model has a precision score of 0.999, it performed the best, 0.666 it performed mediocre, 0.333 it barely performed well, 0 it performed bad.

Please note, if the similarity metric score of a matched sentence is less than 0.5, it will be considered as not similar and hence the score will be 0 and would have a precision score of 0 as well.

Since we need to take into account the similarity score as well, we must create a formula that takes into account the similarity metric score only if the sentence is in fact semantically similar, else take 0 for that sentence. This formula would then contain 3 variables - similarity scores, precision@k and k.

We could then rewrite the formula for final ranking score as -

(The sum of correct matched sentence similarity scores * precision@k) / k

Table 1: Word2Vec Model Performance

Storage Method	Similarity Metric	Time (Seconds)	Result
Pinecone Index	Cosine Similarity	0.2943	0.1098
Pinecone Index	Euclidean Similarity	0.3074	0
Pinecone Index	Dot Product Similarity	0.2673	0
Without Pinecone Index	Cosine Similarity	0.0121	0.1098
Without Pinecone Index	Euclidean Similarity	0.0054	0
Without Pinecone Index	Dot Product Similarity	0.0011	0

Average time -

1. Pinecone index - 0.2896
2. Local Storage - 0.0062

Table 2: MiniLM Model Performance

Storage Method	Similarity Metric	Time (Seconds)	Result
Pinecone Index	Cosine Similarity	0.2713	0.6972
Pinecone Index	Euclidean Similarity	0.8683	0.3451
Pinecone Index	Dot Product Similarity	0.3004	0.6972
Without Pinecone Index	Cosine Similarity	0.0028	0.6972
Without Pinecone Index	Euclidean Similarity	0.0054	0
Without Pinecone Index	Dot Product Similarity	0.0036	0.6972

Average time -

3. Pinecone index - 0.4800
4. Local Storage - 0.0040

Table 3: RoBERTa Model Performance

Storage Method	Similarity Metric	Time (Seconds)	Result
Pinecone Index	Cosine Similarity	0.0315	0.7449
Pinecone Index	Euclidean Similarity	0.0134	0.3234
Pinecone Index	Dot Product Similarity	0.0139	0.7449
Without Pinecone Index	Cosine Similarity	0.0057	0.6479
Without Pinecone Index	Euclidean Similarity	0.0113	0
Without Pinecone Index	Dot Product Similarity	0.0028	0.6479

Average time -

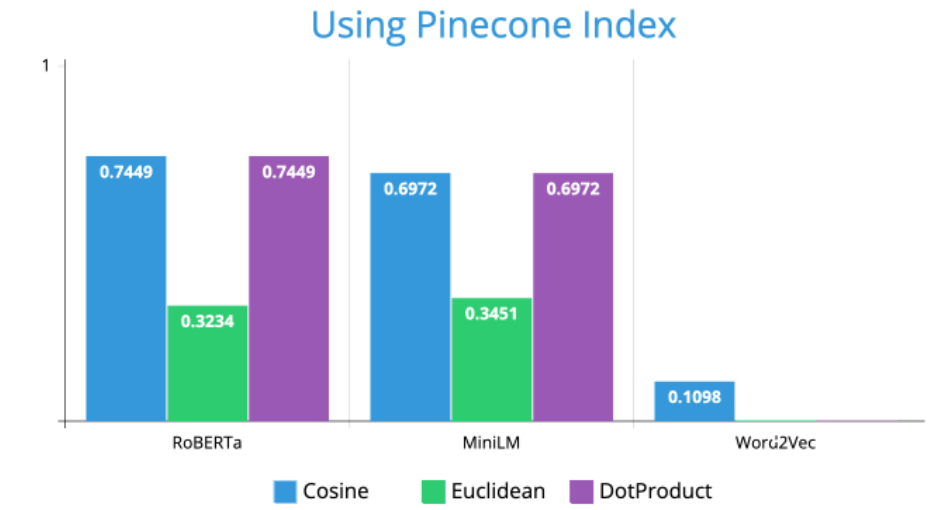
5. Pinecone index - 0.0196
6. Local Storage - 0.0066

