Analyzing Writing Style of Ali Teoman for Testing Uniqueness of Individual Works

Aqsa Shabbir, 22301043
Kousar 22301044
Burak Ferit Aktan, 22301424
Noor Muhammad, 22301048
Yasir Ali Khan, 22301049

Contents

Problem Description

Background

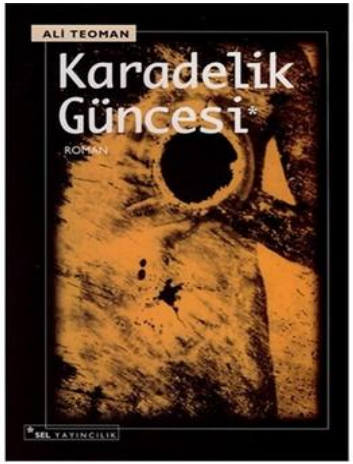
- Books showcase unique styles
- Each work stands apart

Our Aim

- Investigate and verify the validity of this claim through a comprehensive analysis.

Ali Teoman Books:

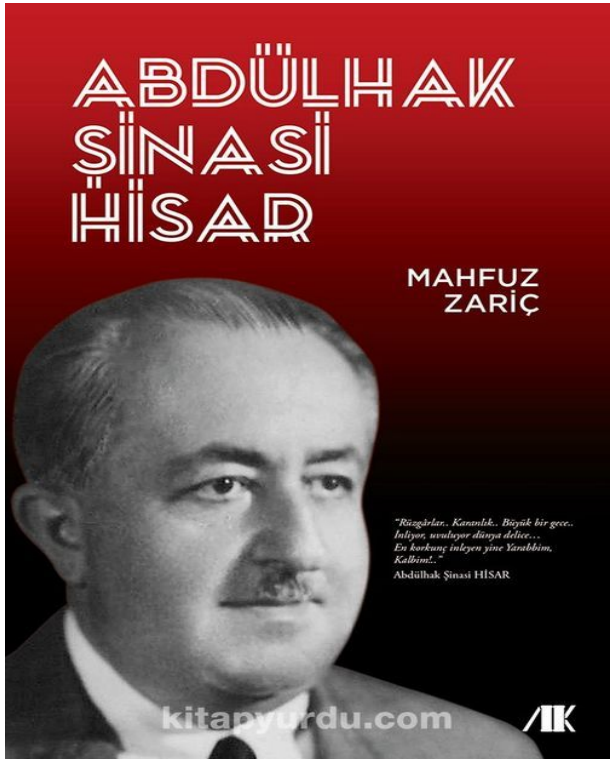
- Karadelik Güncesi
- İnsansız Konak
- Gizli Kalmış Bir İstanbul Masalı



Other Authors:

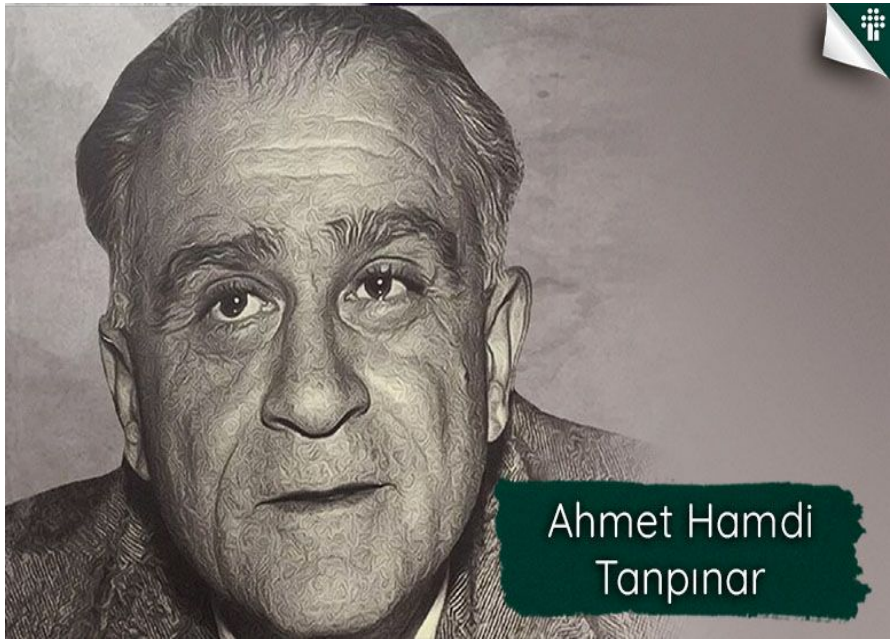
Abdulhak Sinasi Hisar

1. Ali Nizami Beyin Alafrangalığı ve Şeyhliği
2. Çamlıcadaki Enişte mi z
3. Fahim Bey ve Biz



Ahmet Hamdi Tanpinar

- Huzur
- Mahur Beste
- Sahnenin Dışındakiler



Halid Ziya Usakligil

- Aşk-ı Memnu
- Kadın Pençesi
- Kırık Hayatlar

Refik Halid Karay

- Anahtar
- Bu Bizim Hayatimiz
- Bugunun Saraylisi

Pre-Processing

- We have used necessary resources from NLTK, specifically the Punkt tokenizer models for the removal of English stopwords.
- We also utilized PyPDF2 to extract text from each page of the PDF and concatenate it into a single string.
- This preprocessing have all the standard steps involved:
    - Stop Word Removal
    - Removing Numbers and Special Characters
    - Lowercasing of the text
    - Tokenization

