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

AQSA SHABBIR, Bilkent University, Turkey
KOUSAR KOUSAR, Bilkent University, Turkey
BURAK FERIT AKTAN, Bilkent University, Turkey
YASIR ALI KHAN, Bilkent University, Turkey
NOOR MUHAMMAD, Bilkent University, Turkey

This research project aims to analyze the writing style of the renowned Turkish author Ali Teoman across several of his works. Challenging the very essence of his claim – the declaration that each of his books possesses a distinctive, unparalleled style.

The data set comprises digital copies of Ali Teoman's books. Our methodology involves extensive pre-processing steps such as stop-word removal, lower-casing, removing numbers, special characters, and tokenization.

To confirm and validate our algorithm, we are also using works from 4 different authors

For extracting feature vectors, we are using three main methods. TF-IDF, the Turkish BERT model (BERTurk), and the RoBERTa Model. Different parameters for each model are explored to understand their impact on the analysis. We then employ the k-means clustering algorithm to group these feature vectors.

The evaluation of our results involves calculating cluster purity, utilizing the Rand Index method and Silhouette Score. Cluster purity assesses the extent to which clusters are pure, supporting or refuting Ali Teoman's claim of unique writing styles.

By employing these comprehensive methods, this study contributes to the field of information retrieval and authorship analysis, offering insights into the consistency and distinctiveness of writing styles in Ali Teoman's literary works.

Our results indicate Purity of clusters for Ali Teoman's work using K-Means Clustering of 0.81. Rand index score of 0.12, 0.46 and 0.45 and Silhouette Score of 0.00, 0.05 and 0.02 are obtained for TFIDF, BERT and RoBERTa respectively.

Authors' addresses: Aqsa Shabbir, aqsa.shabbir@bilkent.edu.tr, Bilkent University, Ankara, Turkey; Kousar Kousar, kousar.kousar@bilkent.edu.tr, Bilkent University, Ankara, Turkey; Burak Ferit Aktan, ferit.aktan@bilkent.edu.tr, Bilkent University, Ankara, Turkey; Yasir Ali Khan, yasir517390@gmail.com, Bilkent University, Ankara, Turkey; Noor Muhammad, noor.muhammad@bilkent.edu.tr, Bilkent University, Ankara, Turkey.

Manuscript submitted to ACM

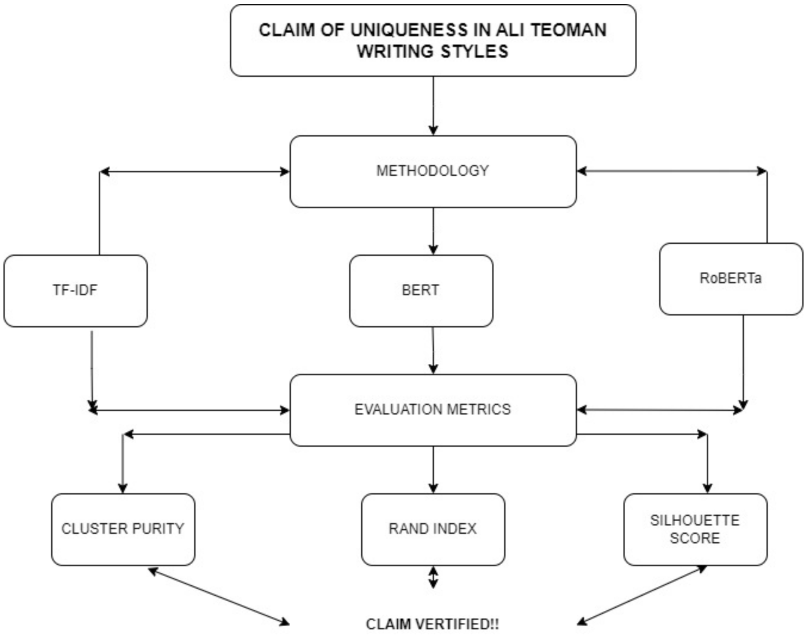

Fig. 1. Project Graphics

CCS Concepts: • **Information Retrieval**; • **Feature Vectors**; • **BERT**; • **TF-IDF**;

Additional Key Words and Phrases: word embeddings, unique style, BERTurk

ACM Reference Format:

Aqsa Shabbir, Kousar Kousar, Burak Ferit Aktan, Yasir Ali Khan, and Noor Muhammad. 2023. Analyzing Writing Style of Ali Teoman for Testing Uniqueness of Individual Works . 1, 1 (December 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This project observes the intricate task of textual analysis and utilization of advanced computational techniques to determine the authenticity of the Ali Teoman's claim. Ali Teoman is a famous Turkish author and poet. He makes a very interesting and unique claim that all of his works are stylistically different from each other. Anchored by the motivation to empirically validate this claim of his, our research purpose is to apply a combination of traditional and modern linguistic processing methods to validate said claim. We have used few literary works of some other authors as well in order to build a strong foundation for comparison.

