

Painter Profile Clustering Using NLP Features

N. Yagmur Ilba^a, U. Mahir Yıldırım^b and Doruk Sen^c

Department of Industrial Engineering, Istanbul Bilgi University, Eyupsultan, Istanbul, Turkey

Keywords: Natural Language Processing, Clustering, Text Analysis, XAI.

Abstract: This study introduces a practice for clustering painter profiles using features obtained from natural language processing (NLP) techniques. The investigation of similarities among painters plays an essential function in art history. While most existing research generally focuses on the visual comparison of the artists' work, more studies should examine the textual content available for artists. As the volume of online textual information grows, the frequency of discussions about artists and their creations has gained importance, underscoring the connection between social visibility through digital discourse and an artist's recognition. This research provides a method for investigating Wikipedia profiles of painters using NLP attributes. Among unsupervised machine learning algorithms, the K-means is adopted to group the painters using the driven attributes from the content details of their profile pages. The clustering results are evaluated through a benchmark painter list and a qualitative review. The model findings reveal that the suggested approach effectively clusters the presented benchmark painter profiles, highlighting the potential of textual data analysis on painter profile similarities.

1 INTRODUCTION


1.1 Understanding Painters' Similarities


Painters of visual arts have played an essential role in the development of society, both economically and ideologically. In the history of art, it has always been necessary to understand the similarities of artists and to be able to examine and group them under specific characteristics. While there are studies in the literature that segment artists' works, and therefore artists, by finding similarities through visual analysis of their works, analysis studies based on the media visibility and lives of artists are limited. Within the former line of research, a methodology to link the similarities of paintings using a pre-trained Convolutional Neural Network (CNN) model is developed; similarly, computer vision techniques are employed to determine visual similarities among paintings and applied graph-based analysis to understand the centrality of the top artists (Seguin et al., 2016; Castellano et al., 2021).


Certain painters have distinguished themselves; some pioneered new movements, while others left

a permanent mark with their distinctive works. Today, with the growth of information on the web, the frequency with which artists and their works are discussed and the manner of such discussions have gained significance, becoming one of the factors shaping an artist's fame. An artist's fame acts as a driving force in the market, drawing in audiences and collaborators, thus marking success. Creativity is often linked with fame; however, there is a notable lack of research on the other factors influencing an artist's renown in creative markets. Determining why a work of art or an artist becomes more recognised can be challenging. Objective characteristics frequently take centre stage. These characteristics may be determined by the artist's skills or the materials used. Painting's unique appeal, viewed as infused with the artist's essence, justifies its ongoing popularity and ability to fetch high prices at auctions, keeping its top spot in the art scene. (Mitali and Ingram, 2018; Graw, 2016)

Furthermore, the popularity of a work of art can result from a combination of historical context, human psychology, and factors that may be simply coincidental. For instance, the Mona Lisa, one of history's most famous paintings, made headlines due to vandalism and gained widespread attention after being stolen from a museum. During its absence, visitors headed to the gallery to look at the empty spot

^a  <https://orcid.org/0009-0001-6901-7105>

^b  <https://orcid.org/0000-0003-3469-8112>

^c  <https://orcid.org/0000-0003-3353-5952>

where it once hung. As a result, when it comes to reputation in society, the fame of artworks is frequently determined not only by their inherent quality but also by the discussions surrounding them (Leslie, 2014). While the popularity of artists in the music industry has been measured (Schedl et al., 2010; Bellogin et al., 2013; Krasanakis et al., 2018), to the best of our knowledge, the reasons behind the popularity of artists in the visual arts have yet to be assessed using big data analysis techniques.

1.2 Clustering Algorithms in Machine Learning

Machine learning is an essential element of the developing domain of data science. It comprises algorithms that learn by establishing relationships from data designated as model input, aiming to minimise the margin of error to enhance human decision-making. These algorithms are trained using statistical and algebraic methods. Machine learning algorithms are categorised into three main types: supervised, unsupervised, and semi-supervised. Selecting a suitable method is contingent upon the availability of labelled data. Supervised learning methods are employed for predictive analysis of tagged data, unsupervised learning methods are ideal for unlabelled data, and semi-supervised methods are utilized for partially labelled datasets. Within the scope of this study, the K-means algorithm was used to implement the clustering method, an unsupervised machine learning technique. (Maleki et al., 2020)

The clustering method is a powerful tool that groups similar data points based on specified metrics. There are four main types of clustering algorithms: hierarchical, density-based, partitioning, and grid-based, each with unique advantages and use cases (Chaudhry et al., 2023). The K-means algorithm is widely adopted among these algorithms due to its simplicity, low computational complexity, and versatility. Recent research in this domain has focused on improving the algorithm's overall performance, and a comprehensive review of these methods has been conducted (Ikotun et al., 2022).

