Breaking Down SEO Complexity: Bridging PCA and Bayesian-Optimized t-SNE

Amin Karami*, Setinaz Foroudi Ghasemabadi†, Mohammad Hossein Amirhosseini‡
Computer Science and Digital Technologies, University of East London (UEL), London, UK
*a.karami@uel.ac.uk, †Sforoudi@shbre.co.uk, ‡m.h.amirhosseini@uel.ac.uk

***Abstract*—The complexity of Search Engine Optimization (SEO) data requires sophisticated analytical tools. This study integrates Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), optimized by Bayesian methods, to enhance SEO data analysis. PCA is a technique that minimises the number of dimensions in data, allowing for the identification of important aspects related to search engine optimisation (SEO). On the other hand, optimised t-SNE gives a visual representation of data clustering and correlations in a way that is easy to understand and interpret. Our methodology enhances computing efficiency and interpretability, surpassing conventional techniques in analysing both linear and non-linear data. The results develop more strategic decision-making in the field of SEO, indicating a remarkable advancement in SEO analytics.**

***Index Terms*—Search Engine Optimization, PCA, t-SNE, Bayesian Optimization, Clustering, Bokeh, Interactive Visualizations**

I. Introduction

Search Engine Optimization (SEO) has become an essential aspect for businesses and organizations to enhance their online visibility and reach [1], [2]. The complexity of SEO, influenced by various factors such as website structure, content quality, keyword optimization, and backlink profiles, presents significant analytical challenges due to the high-dimensional nature of the data involved [3]–[5]. Dimensionality reduction is a strategic method used to simplify complicated data by finding the most important elements [6], [7].

Principal Component Analysis (PCA) is a well-known method for reducing the dimensions of data by breaking down it into a smaller number of dimensions. The goal is to sustain as much variation as feasible in the process [8]. Although PCA efficiently reduces dimensionality, it inherently fails to capture non-linear relationships, which are often crucial in complex datasets like those in SEO. To address this, enhancements in techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) have been introduced. This research adopts an optimized version of t-SNE, utilizing Bayesian Optimization, to improve the accuracy and efficiency of 2D visualizations. This optimization helps in tuning the perplexity and learning rates, traditionally challenging parameters in t-SNE, to better maintain the local structure of data in visual representations [9]. While dimensionality reduction techniques like PCA and t-SNE are extensively applied in domains such as computer vision, intelligent agents, optimization, robotics and natural language processing, their utilization in SEO data analysis remains limited. This paper seeks to bridge this gap by deploying PCA combined with Optimized t-SNE on a large dataset of SEO metrics collected from Alexa [10]. Alexa's data [11] is renowned for its comprehensive coverage and reliability, making it an exemplary source for analyzing and measuring SEO effectiveness. This dataset provides a robust foundation for our analysis, offering extensive metrics that reflect website performance and SEO strategies across a diverse range of domains. The contributions of this study are as follows:

- A novel application of PCA integrated with Optimized t-SNE for SEO data, providing a methodological advancement in the field.
- The application of proposed technique on real-world SEO data, aiming to identify the most effective approach for visualizing complex relationships between SEO metrics.
- a comprehensive understanding of the essential SEO elements, enabling SEO professionals and decision-makers to get valuable and practical information for improving tactics and strengthening their online visibility.

To further guide this investigation, we pose the following research questions:

1) How do SEO keyword opportunities, site rankings, and categories correlate with each other in lower dimensional spaces?
2) How do SEO keyword metrics, site performance indicators, audience attributes and category associate with each other in lower dimensions?

The literature reveals specific limitations in existing dimensionality reduction techniques [12]–[14]. PCA, while effective for linear dimensionality reduction, often fails to capture non-linear patterns that are crucial in intricate datasets such as SEO [15]. Traditional t-SNE, although effective in maintaining local relationships, can be computationally intensive and its results are highly sensitive to the choice of hyperparameters [16]. This study employs Bayesian Optimisation with t-SNE to effectively tackle these problems by methodically optimising the hyperparameters in order to get consistent and meaningful visual results. Furthermore, the field of SEO is insufficient in sophisticated analytical instruments capable of efficiently managing the intricate and varied characteristics of SEO data. This research seeks to address these deficiencies by introducing sophisticated visualisation techniques using Bokeh [17] that aid stakeholders in discovering diverse patterns and linkages, which are crucial for making informed strategic

choices. Bokeh is a robust visualisation library that specialises in presenting intricate SEO data in an interactive manner. Bokeh offers a wide range of customisation options, enabling users to create visualisations that are especially designed for analysing patterns and performance indicators in SEO data. The interactive nature of Bokeh data visualisations allows individuals to dynamically analyse SEO indicators, facilitating more informed decision-making processes. In summary, The main contribution of this research is the novel application of PCA integrated with optimized t-SNE using Bayesian Optimization for analyzing and visualizing complex SEO data for an effective SEO strategies. The rest of the paper is organized as follows: Section II presents the mathematical underpinnings of the methods utilized in our approach, Section III discusses related work. Section IV describes the data used in this study and the methodology and the performance measurements. Section V presents results and Section VI concludes the paper.

