Assignment 12: Clustering Text Embeddings Obtained from Fine-tuned Language Model

Name: Shiqi Hu

GTID: 904061372

1. Insights

1.1 Analyze the embedding cluster plots

Based on the clustering visualization, which model do you think performs the best (i.e., which model's embeddings are the most effective)? Is this consistent with its F1 score or accuracy?

- BERT appears to perform the best based on both the embedding cluster separations and the evaluation metrics (F1 score: 0.94 and accuracy: 0.94). It shows relatively distinct clusters with a few overlapping areas. This consistency suggests that BERT embeddings are the most effective in capturing the separable structure of the data.

What are your takeways when comparing decoder-only model embeddings with encoder-only model embeddings?

- Decoder-Only Model (GPT-2): GPT-2, as a decoder-only model, generates embeddings that are more contextual and sequential. While it performs well, its embeddings show some overlap, possibly because decoder models are optimized for generation rather than distinguishing class boundaries. Additionally, the high overlap in GPT-2's clusters might also be influenced by the PCA dimensionality reduction. Since GPT-2 embeddings capture many complex sequential dependencies, PCA might have removed some of these important features, leading to a loss in class-distinct information and resulting in clusters that appear more merged and less defined.
- Encoder-Only Models (BERT, SBERT, Longformer): Encoder models like BERT are optimized for extracting contextualized embeddings, which often capture distinctions between classes more effectively. BERT, in particular, shows clear separations between clusters, likely due to its bidirectional attention mechanism that excels at understanding relationships within the entire input context.

Some clusters are distinct, while others overlap with each other. What does this overlap indicate?

- Overlapping clusters indicate that certain embeddings are not easily distinguishable between classes. This overlap might mean that some classes are inherently similar in content or context, making it challenging for the model to differentiate them.
- For instance, Manufacturing and Wholesale Trade tend to overlap across models, as these sectors share characteristics, making them harder for the model to differentiate.
- This overlap may also highlight limitations in the model's ability to capture subtle distinctions, indicating a need for further tuning to improve class separation.

1.2 Analyze the K-Means Cluster Plots (3 cluster centroids)

Based on the K-Means clustering visualization, which model do you think performs the best? Is this consistent with its F1 score or accuracy?

- GPT-2 shows the clearest separation with minimal overlap between the 3 clusters, suggesting it captures broad, distinct groupings effectively.
- However, this clustering performance does not perfectly align with its slightly lower F1 score compared to BERT, indicating that while GPT-2 groups data well, BERT may still excel at fine-grained classification.

What are your takeways when comparing decoder-only model embeddings with encoder-only model embeddings?

- Decoder-only models like GPT-2 are effective at generating distinct groups in embeddings, which benefits broader clustering.
- Encoder-only models (e.g., BERT) produce embeddings optimized for contextual distinctions, often enhancing accuracy in detailed classification tasks but not always achieving as clear separations in broader clusters.

Some clusters are distinct, while others overlap with each other. What does this overlap indicate?

- Overlapping clusters suggest some similarity or ambiguity between classes, which the model struggles to fully distinguish. This could indicate that certain classes may share overlapping characteristics, highlighting either natural similarities in the data or limitations in the model's ability to separate subtle differences.

1.3 Provide Comments on What You Observe from Equal Weighted Portfolio Cumulative Returns vs Market Plot

Portfolio Performance Relative to Market:

- Portfolio 2 outperformed both the market and the other portfolios,showing a significant rise in cumulative returns, particularly after 2017.
- Portfolio 0 consistently underperformed compared to the market, with returns generally below the market trend throughout the observed period.
- Portfolio 1 also underperformed compared to the market, showing relatively flat returns over time.

Industry Composition and Performance Implications:

- Portfolio 2 has a high concentration in sectors like Transportation and other Utilities & Finance, Insurance and Real Estate, which might have contributed to its strong performance, especially if these sectors experienced growth in the observed period.
- Portfolio 0 includes significant representation from Manufacturing , Wholesale Trade, and Retail Trade. Despite having a balanced mix, the industries in this portfolio underperformed relative to the market, possibly due to slowdowns in traditional manufacturing or wholesale trade.
- Portfolio 1 contains a large number of Mining stocks, which might have contributed to its underperformance, as the mining sector can be volatile and dependent on global commodity cycles.