1.3 NLP Applications

While being a subfield of artificial intelligence, NLP assigns computers to understand, interpret, and manipulate natural language. This technology has progressed in several areas, such as automatic text summarization, translation, semantic analysis, emotion analysis, and speech recognition. The challenges of NLP broadly fall under two main approaches. The

first is the linguistic approach, where words or groups of words serve as features and texts are analysed in distinct categories. Secondly, there are algorithms based on machine learning and deep learning developed to create representations that better understand the holistic meaning of the text (Ramakrishnan et al., 2020).

Analysing explanatory texts about artists using NLP methods and understanding their impact on their recognition by discovering distinctive information about the artist is crucial for evaluating their works. However, studies that detail the various language processing techniques mainly used in analysing text data and how these techniques can be applied in arts and cultural management are far from ordinary. In the field of natural language processing applied to arts and culture, a notable study was presented by Cieliebak in 2019, focusing on topic modelling, trend detection, and determining demographic characteristics. The study also illustrates how these techniques can be utilised to analyse online feedback about a music event or a new exhibition. (Cieliebak et al., 2019)

Since social media has become an effective platform for discovering artists, artists are clustered based on their presence on social media and vectorised textual data from their biographies, as highlighted by Powell et al. in their research. This approach underscores the significance of an artist's digital footprint for understanding their influence and reach. Yet, it still leaves a gap in comprehending the factors contributing to an artist's recognition and success. (Powell et al., 2020)

Furthermore, an automated system to review NLP literature was developed in 2023 and commonly used approaches were identified (Sawicki et al., 2023). While Wikipedia has been one of the most frequently employed datasets for NLP applications, Spacy has been selected amongst other Named Entity Recognition (NER) models.

The study continues with methodology with the introduction of data collection in section 2.1, preprocessing in section 2.2, feature extraction, and clustering algorithm in sections 2.3 and 2.4 respectively. The results and findings section provides extracted features in 3.1 and clustering outputs in 3.2, followed by the conclusion.

2 METHODOLOGY

2.1 Data Collection

Several websites associated with art, like Wikipedia and WikiArt, are valuable sources of extensive in-

formation. WikiArt is a resource that provides a comprehensive knowledge database about visual art. The artists' Wikipedia links on their artist pages were recorded in February and March 2023, along with details about the filtered painters on WikiArt. The BeautifulSoup package of Python was utilised to gather this data. Afterwards, using the urllib and openpyxl packages, an approach has been created to use the retrieved links to add all of the text from the artists' Wikipedia pages to the corresponding Excel sheets. This method provided information from the Wikipedia entries of 2,827 different artists, totalling 3,776,496 words.

Regarding popularity data, the first study that converts web data to popularity belongs to Knebel (2007). In his study, popularity is defined as the amount of media attention attracted by an artist, measured by the count of total hits for the artist at Google search in English and German, which only differ by the word "artist". The hits include exhibitions, artist biographies and information about the latest sales as well (Knebel, 2007). Further, Tekindor and McCracken (2012) also used Google hits as an independent variable, among others, to investigate the relationship between the value of painting and the artist's fame. As Google no longer provides the exact number of hits but rather the normalized popularity score, we used the aforementioned Wikipedia pages as an alternative approach to measuring an artist's popularity (Tekindor and McCracken, 2012).

Although the artists in the Wikiart archive are well-known, the site has different listings made by users. Artists who stand out in media searches are considered for the list, taking into account the media recognition of artists, and the publicly available artists list having the top 40 was investigated (Cole, 2022). The artists on this list correspond to those on Wikipedia, except Ai Weiwei, who is omitted because he is categorised as a performance artist.