We have started this task by preprocessing the text into a form amenable for computation analysis. This procedure includes the removal of stop words and non-alphabetic characters from the text and normalizing case, thus preparing our dataset for the application of computational analysis. Our research encompasses the implementation of Term-Frequency-Inverse Document Frequency (TF-IDF), a statistical measure that evaluates the importance of a word in the context of a corpus by emphasizing its importance based on the commonality across documents. This traditional information retrieval methodology is complemented by the deployment of BERT and Roberta models. BERT is a state-of-the-art language processing framework that can understand nuances of language through the use of contextually-rich embeddings, while Roberta, another advanced language model, enhances our understanding of text by employing masked language modeling and contextualized representations, thereby improving various natural language processing tasks. The final results of the implementation are promising and suggest the feasibility of our approach.

2 RELATED WORK

A lot of work has been done on authorship identification of online messages [9]. In those frameworks, four types of writing-style features (lexical, syntactic, structural, and content-specific features) are extracted and inductive learning algorithms are used to build feature-based classification models to identify authorship of online messages. Their experimental results showed that the proposed approach was able to identify authors of online messages with satisfactory accuracy of 70 to 95 %. Another work [8] has been found that uses the LibLINEAR SVM algorithm to classify 50 authors.The dataset used for this experiment is a subset of the popular and well-established RCV1 (Reuters Corpus Volume 1) dataset, used as a benchmark for research in information retrieval, called the Reuter-50-50 dataset. This gave an accuracy of 88.83 %. But most of this work has been done for English authors.

3 DATA SETS

In order to check the correctness of our algorithm along with Ali Teoman we have used literature of four different authors.

- Abdulhak Sinasi Hisar
- Ahmet Hamdi Tanpinar
- Halid Ziya Usakligil
- Refik Halid Karay

3.1 Text Extraction and Preprocessing:

- Extracts text from the PDF using PyPDF2.
- Preprocesses the text by:
  - Converting to lowercase.
  - Removing non-alphabetic characters.
  - Tokenizing using NLTK.
  - Eliminating stopwords.

3.2 Chunk Creation:

- Splits the preprocessed text into chunks for analysis.
- Each chunk contains a specified number of words (chunk size: 32,64 and 128 words).

4 METHODS

4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF stands for Term Frequency-Inverse Document Frequency, and it is a numerical statistic used in information retrieval and text analysis. It is commonly used in natural language processing and text mining to evaluate the importance of a word in a document relative to a collection of documents (corpus)

Term Frequency (TF): It measures how often a term (word) appears in a document.

Inverse Document Frequency (IDF): This measures how important a term is across the entire document collection. The product of TF and IDF yields a weight, emphasizing unique terms and denigrating the common ones. This process results in TF-IDF matrices representing each chunk, capturing the uniqueness of writing styles in Ali Teoman's works. On created chunks, TF-IDF is applied to extract unique feature vectors and we found below results:

Table 1. Feature Extraction of Literature of Ali Teoman

Ali Teoman		
Book Name	Total Words	Extracted Features
Gizli Kalmis Istanbul Masali	70564	5467
Karadelik Güncesi	907857	31894
Insansiz Konak	151025	9312

4.2 Feature Vector Extraction with BERT and RoBERTa

4.2.1 General Overview of BERT and RoBERTa. BERT is a short-hand for Bidirectional Encoder Representations from Transformers. BERT is a transformer-based large language model, trained on text data collected from Wikipedia in a self-supervised manner by Google [4]. As can be easily imagined, labeling large text corpora is not only very time-consuming but also requires a human source. Therefore, being a self-supervised method is a significant advantage for BERT. There are 2 main training objectives of BERT, these are masked language modeling and next-sentence prediction. While masked language modeling training objective improves BERT for understanding semantic relationships between words better, being trained with next-sentence prediction data allows BERT to understand relations between sentences. RoBERTa is a BERT variant where the main differences from BERT are the following: removal of next sentence prediction (i.e., the only pretraining objective of RoBERTa is masked language modeling) and dynamic masking. [7]

There are numerous use cases of BERT and RoBERTa, it can be used for almost any natural language processing task except text generation. One of the most important utilities of BERT is the following: it enables us to extract feature vectors from text.