Methodology

TF-IDF

- Assigns more weight to less frequent terms in the text.
- Extracted Feature Names by creating TFIDF Matrix

Author	Book	Extracted Features
Ali Teoman	GizliKalmisIstanbulMasali	5467
	Karadelik Güncesi	31894
	InsansizKonak	9312
Abdulhak Sinasi Hisar	Fahim Bey ve Biz	11508
	Ali Nizami Beyin Alafrangaligi ve Seyhligi	7065
	Camlicadaki Enistemiz	17280
Ahmet Hamdi Tanpinar	Huzur	22983
	Mahur Beste.	13036
	Sahnenin Disindakiler	19283
Halid Ziya Usakligil	Ask-i Memnu	18242
	Kadin Pencesi.	7679
	Kirik Hayatlar.	19370
Refik Halid Karay	Anahtar	15887
	Bu Bizim Hayatimiz	17260
	Bugunun Saraylisi.	17492

Feature Vector Extraction with BERT and RoBERTa

- BERT is an encoder-only Large Language Model, this makes it successful in feature vector extraction.
- To extract feature vectors, the output of the last hidden state could be used.
- The paper proposing BERT mentions that feature vectors obtained by BERT could be used for unsupervised learning.
- We will use the Hugging Face library to utilize these methods.
- RoBERTa is a BERT variant, outperforming BERT in numerous tasks.

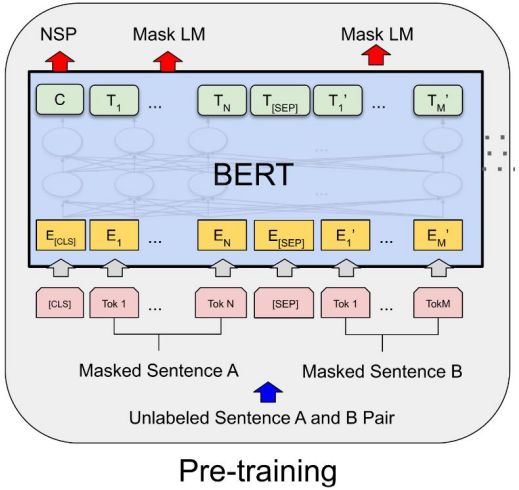
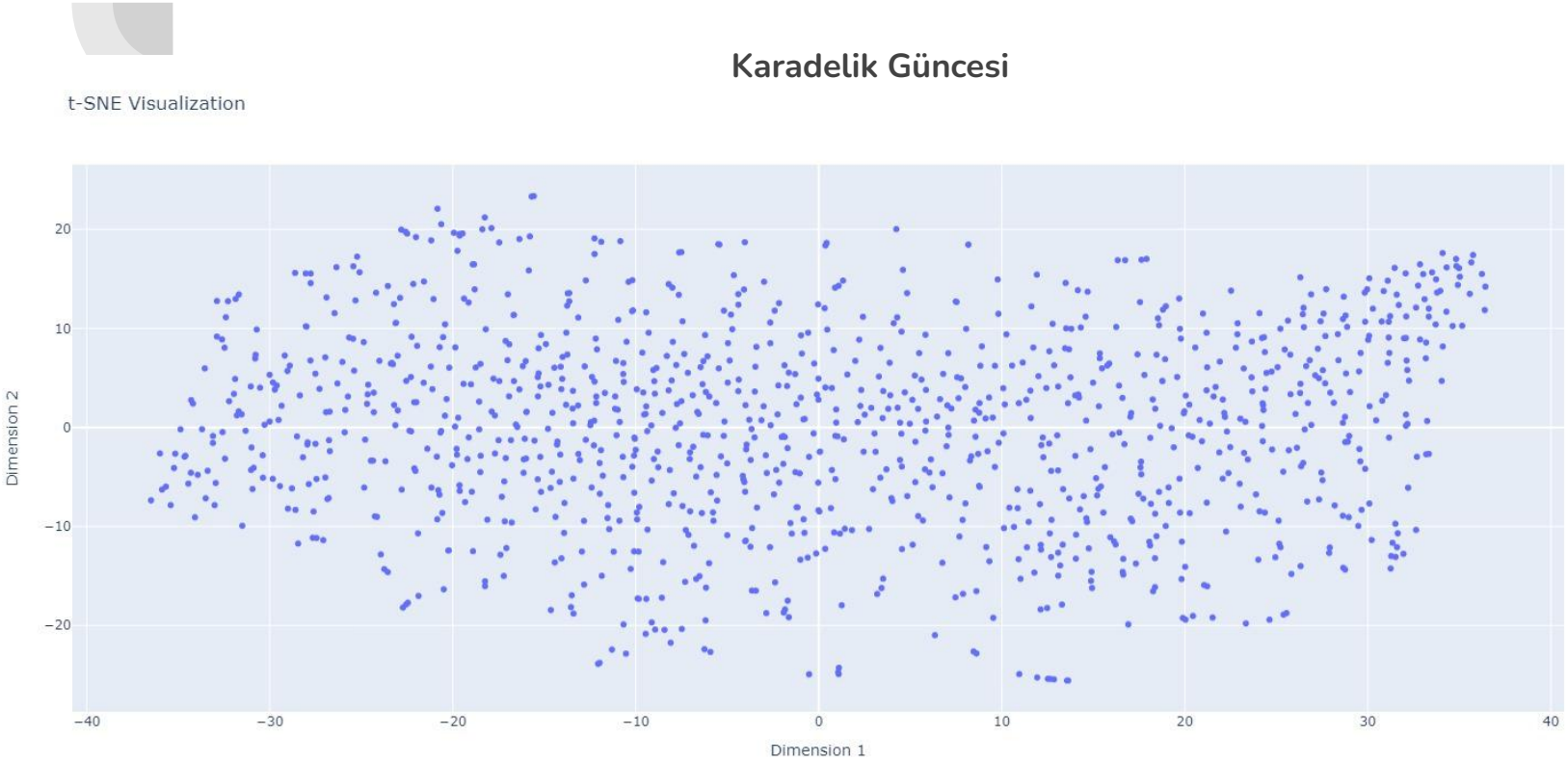