2.2 Data Preprocessing

Foundational preprocessing techniques are primarily used for textual analysis when sourced from platforms like Wikipedia. The obtained data must be curated and refined to assure accuracy and reliability in subsequent analyses. The journey of data preprocessing encompassed several critical steps, ensuring the clean and structured textual data for continuing the further NLP tasks. Implementing these steps aims to eliminate noise and refine the content to its essence (Bird et al., 2009).

At first, a set of fundamental Python libraries was utilised, each conforming to a critical purpose. For instance, the Natural Language Toolkit (nltk) was sig-

nificant for diverse tasks related to linguistic processing. Other libraries, such as gensim and scikit-learn, were also used for modelling duties. Upon loading the dataset, a phase for data exploration ensures familiarity with its structure. This phase is also beneficial for identifying potential anomalies that deserve attention (McKinney, 2012).

The textual data is segmented into discrete linguistic units through tokenisation, a process grounded in linguistic theories. This way, the text is more easily digested by algorithms, enhancing the accuracy of the subsequent analysis (Meurers, 2012). As a step further in preprocessing, stopwords, often defined as the widespread words that typically add little or no semantic value, were filtered out from the dataset. This streamlining allows for a clearer understanding of the content presented in the dataset (Bird et al., 2009). Lemmatisation was then utilised to filter words to their lexical essence. In light of linguistic morphology, this technique secures uniformity in presenting words to increase analytical accuracy and minimise redundancy (Navigli and Ponzetto, 2012). These procedures form the steps of a comprehensive text-cleaning approach. Each artist profile is transformed using these procedures to create a groundwork for further analysis.

2.3 Feature Extraction

Feature extraction from the Wikipedia profiles of the painters enables the transformation of qualitative textual descriptions into quantitative measures. NLP techniques were applied to provide meaningful insights into the artists' profiles for conducting further analyses. As a first step, the Sentiment Intensity Analyzer is used to measure the general tone of the artist profiles. The technique offers a compound sentiment score while observing positive and negative nuances within the text in artist profiles, also providing neutral scores for each input. When looking for a single sentiment measure, the normalised compound score is a highly helpful statistic as it adds up all lexical assessments while taking scores between -1 and 1. The primary motivation for using this technique is to shed light on the general perception of the artist's profile.

Furthermore, the text length feature is generated from the preprocessed artist description. This metric can provide fundamental insights into the comprehensiveness of the artist's profile, pointing at the prominence or depth of the text in the art world (Liu, 2012).

Additionally, NER, one of the most fundamental techniques in NLP, is employed to find specific organisations, individuals, and events within the preprocessed dataset of artist profiles. Finding the number

of observations is elaborated to aid in demonstrating the association, affiliation, and potential influences of an artist. Identifying additional individuals in the artist description text can emphasise noteworthy partnerships or influential associations that shaped artists' careers. Moreover, the number of art-specific events and awards in the Wikipedia descriptions can be remarkably insightful. Frequent mentions of these honours and events in the text can indicate information on the artist's recognition, reputation, and participatory activeness in the art community and high-profile art-specific events (Nadeau and Sekine, 2007).

Another metric is introduced by measuring the linguistic diversity within each artist's profile. One of the techniques to discuss the range and deepness of topics is the breakdown of the ratio of unique words that an artist has within the preprocessed corpus. High lexical diversity might underline varying subjects in the artist's profile, while less diverse and repetitive narratives are expected to produce a lower score (McCarthy and Jarvis, 2010).

Along with the several introduced contexts, semantic and syntactic attributes of the dataset are sought when implementing the Word2Vec model. The algorithm was introduced in 2013. It is known for generating vector representations of words to seize their semantic and syntactic characteristics by creating dense vectors in a continuous vector space (Mikolov et al., 2013). The tokenised sentences of artist profiles are used as a training corpus. The corpus is trained using a window size of five to define the contextual boundary for word prediction, a minimum word frequency threshold set at one, and the four workers are utilised to efficiently seize the power of parallel processing of the dataset.

The first step of similarity computation is introducing Word2Vec to vectorise each sentence in the artist descriptions and reference sentences. Then, cosine similarity is computed for each vectorised pair to capture the resemblance between these vectors while considering the distance between the pairs. The process is completed by calculating the average cosine similarity with each set of reference sentences. It is conducted to obtain a quantifiable criterion of the closeness of reference phrases with the profile of artists. The generated feature using this process is the difference between each artist profile's high and low average similarity scores. This approach indicates the closeness of these profiles to two poles of the phrase groups. The features extracted from the artist profiles form a solid foundation for further analyses.