Since the training dataset of BERT consists of articles in English, we used BERTurk. BERTurk has the same architecture as BERT, however, it is trained with text data in Turkish. [10] Similarly, we used RoBERTaTURK for being able to use the RoBERTa model for Turkish. [1]

4.2.2 Obtaining Features Vectors with BERT and RoBERTa. The paper in which BERT is proposed, claims that feature vectors obtained by BERT could be used for unsupervised machine learning tasks, in addition to supervised machine learning tasks [4]. There are numerous academic works utilizing feature vectors extracted using BERT. [11] [3] [6] [5] To download the pre-trained BERT model with its weights, the Transformers library [12] is used. Using the pipeline function of the Transformers library, feature vectors are extracted from the text chunks of books. Each feature vector has 768 dimensions. Note that the number of feature vectors obtained by this method for a chunk of text is equal to the number of tokens in it. However, our methodology requires obtaining only one feature vector per chunk. There are 4

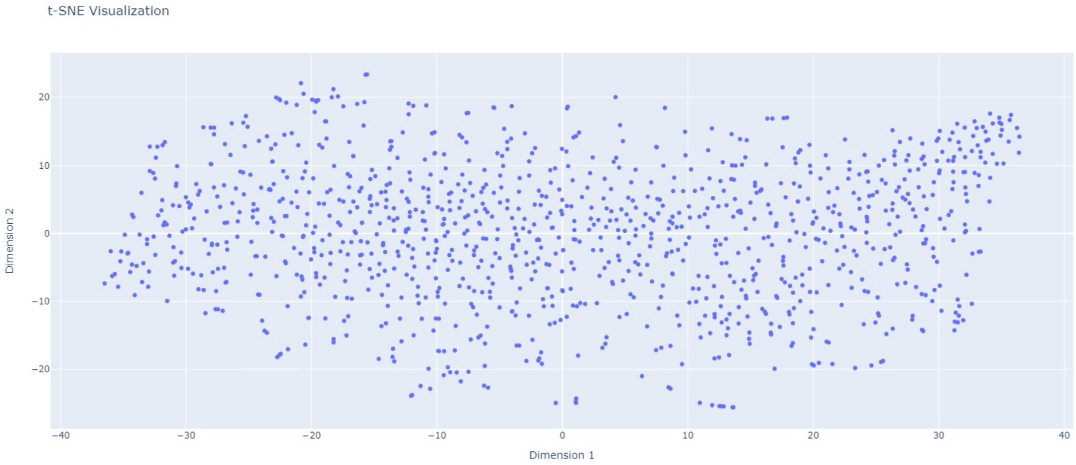

Fig. 2. t-SNE visualization for feature vectors of text chunks from Karadelik Güncesi

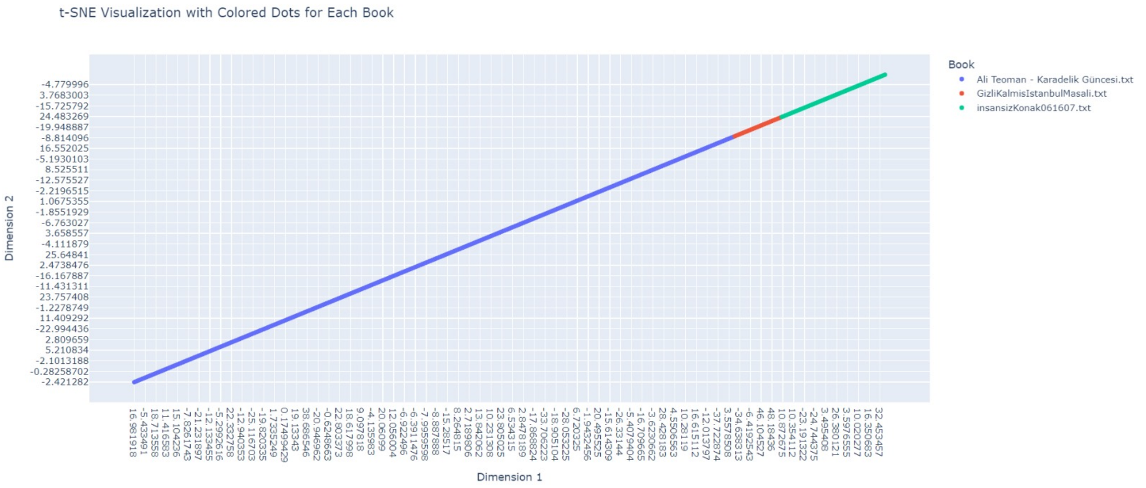

Fig. 3. t-SNE visualization of text chunks obtained by Ali Teoman's books in our dataset