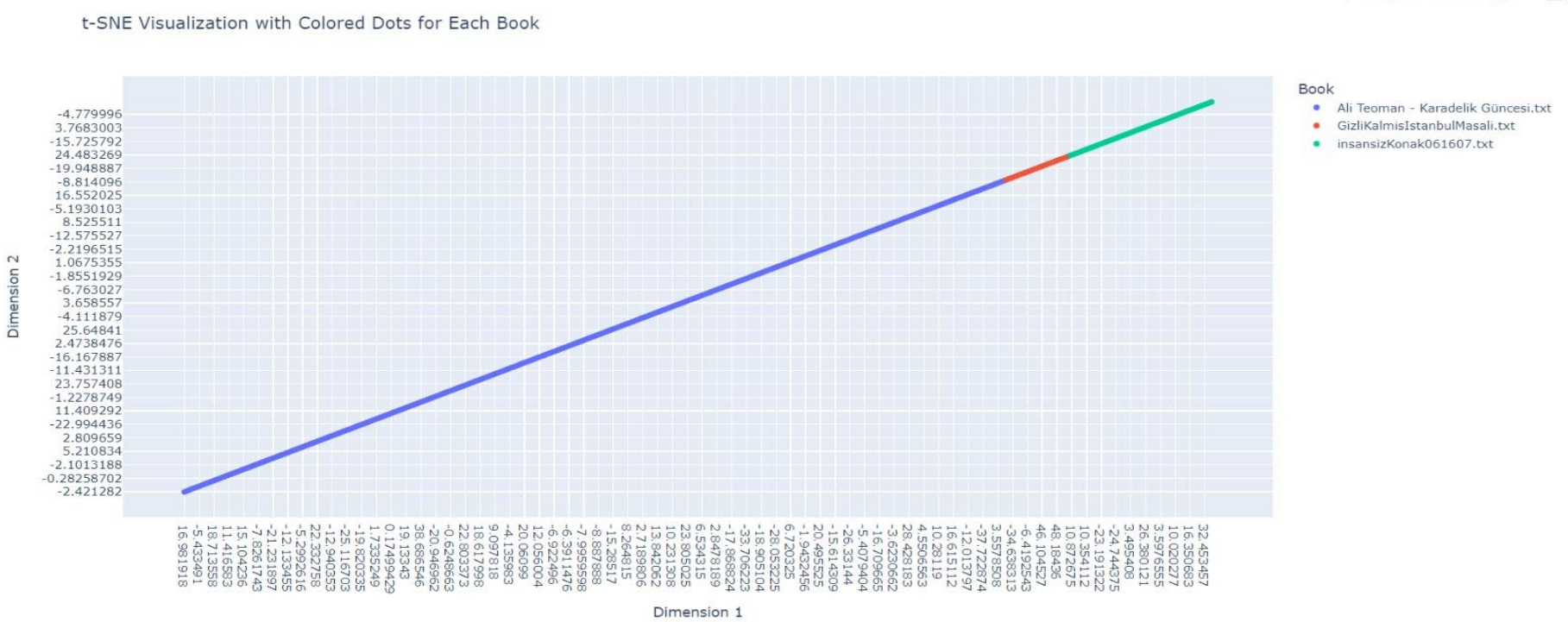
Figure: BERT Architecture[1]

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Visualization of Feature Vectors



Visualization of Feature Vectors (cont'd)



Clustering

- We cluster feature vectors obtained from the chunks of text.
- K-means algorithm will be used for clustering.
- Since each book has its own style, feature vectors of text chunks obtained from the same book should be in the same cluster.