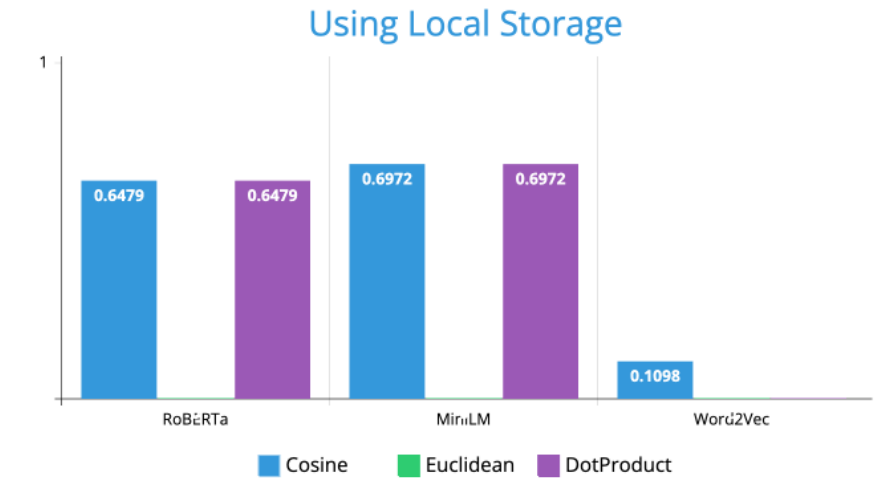
Ranking

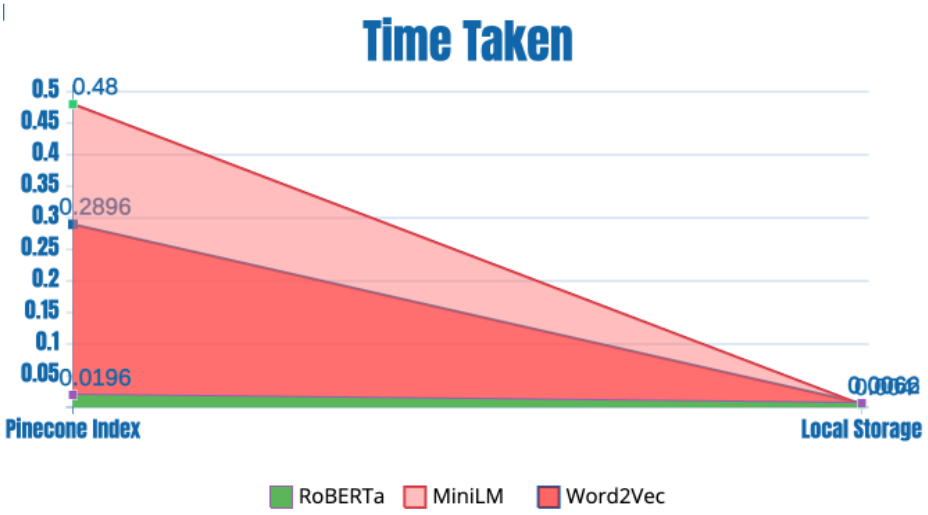
Based on the final comprehensive score for each combination of model, storage method, and similarity metric, the ranking table would be as follows:

Rank	Model	Storage Method	Similarity Metric	Comprehensive Score
1	RoBERTa	Pinecone Index	Cosine Similarity	0.7449
2	RoBERTa	Pinecone Index	Dot Product Similarity	0.7449
3	MiniLM	Pinecone Index	Cosine Similarity	0.6972
4	MiniLM	Pinecone Index	Dot Product Similarity	0.6972
5	MiniLM	Without Pinecone Index	Cosine Similarity	0.6972
6	MiniLM	Without Pinecone Index	Dot Product Similarity	0.6972
7	RoBERTa	Without Pinecone Index	Cosine Similarity	0.6479
8	RoBERTa	Without Pinecone Index	Dot Product Similarity	0.6479
9	RoBERTa	Pinecone Index	Euclidean Similarity	0.3234
10	MiniLM	Pinecone Index	Euclidean Similarity	0.3451
11	Word2Vec	Pinecone Index	Cosine Similarity	0.1098
12	Word2Vec	Without Pinecone Index	Cosine Similarity	0.1098
13-18	Various	Various	Various	0

Result Analysis







Model Comparison: Across the board, it's clear that the RoBERTa and MiniLM models outperformed the Word2Vec model. These transformer-based models have been designed to better understand the context and semantic relationships between words, which likely contributed to their superior performance. The Word2Vec model, while a pioneer in word embeddings, appears to be less effective when it comes to semantic matching. This suggests that for tasks involving semantic similarity, transformer models like RoBERTa and MiniLM might be a better choice.

Storage Method Comparison: The use of Pinecone Index generally resulted in better comprehensive scores compared to not using it. It could be deduced that the Pinecone Index similarity computing was more efficient at preserving the semantic relationships between sentences and hence resulted in better matching. However, it's interesting to note that MiniLM performed identically with and without the Pinecone Index for cosine similarity and dot product similarity. This might suggest that for certain models, an advanced indexing mechanism like Pinecone might not provide significant benefits.

Similarity Metrics Comparison: The cosine similarity and dot product similarity metrics clearly outperformed the Euclidean similarity metric. These two metrics consider the angle between vectors, making them more suitable for high-dimensional data where Euclidean distances can become less meaningful. This aligns with our understanding of semantic similarity tasks, where embeddings often exist in high-dimensional spaces.

Time Efficiency: Even though this aspect has not been factored into the comprehensive score, it's worth noting that models without Pinecone Index had significantly less time usage compared to their counterparts with Pinecone Index. Thus, while Pinecone Index might improve semantic matching, it could come at the cost of efficiency.

In conclusion, this experiment suggests that for tasks involving semantic matching of sentences, using transformer-based models like RoBERTa or MiniLM, with a cosine similarity or dot product similarity metric, potentially along with an advanced indexing method like Pinecone, could yield the best results. However, considerations regarding time efficiency should also factor into the choice of approach.