2.4 Implementing K-means

The extracted features introduced in the previous section are used to obtain insights for the clustering algorithm. Before the implementation of the algorithm, the standardisation procedure is included. It is often essential for several machine learning algorithms, as they may function erroneously if the features do not approximately correspond to data with standard normal distribution. The process began with the standard scaling of scikit-learn (Pedregosa et al., 2011) to eliminate likely biases towards attributes with larger scales and have consistency in model performance. The analysis continues with determining the number of clusters for the K-means algorithm using the Within-Cluster Sum of Squares (WCSS). This technique aims to specify the number of clusters by identifying the cluster size for the part where the reduction in WCSS begins to decline. A range between one to ten clusters was investigated, and four distinct clusters were found using the elbow method from the visually represented clusters-WCSS chart using scikit-learn (Pedregosa et al., 2011). The obtained clusters were appended to the dataset as a new column for further analysis.

3 RESULTS AND FINDINGS

3.1 Extracted Features

After the vectorisation process, sentiment score, text length, lexical diversity, number of people, organisation, and award-related event entities are investigated, along with their popularity-based differences. The comparative analysis between the descriptive statistics of all artists and the top 40, as represented in Tables 1 and 2, highlights significant differences that outline the unique profiles of the top 40 artists within the broader artistic community.

A notable segregation is observed in the sentiment scores after obtaining compound scores, where discrimination is more precise for a better differentiation among artist entries. Consequently, texts that achieved a maximum compound score of 0.99, despite having a high neutral score, were assigned a higher sentiment score due to their more positive positioning within the lexicon. In this context, the top 40 artists display a higher mean (0.9) than the large artist group (0.66). The resulting increased sentiment score and a reduced standard deviation in the smaller group reveal an increased consistency towards a positive sentiment in the top 40 artist entries. The resulting increased sentiment score and a reduced standard devi-

ation in the smaller group reveal an increased consistency towards a positive sentiment in the top 40 artist entries. The text length additionally emphasises this distinction, with the smaller group of artists yielding an extended coverage (mean of 20806.6 words) than the larger group (6066.8 words). This suggests a more comprehensive discussion or engagement encircling the top 40 artists, as evidenced by the smaller standard deviation, indicating a behaviour towards uniformity in content length.

The analysis advances with the integration of various entities. Compared to the group of all artists, the top 40 indicates a substantial increase in mentions of personal (mean of 181.59) and organisational (mean of 106.38) entities. This difference can also be observed in award-based entities, as there is a notable growth in the small group (mean of 5.08) versus the broader community (1.96). On the other hand, the lexical diversity analysis conveys an opposing trend, as the top 40 group represents a smaller range than the group of all artists (0.65). Among the obtained features, the final investigation is provided with the investigation of popularity-based phrases. Obtained scores in this field present divergence, with a negative value in a small group (-0.032) and slightly positive in the overall artist group (0.002).

In the entity and phrase-related analyses, person and organisation differences highlight a more increased association of the top 40 artists with diverse individuals and organisations. The award-based analysis also reflects the greater recognition and achievement of the same group. Lexical diversity, however, suggests a narrower and more focused scope of language utilisation for the small group.

Table 1: Descriptive Statistics for all artists.

	Mean	Std. Dev.	Min	Max
Sentiment score	0.66	0.61	-0.999	0.999
Text length	6066.8	5763.6	75	28683
Org. entity	39.34	39.87	0	311
Person entity	55.51	53.31	0	469
Award-based entity	1.96	3.40	0	96
Lexical diversity	0.65	0.11	0.23	1
Popularity-based phrases	0.002	0.029	-0.088	0.095

These descriptive insights collectively emphasise the distinctive attributes of the top 40 artists, charac-

Table 2: Descriptive Statistics for Top-40.