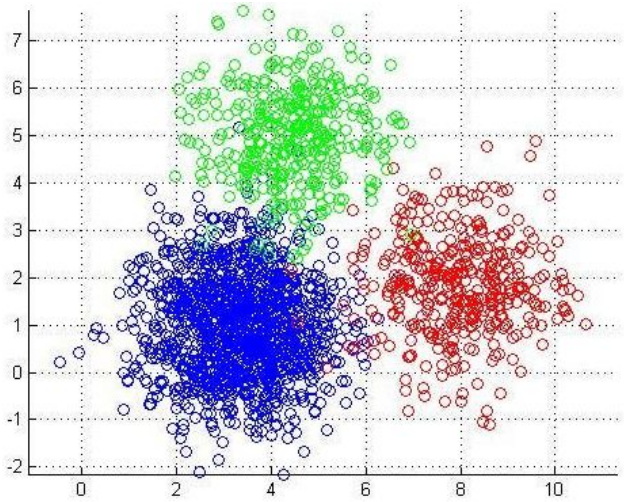
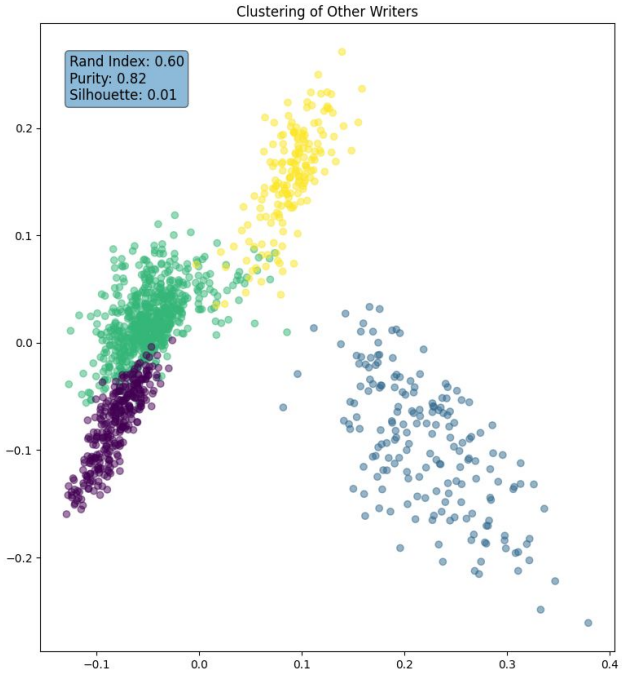
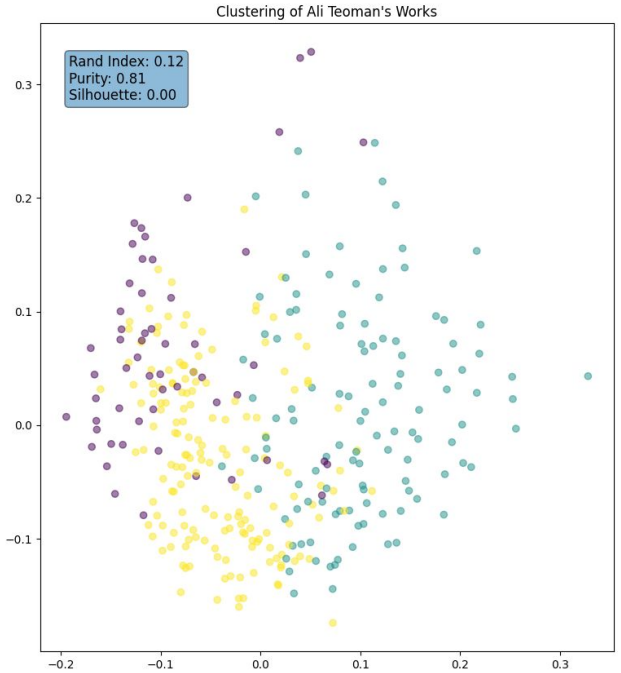


Figure: Example Clustering, taken from [1]

[1] “What Is Hierarchical Clustering?” *KDnuggets*, www.kdnuggets.com/2019/09/hierarchical-clustering.html. Accessed 22 Nov. 2023.