common ways of extracting one feature vector to represent all the input text in large language models, one of them is special for BERT and similar architectures. These are:

- Averaging. This method can cause losing a small number of high activations if there are so many low activations. Therefore, this method may not be good enough.
- Getting maximum value among feature vectors for each dimension.
- Concatenating feature vectors. This method causes too long feature vectors and is computationally ineffective.

- Using feature vector of [CLS] token. [CLS] is a special token used in BERT and similar architectures but not all large language models. It is a special token that represents the whole input text. It is the most common method to obtain a single feature vector from a text.

We used the last and the most common method for BERT. One of the reasons for the existence of [CLS] tokens is to enable users to extract a single feature vector for a text. For RoBERTa, we took the maximum value among feature vectors for each dimension, since there is no [CLS] token for RoBERTa. In future work, we will use these feature vectors for clustering purposes.

Since these feature vectors obtained by BERT and RoBERTa are 768 dimensional, it is not easy to visualize them. To reduce the number of dimensions of the feature vectors, t-SNE is used. t-SNE is a dimensionality reduction technique, ensuring the most similar points will be close to each other and most different points will be far from each other after the dimensionality reduction [2]. In this study, this method is used for visualization purposes.

As can be seen from Figure 2, the feature vectors of text chunks taken from Ali Teoman's books are very discriminative. They are even separable by the human eye.

5 APPLICATION OF K-MEANS CLUSTERING

In our study, we choose to apply K-means algorithm as the clustering methodology. This unsupervised learning method partitions the data into provided number of clusters. For our case, we used feature vector data from TF-IDF, BERT and RoBERTa. Three different chunk size configurations (32, 64 and 128 words) were used for BERT and RoBERTa. Furthermore, clustering was done in two parts. First, K-means was applied to cluster the works of Abdulhak Sinasi Hisar, Ahmet Hamdi Tanpinar, Halid Ziya Usakligil and Refik Halid Karay. Three pieces of writing from each author were distinguished into 4 separate clusters. Second, three literary works of Ali Teoman were clustered individually. Three different cluster evaluation methods were used to judge the clusters and to verify Ali Teoman's claim.

6 METHODS EVALUATION

To evaluate the feature vectors, the following three evaluation metrics were used.

6.1 Cluster Purity

Cluster purity is a measure used in information retrieval and clustering evaluation to assess the quality of clustering results. It is particularly relevant in the context of unsupervised learning, where the algorithm is not provided with labeled data, and the goal is to group similar items together into clusters.

Cluster purity is calculated by comparing the clusters produced by an algorithm to the ground truth or true class labels, if available. The formula for cluster purity is as follows:

$$Purity = \frac{1}{N} \sum_{i=1}^{k} \max_{j} \left| C_i \cap T_j \right| \tag{1}$$

Limitation: The cluster purity has some limitations. For instance, it assumes that each cluster should correspond to a single class, which may not always be the case in real-world data. Additionally, it does not take into account the actual structure of the data, and two different clusterings with the same purity value may have different characteristics. Therefore, it is often used in conjunction with other clustering evaluation metrics to provide a more comprehensive assessment of clustering quality. That is why in our case, we are multiple evaluation metrics.

Table 2. Cluster Purities

Feature Vector Extraction Method	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

6.2 Rand Index

The Rand Index (RI) is a measure used in information retrieval and clustering evaluation to assess the similarity between two clustering. It measures the percentage of agreements between the pairs of data points in the given clustering and the true clustering or ground truth.

The Rand Index is defined in terms of four quantities:

True Positives (TP): The number of pairs of data points that are in the same cluster in both the predicted clustering and the true clustering.

True Negatives (TN): The number of pairs of data points that are in different clusters in both the predicted clustering and the true clustering.

False Positives (FP): The number of pairs of data points that are in the same cluster in the predicted clustering but in different clusters in the true clustering.