	Mean	Std. Dev.	Min	Max
Sentiment score	0.90	0.43	-0.998	0.999
Text length	20806.6	3587.4	10635	24528
Org. entity	106.38	43.10	50	231
Person entity	181.59	57.19	88	292
Award-based entity	5.08	5.64	0	28
Lexical diversity	0.50	0.03	0.43	0.58
Popularity-based phrases	-0.032	0.025	-0.084	0.022

terised by enriched positivity, broader coverage, increased personal and organisational references, and additional specific recognition patterns, positioning them apart from the wider community.

3.2 Clustering Outputs

K-means clustering is widely used due to its efficiency in offering a scalable interpretable solution for partitioning data into distinct groups based on similarity. This algorithm is implemented to observe the possibility of clustering artists based on their similarities in their Wikipedia entries. Given that, our analyses include K-means clustering with k-means++ initialisation for faster convergence, a maximum of 300 iterations, and ten different initial placements of centroids with the random state 42. The model is run for 2-10 clusters, and the result of 4 is obtained using the elbow method, as indicated in Figure 1.

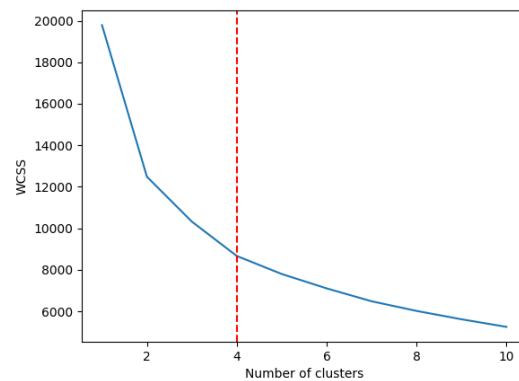


Figure 1: The elbow method.

Figure 2 illustrates the four-cluster split of K-means clustering. It highlights that cluster 0 contains

1261 painters, while clusters 1, 2 and 3 have 803, 440 and 323 painters, respectively. It has been observed that among 2,827 painters, cluster 3 contains more than 87% of the top 40 artist group while only leaving five painters behind, namely Diego Rivera, Giacomo Balla, Hokusai, Rene Magritte, and Umberto Boccioni. Their positions can explain the distinct clustering of these painters in terms of the artistic movements, stylistic element deviations, thematic focus, and content they represent compared to the clustered group of 34 artists.

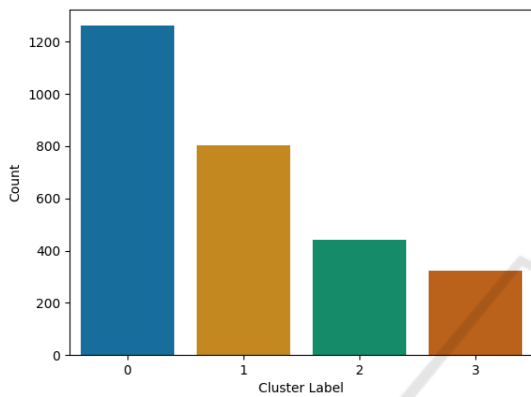


Figure 2: Artist counts in each cluster.

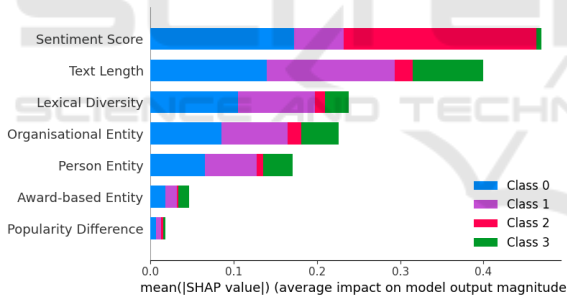


Figure 3: Average impact on model output magnitude.

SHAP (SHapley Additive exPlanations) values are calculated for different features in a model. These values are used to comprehend the influence of each feature on the model’s predictions (Lundberg and Lee, 2017). As illustrated in the stacked horizontal bar chart in Figure 3, the average magnitude of each feature’s SHAP values are quantified using TreeExplainer (Lundberg et al., 2020). The values are provided across all instances; hence, it proposes their relative importance measure. The sentiment score, followed by the text length, forms the two most dominant clusters in the model. The least two significant cluster creation features are award-based entities and the popularity difference.