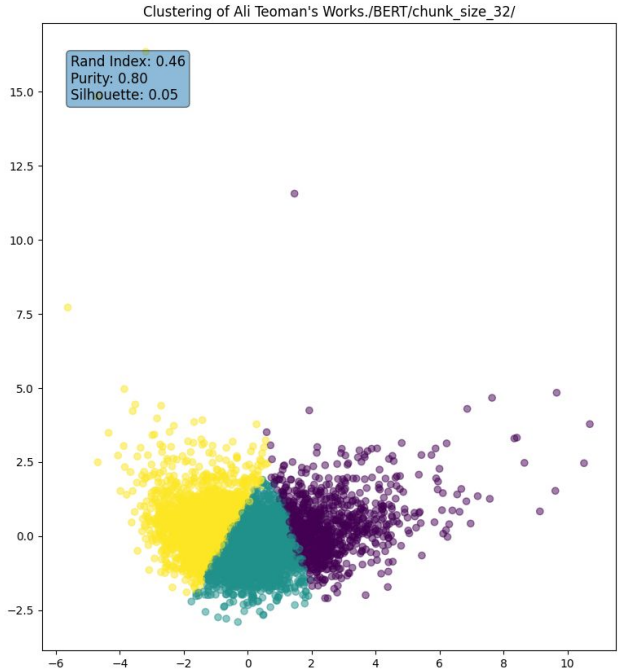
TF-IDF


Clustering of Other Writers
Rand Index: 0.60
Purity: 0.82
Silhouette: 0.01

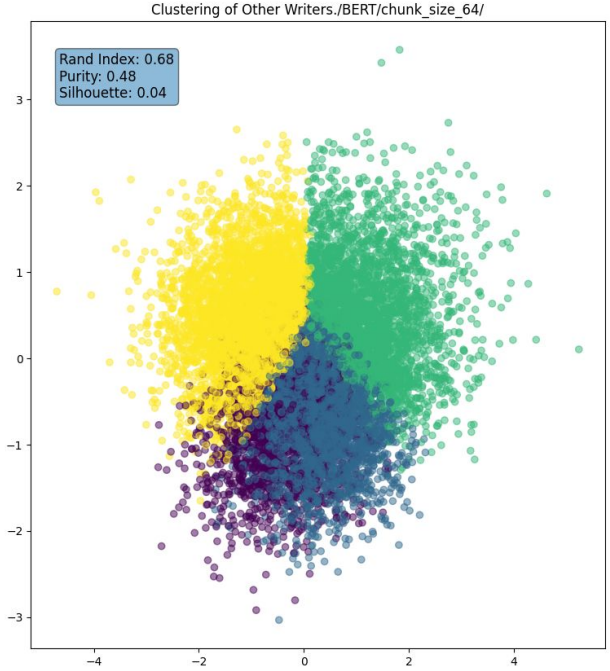

Clustering of Ali Teoman's Works
Rand Index: 0.12
Purity: 0.81
Silhouette: 0.00

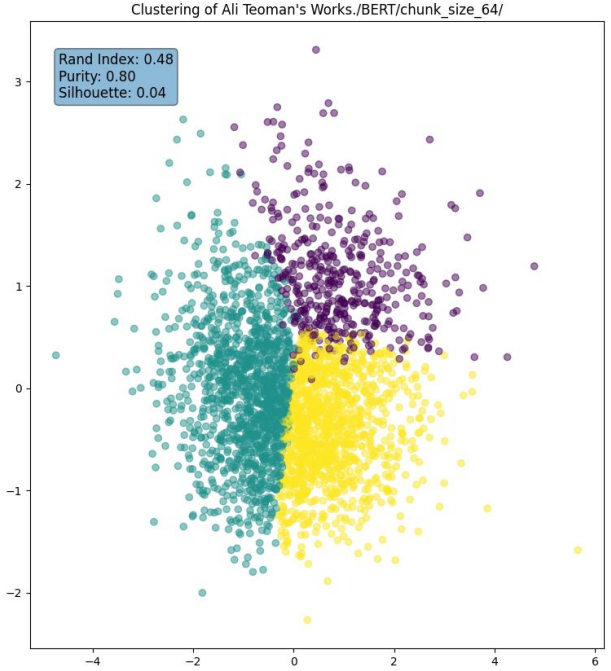
BERT


Clustering of Other Writers./BERT/chunk_size_32/
Rand Index: 0.65
Purity: 0.45
Silhouette: 0.04
6
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