1.4 Provide Comments on What You Observe from Value Weighted Portfolio Cumulative Returns vs Market Plot

Portfolio Performance Relative to Market:

- Portfolio 0 significantly outperformed both the market and the other portfolios, especially from 2016 onwards, indicating that larger companies in this portfolio drove substantial growth.
- Portfolio 2 also performed better than the market but with less pronounced growth than Portfolio 0.
- Portfolio 1 had the weakest performance, remaining relatively flat over time with returns generally tracking below the market.

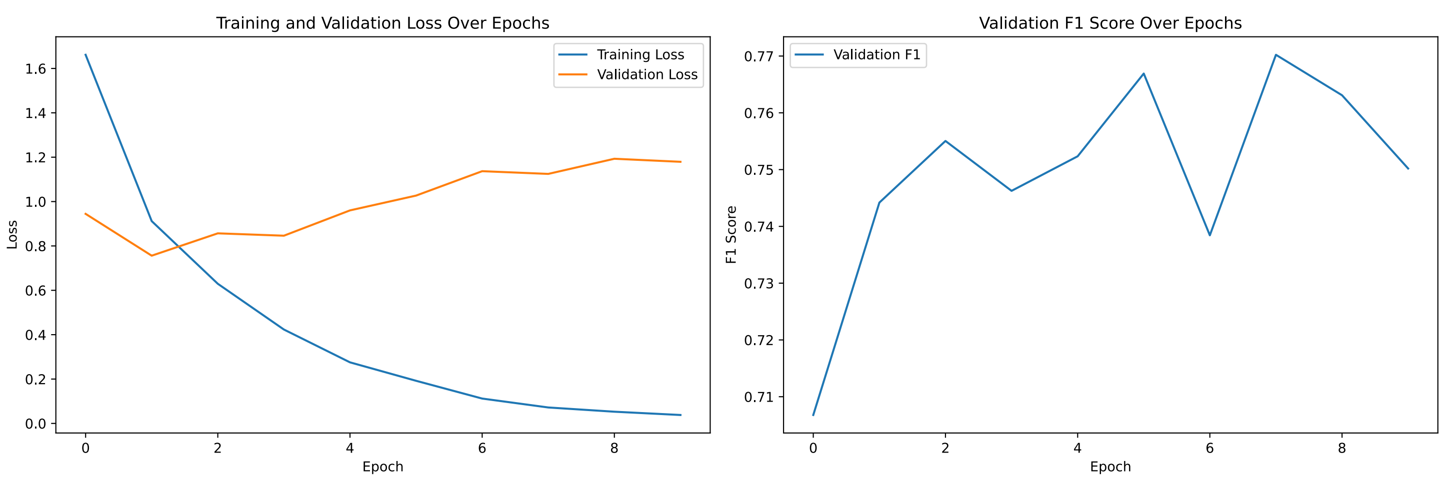
Industry Composition and Performance Implications Under Value-Weighting:

- In a value-weighted portfolio, larger-cap companies have more influence on performance. The strong outperformance of Portfolio 0 suggests that this portfolio includes high-performing large-cap companies, possibly in sectors like Manufacturing and Retail, which had robust growth in recent years.
- Portfolio 1, with a high concentration in Mining, may have struggled due to market volatility in that sector, with larger companies in this portfolio not performing well over the observed period.

2. Fine-tuning BERT

After fine-tuning BERT with provided hyper-parameter, the best validation F1 score is 0.7702.

Plot training loss, validation loss, and validation metric (F1) over epochs



3. Inference using Fine-tuned Models

Report classification metric (F1 or Accuracy) for each model in a table.