Individual SHAP values can be portrayed by

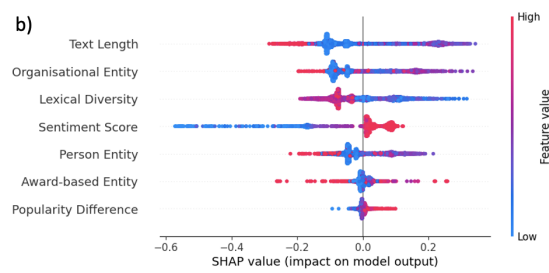
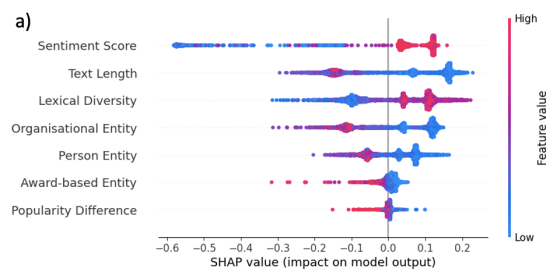


Figure 4: Beeswarm Plots for a) Cluster 0, b) Cluster 1.

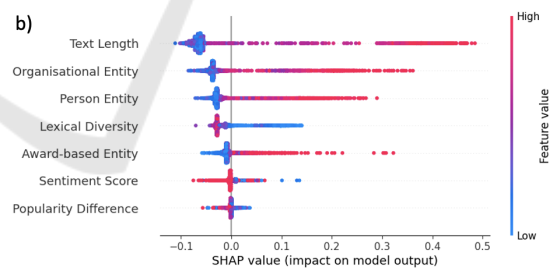
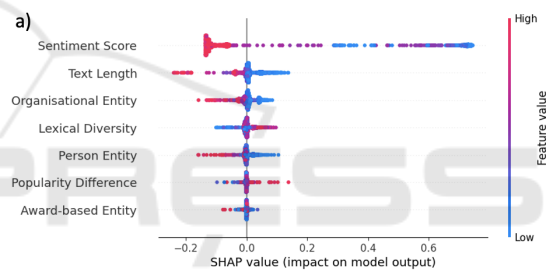


Figure 5: Beeswarm Plots for a) Cluster 2, b) Cluster 3.

beeswarm plots, which can be considered an essential tool in Explainable Artificial Intelligence (XAI). Figures 4 and 5 represent these plots for all clusters. In these figures, the impact and the distribution of the features are presented. Therefore, the model’s interpretability and transparency are enhanced due to their representation of variability and outliers with colour-coded features.

Even though the sentiment score is the predominant feature for clusters 0 (Figure 4a) and 2 (Figure 5a), text length appears to be the most significant in clusters 1 and 3 in Figures 4b and 5b, respectively.

Text length, as the most influencing factor in cluster 3, has been investigated further to observe the bias and the robustness of the results. It has been found that the exclusion of the feature did not yield significant changes in the result in different parameter settings. Text length is followed by the organisational entity in these clusters (1 and 3). This appears substantial because the remaining four artists in the top 40 are not clustered in the same group. As introduced earlier, four of the remaining five artists are clustered in 1. Therefore, it could be inferred that the text length and the organisational entity counts in the Wikipedia page of the artists' form importance within the small group. Although it highlights a wide range, Figure 5b suggests that longer texts are more likely to predict this cluster as the positive values dominate. Organisational entity seems to be the most influential feature, given its position and spread on the plot of cluster 3. While appearing significant in almost all other clusters, the sentiment score does not form a high relative importance for this cluster.

4 CONCLUSION

This study employed an approach for clustering painter profiles using NLP features. By leveraging unsupervised machine learning algorithms such as K-means, painters are clustered based on resemblances in their Wikipedia articles. This method provides an additional perspective in comprehending the textual likenesses and contrasts in articles of painters. Thus, it contributes to further research in art history and analysis. Although the web link for the 40 artists is introduced as a benchmark, it is essential to note that this study does not correlate the artists presented in the link with their appreciation or the value of their artwork. In the analysis, SHapley Additive exPlanations are introduced as it is considered one of the key techniques in XAI. This technique ensures transparency and trust in AI models by investigating the impact of different features on the model output. This study highlights the potential of artists' media visibility. As the online information availability has expanded, so has the frequency of discussions about the artists and their works. Therefore, understanding the similarities among painters using media content is becoming more integral to understanding artists' recognition.