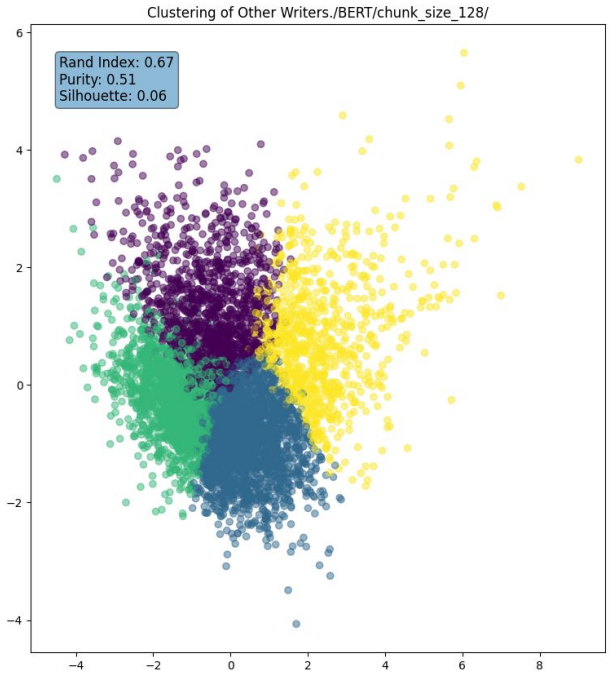

Clustering of Ali Teoman's Works./BERT/chunk_size_32/
Rand Index: 0.46
Purity: 0.80
Silhouette: 0.05
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−6
−4
−2
0
2
4
6
8
10

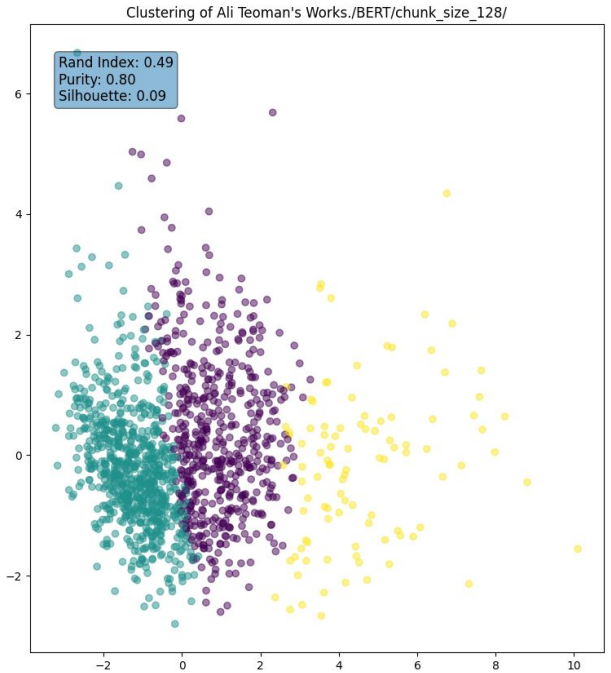
BERT


Clustering of Other Writers./BERT/chunk_size_64/
Rand Index: 0.68
Purity: 0.48
Silhouette: 0.04

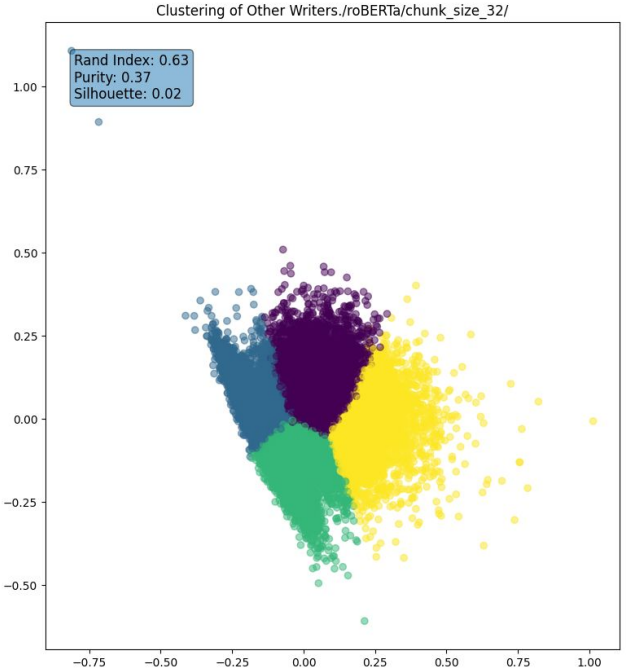

Clustering of Ali Teoman's Works./BERT/chunk_size_64/
Rand Index: 0.48
Purity: 0.80
Silhouette: 0.04