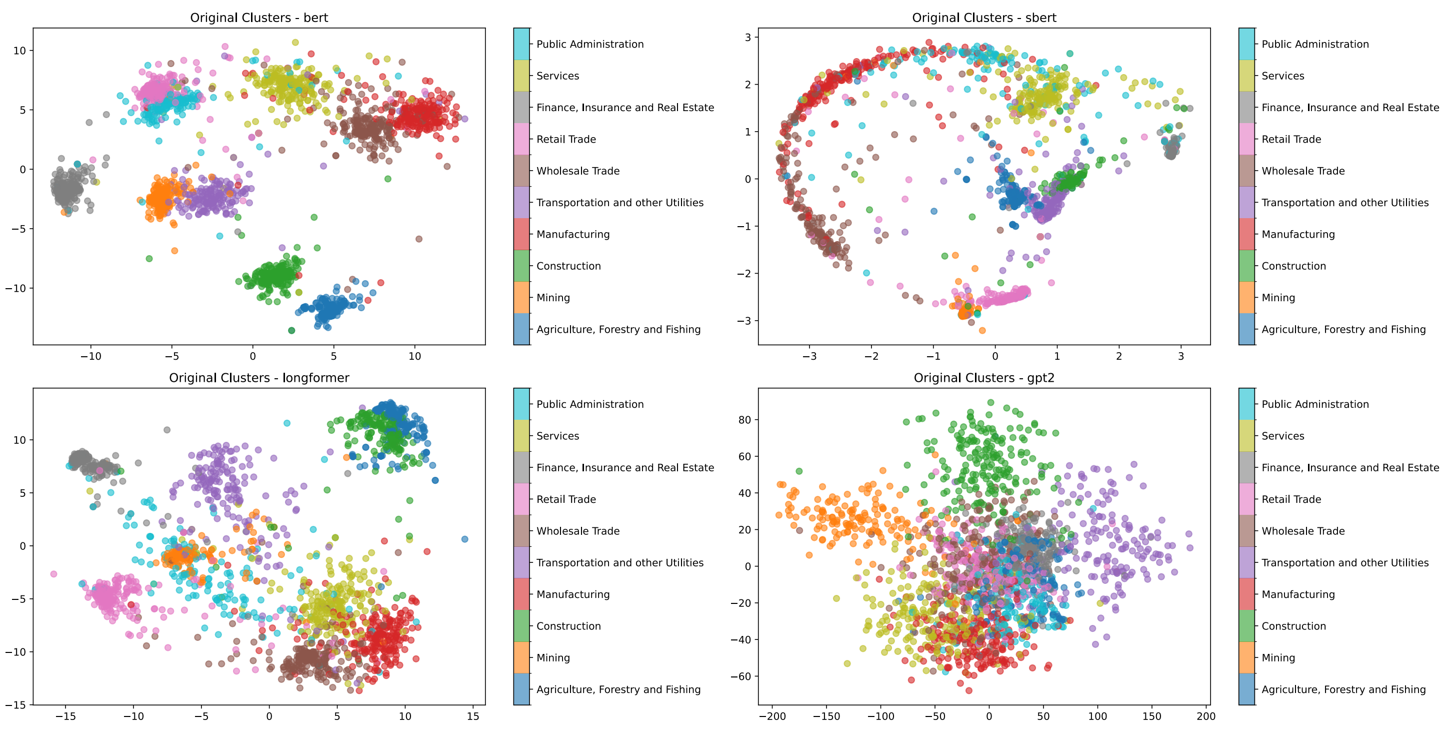
```
Model Performance Comparison:
```

	Model	F1 Score	Accuracy
0	bert	0.944214	0.944359
1	sbert	0.833523	0.839773
2	longformer	0.909898	0.910871
3	gpt2	0.935901	0.935600