False Negatives (FN): The number of pairs of data points that are in different clusters in the predicted clustering but in the same cluster in the true clustering.

The Rand Index (RI) is then calculated as:

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Limitation: Rand Index can be sensitive to the number of clusters and the distribution of elements among them. Adjusted Rand Index (ARI) is a variation of the Rand Index that adjusts for chance, providing a normalized measure that accounts for the expected similarity between random clustering.

Table 3. Rand Index Scores

Feature Vector Extraction Method	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

6.3 Silhouette Score

The Silhouette Score is a metric used to calculate the goodness of a clustering technique, based on how well-separated the clusters are. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The Silhouette Score ranges from -1 to 1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. A Silhouette Score close to 1 indicates a good clustering, while a score close to -1 suggests that the samples may have been assigned to the wrong cluster. A score around 0 indicates overlapping clusters.

The Silhouette Score for a single sample is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

where:

$s(i)$ is the Silhouette Score for the ith sample,

$a(i)$ is the average distance from the ith sample to other samples in the same cluster,

$b(i)$ is the smallest average distance from the ith sample to samples in a different cluster, minimized over clusters.

The overall Silhouette Score for the entire dataset is calculated as:

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{4}$$

where S is the overall Silhouette Score for the dataset and N is the number of samples.

Limitation: Silhouette Score is just one of many metrics available for evaluating clustering quality, and its interpretation should be considered alongside other relevant metrics based on the specific characteristics of your data.

Table 4. Silhouette Scores

Feature Vector Extraction Method	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

7 DISCUSSIONS

The research project aimed to empirically validate the distinctive writing styles claimed by the renowned Turkish author, Ali Teoman, across his literary works. Leveraging a combination of traditional and advanced computational linguistic methods, our study focuses on a comprehensive analysis of textual data from Ali Teoman's writings and works from various other authors.

Three distinct methods—TF-IDF, BERTurk (Turkish BERT model), and RoBERTa Model—were employed to extract feature vectors, exploring diverse parameters to understand their impact on the analysis.

The premises of clustering method in this case was that in order for Ali Teoman's claim of unique individual writings to be true, the extent up to which the literary works of 4 different authors can be separated from each other should be close or less than the distinctiveness among the works of Ali Teoman himself.

Utilizing the K-Means clustering algorithm, feature vectors were grouped, allowing for an evaluation of cluster purity, Rand Index, and Silhouette Score.

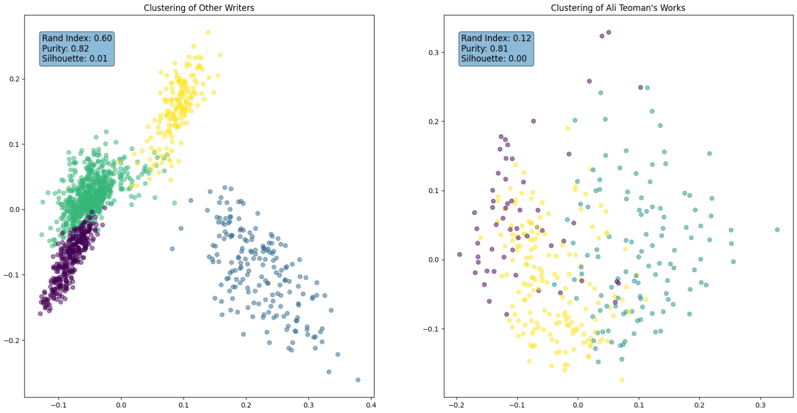

Fig. 4. K-Means cluster visualization for TF-IDF

Fig. 3 shows that clusters from the feature vectors obtained through TF-IDF. The image and the evaluation scores depict poor distinctivness among the works of Ali Teoman.

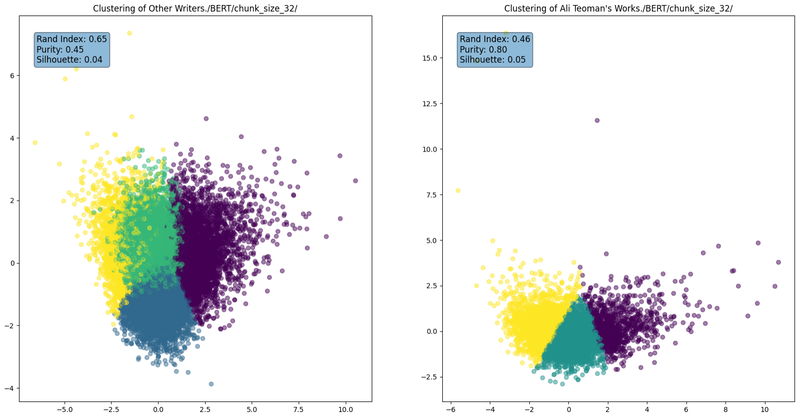

Fig. 5. K-Means cluster visualization for BERT 32 word chunk.