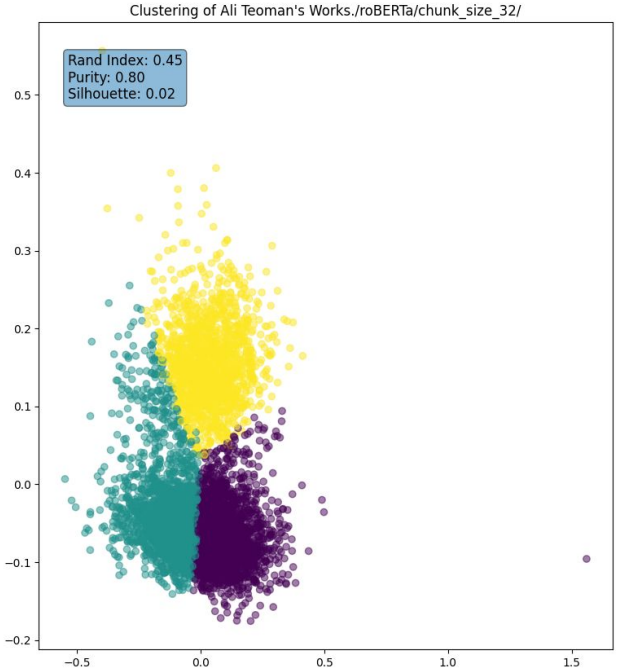
BERT


Clustering of Other Writers./BERT/chunk_size_128/
Rand Index: 0.67
Purity: 0.51
Silhouette: 0.06

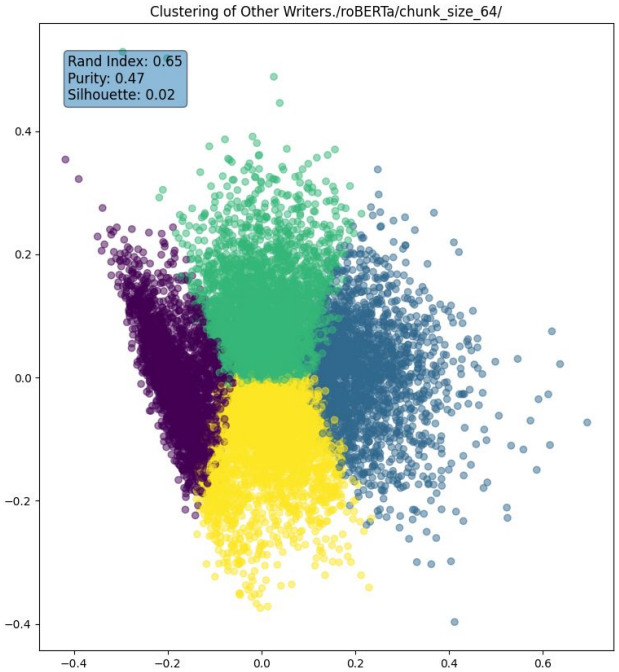

Clustering of Ali Teoman's Works./BERT/chunk_size_128/
Rand Index: 0.49
Purity: 0.80
Silhouette: 0.09

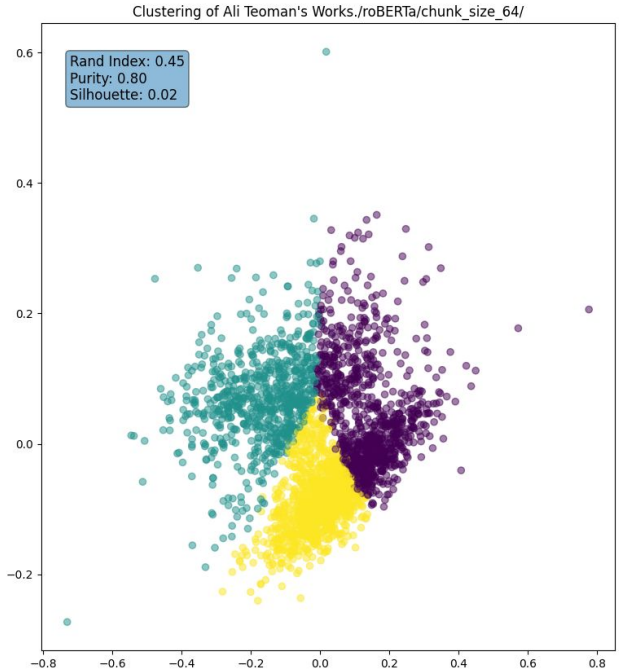
RoBERTa


Clustering of Other Writers./roBERTa/chunk_size_32/
Rand Index: 0.63
Purity: 0.37
Silhouette: 0.02
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