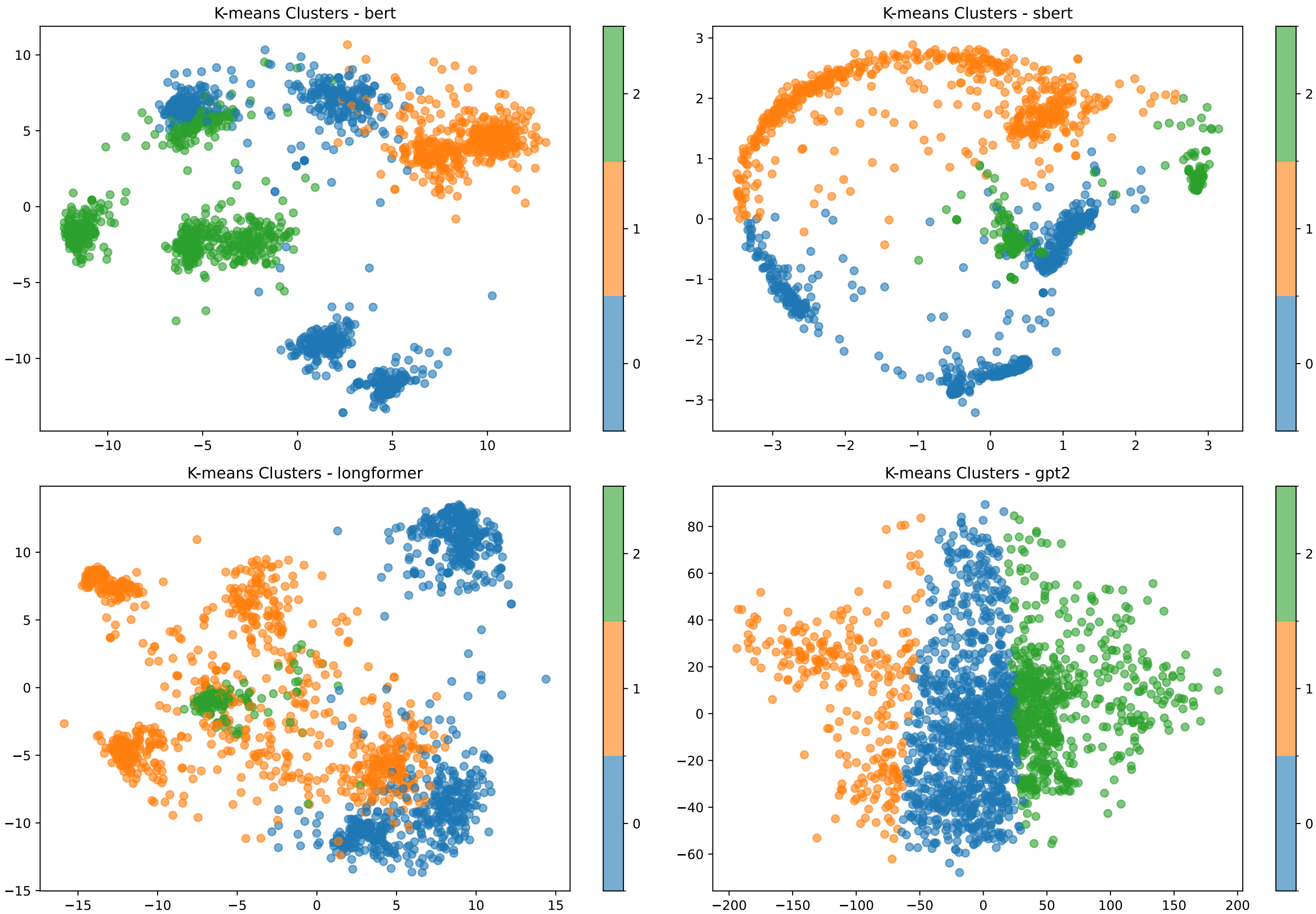
```
Saved embeddings for bert
Saved embeddings for sbert
Saved embeddings for longformer
Saved embeddings for gpt2
```

4. Clustering using Embeddings

Plot the embedding cluster for each model



K-means Clustering Visualization



5. Portfolio Analysis

Regardless of the original 10 class PCA clustering, GPT-2 achieved the best cluster separation under K-Means clustering with K=3, it appears suitable for distinguishing between broader, high-level categories. The strong separation in the K-Means result suggests that GPT-2 embeddings capture underlying patterns or groupings that align well with the high-level clusters needed for portfolio analysis. So I choose GPT-2 for portfolio analysis.

```
Clustered Business Descriptions:
```