Fig. 4 shows that clusters from the feature vectors obtained through BERT. The results depict that when a contextual based feature vector extraction is done using a smarter model such as BERT, Ali Teoman's works are as much as if not even more seperable from each other as compared to all the other authors.

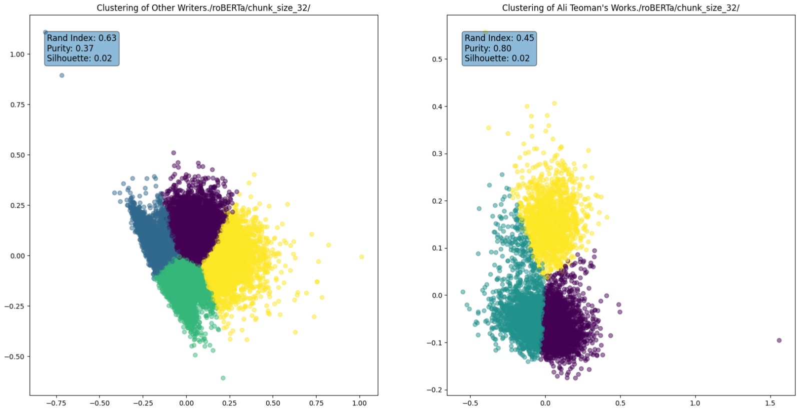

Fig. 6. K-Means cluster visualization for RoBERTa 32 word chunk

Fig. 5 shows a similar output through roBERTa to that of BERT where clusters are visually highly distinct and the scores are very close to each other for different authors and Ali Teoman's work.

8 CONCLUSION

The obtained cluster purity for Ali Teoman's works using K-Means Clustering reached 0.81, indicating a substantial level of purity in the identified clusters. The Rand Index scores of 0.12, 0.46, and 0.45, alongside Silhouette Scores of 0.00, 0.05, and 0.02 for TF-IDF, BERT, and RoBERTa respectively, provided insight into the similarity and dissimilarity within clusters, contributing to the validation of distinct writing styles. Implementation of Term-Frequency-Inverse Document Frequency (TF-IDF) and advanced language models—BERT and RoBERTa—proved insights into the consistency and uniqueness of Ali Teoman's writing styles. The findings not only support the claim of distinctive writing styles across his works but also demonstrate the efficacy of computational linguistic models in textual analysis and authorship attribution tasks.

REFERENCES

[1] Burak Aytan and C Okan Sakar. 2022. Comparison of Transformer-Based Models Trained in Turkish and Different Languages on Turkish Natural Language Processing Problems. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.

[2] T Tony Cai and Rong Ma. 2022. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *The Journal of Machine Learning Research* 23, 1 (2022), 13581–13634.

[3] Xinying Chen, Peimin Cong, and Shuo Lv. 2022. A long-text classification method of Chinese news based on BERT and CNN. *IEEE Access* 10 (2022), 34046–34057.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935* (2021).

[6] Aparna Sunil Kale, Vinay Pandya, Fabio Di Troia, and Mark Stamp. 2023. Malware classification with word2vec, hmm2vec, bert, and elmo. *Journal of Computer Virology and Hacking Techniques* 19, 1 (2023), 1–16.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[8] Carolyn Penstein Rose Rahul Radhakrishnan Iyer. 2005. A Machine Learning Framework for Authorship Identification From Texts. *Journal of the American Society for Information Science and Technology* 30 (2005), 34046–34057. https://doi.org/pdf/1912.10204

[9] Hsinchun Chen Zan Huang Rong Zheng, Jiexun Li. 2005. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 57 (2005), 34046–34057.

[10] Stefan Schweter. 2020. *BERTurk - BERT models for Turkish.* https://doi.org/10.5281/zenodo.3770924

[11] Hirotaka Tanaka, Hiroyuki Shinnou, Rui Cao, Jing Bai, and Wen Ma. 2020. Document classification by word embeddings of bert. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*. Springer, 145–154.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).