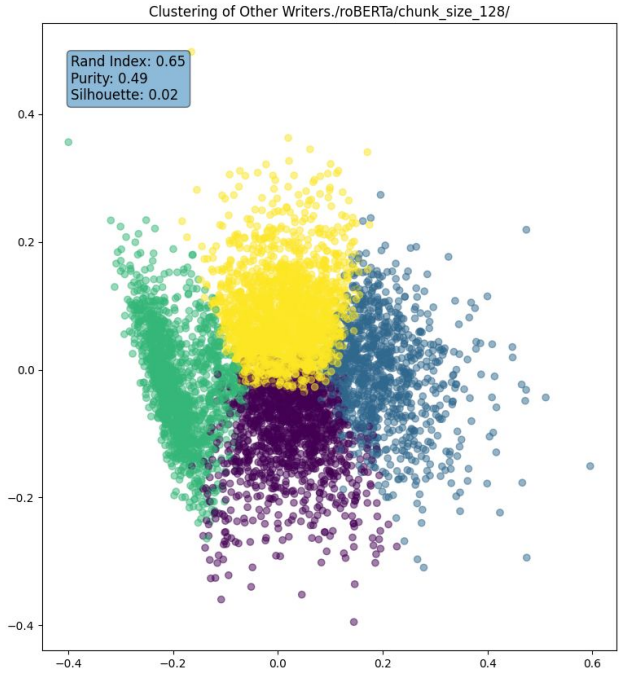

Clustering of Ali Teoman's Works./roBERTa/chunk_size_32/
Rand Index: 0.45
Purity: 0.80
Silhouette: 0.02
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.5
0.0
0.5
1.0
1.5

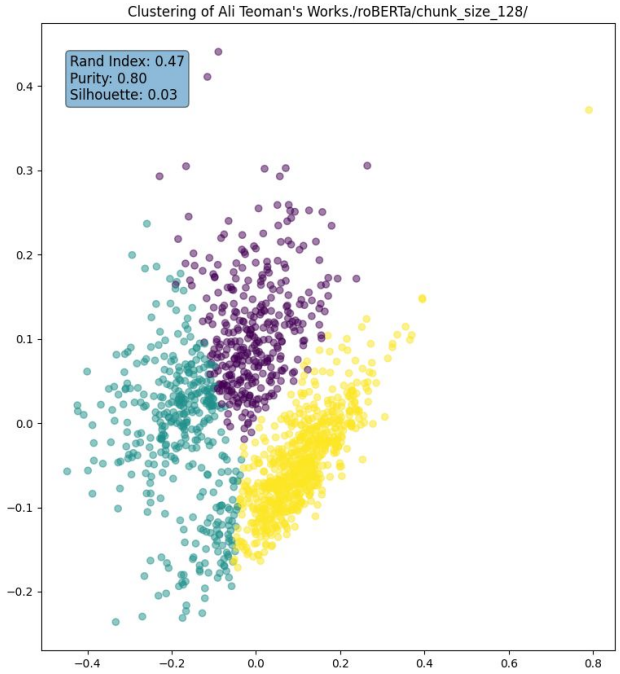
RoBERTa


Clustering of Other Writers./roBERTa/chunk_size_64/
Rand Index: 0.65
Purity: 0.47
Silhouette: 0.02
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6


Clustering of Ali Teoman's Works./roBERTa/chunk_size_64/
Rand Index: 0.45
Purity: 0.80
Silhouette: 0.02
0.6
0.4
0.2
0.0
−0.2
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

RoBERTa


Clustering of Other Writers./roBERTa/chunk_size_128/
Rand Index: 0.65
Purity: 0.49
Silhouette: 0.02
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6


Clustering of Ali Teoman's Works./roBERTa/chunk_size_128/
Rand Index: 0.47
Purity: 0.80
Silhouette: 0.03
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Evaluation Criteria

- Pure Clusters
- Rand Index
- Silhouette Score

Cluster Purity

Feature Vector Extraction Method / Clustered Data	Clustering of Ali Teoman's Work	Clustering of Other Authors
TF-IDF	0.81	0.82
BERT - chunk size 32	0.80	0.45
BERT - chunk size 64	0.80	0.48
BERT - chunk size 128	0.80	0.51
RoBERTa - chunk size 32	0.80	0.37
RoBERTa - chunk size 64	0.80	0.47
RoBERTa - chunk size 128	0.80	0.49

Rand Index

Feature Vector Extraction Method / Clustered Data	Clustering of Ali Teoman's Work	Clustering of Other Authors
TF-IDF	0.12	0.60
BERT - chunk size 32	0.46	0.65
BERT - chunk size 64	0.48	0.68
BERT - chunk size 128	0.49	0.67
RoBERTa - chunk size 32	0.45	0.63
RoBERTa - chunk size 64	0.45	0.65
RoBERTa - chunk size 128	0.47	0.65

Silhouette Score

Feature Vector Extraction Method / Clustered Data	Clustering of Ali Teoman's Work	Clustering of Other Authors
TF-IDF	0.00	0.01
BERT - chunk size 32	0.05	0.04
BERT - chunk size 64	0.04	0.04
BERT - chunk size 128	0.09	0.06
RoBERTa - chunk size 32	0.02	0.02
RoBERTa - chunk size 64	0.02	0.02
RoBERTa - chunk size 128	0.03	0.02

Summary

- **Aim**: Investigate and verify the claim of Ali Teoman
- **Methodology**: Pre-processing, TD-IDF, BERT, RoBERTa, Clustering
- **Evaluation**: Cluster Purity, Rand Index, Silhouette Score
- **Expectation**: All unique clusters
- **Conclusion**: Validate the claim

Based on our dataset, we can say that the claim of Ali Teoman that all of his works are unique, has finally been validated :)

Thank You!