	filing_date	item_1	PERMNO	year	label	cluster
2599	1996-02-23	ITEM 1. BUSINESS\nGENERAL The Company is a bui...	70092	1996	2.0	0
2709	1996-03-13	ITEM 1. BUSINESS\nGeneral\nEMCOR Group, Inc. (...	82694	1996	4.0	2
2835	1996-03-22	ITEM 1. BUSINESS.\nGENERAL\nUSA Waste Services...	11955	1996	9.0	2
2844	1996-03-22	ITEM 1. DESCRIPTION OF BUSINESS\nPolaris Indus...	75182	1996	6.0	0
2962	1996-03-26	Item 1. Business.\nPrincipal Products\nThe reg...	16468	1996	0.0	2
...	...	...	...	...	...	...
110249	2022-12-13	Item 1. Business\nAlico, Inc. (“Alico”) was in...	11790	2022	0.0	2
110274	2022-12-19	ITEM 1\nBUSINESS\nBusiness Overview\nHovnanian...	65285	2022	2.0	0
110275	2022-12-19	ITEM 1. BUSINESS\nToll Brothers, Inc., a corpo...	70228	2022	2.0	0
110279	2022-12-20	Item 1. Business\nGeneral development of the b...	89447	2022	0.0	2

	filing_date	item_1	PERMNO	year	label	cluster
110281	2022-12-21	ITEM 1. BUSINESS\nOverview\nNeuBase Therapeuti...	13965	2022	7.0	2

1941 rows × 6 columns

```
Monthly msf data with cluster for selected companies:
```

	PERMNO	date	SHRCD	SICCD	TICKER	COMNAM	PERMCO	CUSIP	BIDLO	ASKHI	...	SHROUT
0	10019	1996-01-31	11	3610	IFRS	I F R SYSTEMS INC	7971	44950710	9.3750	11.875	...	5472.0
1	10019	1996-02-29	11	3610	IFRS	I F R SYSTEMS INC	7971	44950710	11.6250	12.625	...	5472.0
2	10019	1996-03-29	11	3610	IFRS	I F R SYSTEMS INC	7971	44950710	11.7500	13.750	...	5503.0
3	10019	1996-04-30	11	3610	IFRS	I F R SYSTEMS INC	7971	44950710	12.6875	15.625	...	5503.0
4	10019	1996-05-31	11	3610	IFRS	I F R SYSTEMS INC	7971	44950710	14.0000	16.000	...	5503.0
...	...	...	...	...	...	...	...	...	...	...	...	...
15300	93428	2012-08-31	11	9999	BSFT	BROADSOFT INC	53446	11133B40	23.8400	39.870	...	27590.0
15301	93428	2012-09-28	11	9999	BSFT	BROADSOFT INC	53446	11133B40	37.2600	42.600	...	27811.0
15302	93428	2012-10-31	11	9999	BSFT	BROADSOFT INC	53446	11133B40	34.7400	40.010	...	27811.0
15303	93428	2012-11-30	11	9999	BSFT	BROADSOFT INC	53446	11133B40	29.8100	38.960	...	27835.0
15304	93428	2012-12-31	11	9999	BSFT	BROADSOFT INC	53446	11133B40	31.6900	36.330	...	27913.0