As future research, feeding these clusters and similarities into a pricing algorithm as a feature to value artworks is an interesting direction. This becomes critical, especially if no previous auction has been completed for the artwork of the artist in consider-

ation. Also, the proposed clustering approach can be enriched by including further features such as the perception of quality of artists and their works (which might be obtained by using surveys or resorting to expert opinions), or the reputation of the galleries where the corresponding artworks are frequently sold.

ACKNOWLEDGEMENTS

We acknowledge the financial support for this project provided by the Istanbul Bilgi University Scientific Research Projects Fund under grant number AK85098.

REFERENCES

- Bellogin, A., de Vries, A., and He, J. (2013). Artist popularity: Do web and social music services agree? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 673–676.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- Castellano, G., Lella, E., and Vessio, G. (2021). Visual link retrieval and knowledge discovery in painting datasets. *Multimedia Tools and Applications*, 80:6599–6616.
- Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., and Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry*, 15(9):1679.
- Cieliebak, M., Benites, F., Leuschen, L., Hnizda, M., and Betzler, D. (2019). Natural language processing in arts management. *Zeitschrift für Kulturmanagement*, 5(1):119–142.
- Cole, M. (2022). 40 Famous Artists Everyone Should Know, From Michelangelo to Frida Kahlo — mymodernmet.com. <https://mymodernmet.com/famous-artists/>. [Accessed 01-08-2023].
- Graw, I. (2016). The value of liveliness: Painting as an index of agency in the new economy. *Painting beyond itself: The medium in the post-medium condition*, pages 79–101.
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhajja, B., and Heming, J. (2022). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*.
- Knebel, C. (2007). Anomalies in fine art markets -three examples of an imperfect market for perfect goods. Master's thesis, University of Paderborn, Faculty of Business Administration and Economics, Paderborn, Germany.

- Krasanakis, E., Schinas, E., Papadopoulos, S., Kompatiaris, Y., and Mitkas, P. A. (2018). Venuerank: Identifying venues that contribute to artist popularity. In *ISMIR*, pages 702–708.
- Leslie, I. (2014). Why the mona lisa stands out. *Intelligent Life*.
- Liu, B. (2012). Sentiment analysis: A fascinating problem. In *Sentiment Analysis and Opinion Mining*, pages 1–8. Springer.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., and Forghani, R. (2020). Overview of machine learning part 1: fundamentals and classic approaches. *Neuroimaging Clinics*, 30(4):e17–e32.
- McCarthy, P. M. and Jarvis, S. (2010). Mtl-d, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. ” O’Reilly Media, Inc.”.
- Meurers, D. (2012). Natural language processing and language learning. *Encyclopedia of applied linguistics*, pages 4193–4205.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Mitali, B. and Ingram, P. L. (2018). Fame as an illusion of creativity: Evidence from the pioneers of abstract art. *HEC Paris Research Paper No. SPE-2018-1305, Columbia Business School Research Paper*, (18-74).
- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Powell, L., Gelich, A., and Ras, Z. W. (2020). Applying analytics to artist provided text to model prices of fine art. *Complex Pattern Mining: New Challenges, Methods and Applications*, pages 189–211.
- Ramakrishnan, J., Mavaluru, D., Srinivasan, K., Mubarakali, A., Narmatha, C., and Malathi, G. (2020). Opinion mining using machine learning approaches: a critical study. In *2020 international conference on computing and information technology (ICCIIT-1441)*, pages 1–4. IEEE.
- Sawicki, J., Ganzha, M., and Paprzycki, M. (2023). The state of the art of natural language processing—a systematic automated review of nlp literature using nlp techniques. *Data Intelligence*, pages 1–47.
- Schedl, M., Pohle, T., Koenigstein, N., and Knees, P. (2010). What’s hot? estimating country-specific artist popularity. In *ISMIR*, pages 117–122.
- Seguin, B., Striolo, C., diLenardo, I., and Kaplan, F. (2016). Visual link retrieval in a database of paintings. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part I 14*, pages 753–767. Springer.
- Tekindor, A. A. and McCracken, V. (2012). Uniqueness in art market: Specialization in visual art. In *Agricultural & Applied Economics Association’s 2012 AAEA Annual Meeting: Seattle, August 12–14, 2012, Working Paper*.