15305 rows × 25 columns

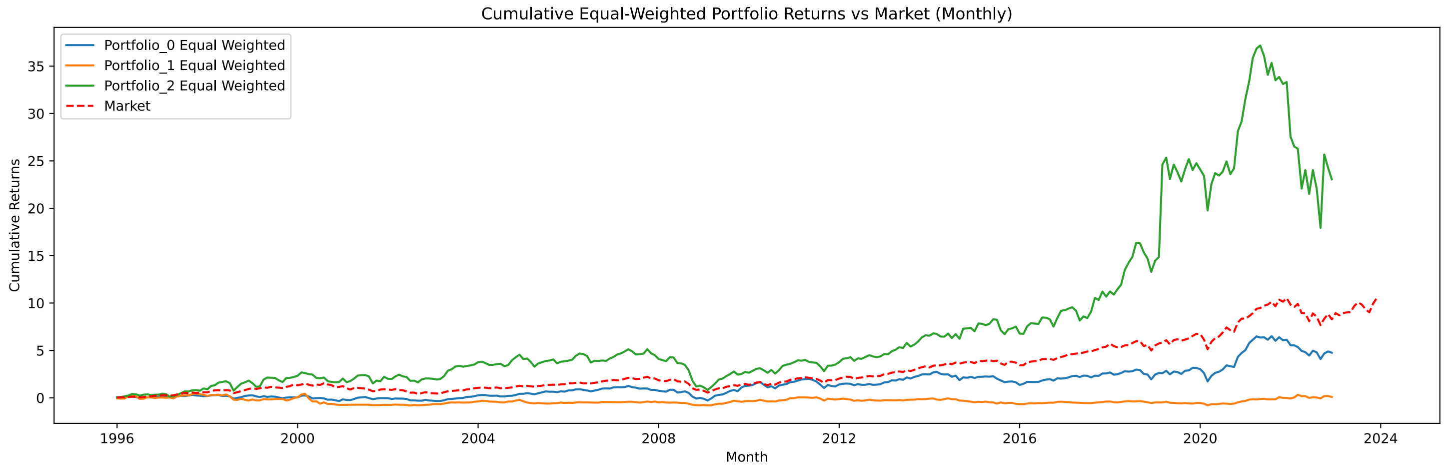
Cumulative Returns:

```
{'Portfolio_0': {'EqualWeighted': Month
  1996-01    0.062298
  1996-02    0.066330
  1996-03    0.086639
  1996-04    0.103986
  1996-05    0.146954
               ...
  2022-08    4.792494
  2022-09    4.082565
  2022-10    4.676993
  2022-11    4.896639
  2022-12    4.756046
  Freq: M, Name: RET, Length: 324, dtype: float64,
  'ValueWeighted': Month
  1996-01      0.025025
  1996-02      0.021245
  1996-03      0.047773
  1996-04      0.057906
  1996-05      0.102764
                 ...
  2022-08    138.289039
  2022-09    127.465530
  2022-10    142.696304
  2022-11    151.870974
  2022-12    144.274516
  Freq: M, Length: 324, dtype: float64},
 'Portfolio_1': {'EqualWeighted': Month
  1996-01   -0.046667
  1996-02   -0.093334
  1996-03   -0.072934
  1996-04    0.146765
  1996-05    0.249030
               ...
  2022-08    0.035453
  2022-09   -0.058957
  2022-10    0.180781
  2022-11    0.191237
  2022-12    0.095449
  Freq: M, Name: RET, Length: 324, dtype: float64,
  'ValueWeighted': Month
  1996-01   -0.046667
  1996-02   -0.093334
  1996-03   -0.085139
  1996-04    0.239698
  1996-05    0.331180
               ...
  2022-08    3.530293
  2022-09    3.081534
  2022-10    3.741607
  2022-11    3.867555
  2022-12    3.753575
  Freq: M, Length: 324, dtype: float64},
 'Portfolio_2': {'EqualWeighted': Month
  1996-01     0.032105
  1996-02     0.077142
  1996-03     0.160438
  1996-04     0.238484
  1996-05     0.400046
               ...
```

```
2022-08    22.067410
2022-09    17.921875
2022-10    25.677317
2022-11    24.274211
2022-12    23.049817
Freq: M, Name: RET, Length: 324, dtype: float64,
'ValueWeighted': Month
1996-01     0.068975
1996-02     0.128995
1996-03     0.302026
1996-04     0.360149
1996-05     0.564841
              ...
2022-08    31.232649
2022-09    28.570031
2022-10    27.893108
2022-11    27.223170
2022-12    25.794064
Freq: M, Length: 324, dtype: float64}}
```

```
Industry label counts for cluster = 0:
label
Manufacturing                        1692
Retail Trade                         1441
Wholesale Trade                      1384
Construction                         1250
Services                             1046
Public Administration                 619
Agriculture, Forestry and Fishing     581
Finance, Insurance and Real Estate    364
Mining                                 96
Transportation and other Utilities     64
Name: count, dtype: int64
Industry label counts for cluster = 1:
label
Mining                              1285
Services                             580
Construction                         156
Wholesale Trade                       96
Manufacturing                         84
Retail Trade                          48
Public Administration                 12
Agriculture, Forestry and Fishing     12
Name: count, dtype: int64
Industry label counts for cluster = 2:
label
Transportation and other Utilities    1477
Finance, Insurance and Real Estate     751
Agriculture, Forestry and Fishing      566
Public Administration                  559
Retail Trade                           420
Construction                           314
Wholesale Trade                        168
Manufacturing                          120
Services                                96
Mining                                  24
Name: count, dtype: int64
```

Equal Weighted Portfolio Cumulative Returns vs Market Plot



Value Weighted Portfolio Cumulative Returns vs Market Plot