II. Mathematical Foundations

This section explores the mathematical foundations of the techniques used in our approach, including Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Bayesian Optimisation. These strategies are fundamental for improving the processes of reducing dimensionality and visualising SEO data analysis.

A. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to reduce the dimensions of a data set, simplifying the complexity while retaining the variation present in the dataset.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Where: - $\mathbf{X}$ is the data matrix, - $\mathbf{U}$ is the left singular vectors, - $\mathbf{\Sigma}$ is the diagonal matrix of singular values, - $\mathbf{V}^T$ is the right singular vectors (transpose of $\mathbf{V}$).

The goal is to project $\mathbf{X}$ onto a space defined by the first few principal components that capture the most variance in the data.

B. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

The t-SNE algorithm comprises two main stages:

1. Computation of the conditional probability $p_{j|i}$ that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i.

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

2. Minimization of the Kullback-Leibler divergence between the joint probabilities p_{ij} of the low-dimensional embedding and the high-dimensional data.

$$C = \mathrm{KL}(P||Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

where q_{ij} is defined similarly to p_{ij} but in the lower-dimensional space.

C. Bayesian Optimization

Bayesian Optimization is a strategy for the global optimization of noisy black-box functions. It builds a probabilistic model for the objective function and uses it to make decisions about where next to evaluate the function [18]. The typical steps involved are:

- Choose a prior over functions that captures our beliefs about the behavior of the function being optimized.
- Observe the function at new points according to a policy based on the current model.
- Update the posterior probability distribution over functions given the observed data.

$$\mu_{t+1}(x), \sigma^2_{t+1}(x) = \text{BayesianUpdate}(\mu_t(x), \sigma^2_t(x), x_t, y_t)$$

Where $\mu_t(x)$ and $\sigma^2_t(x)$ are the mean and variance of the model at step t, and x_t, y_t are the new observation points. Acquisition functions, such as Expected Improvement, guide where to sample next:

$$EI(x) = \mathbb{E}[\max(f(x) - f(x^+), 0)]$$

III. Literature Review

SEO analysis involves diverse strategies, from basic keyword optimization to advanced machine learning for predicting search engine outcomes. This study focuses on incorporating dimensionality reduction into SEO analysis, highlighting key research contributions, strengths, and limitations.

Tran (2024) discusses SEO poisoning risks for SMEs, including tactics like malicious backlinks and cloaking, which harm reputation and finances. However, it overlooks using lower-dimensional visualization to illustrate these features. Paramasivam (2022) explores how big data analytics enhance SEO by improving content analysis and result relevance. The study struggles with managing complex, high-dimensional data and doesn't address simplifying this data while retaining key insights. Shayegan (2020) finds that metrics like backlinks and Page Rank significantly impact university website rankings. By evaluating 38 SEO metrics, it uncovers patterns to improve rankings but fails to reduce data to fewer dimensions for critical pattern visibility. Vyas (2019) ranks Indian tourism websites using SEO tools like TrafficEstimate and SEMRUSH, offering a method for future evaluations. The study lacks lower-dimension visualization techniques, which could simplify complex data, revealing patterns and trends in SEO performance.

TABLE I: Description of Alexa data

Column Name	Description
Category	Category / Subcategory of site
all topics keyword gaps Avg traffic *	An estimate of the traffic that competitors are getting for this keyword. Scores by competitor traffic rank in range(1, 100)
all topics keyword gaps search popularity *	An estimate of how frequently this keyword is searched across all search engines. Scores by popularity in range(1, 100)
all topics easy to rank keywords search pop *	An estimate of how frequently this keyword is searched across all search engines. Scores by popularity in range(1, 100)
all topics buyer keywords Avg traffic *	An estimate of the traffic that competitors are getting for this keyword. The score is based on the popularity of the keyword, and how well competitors rank for it. Scores by traffic in range(1, 100)
all topics buyer keywords organic competition *	An estimate of how difficult it is to rank highly for this keyword in organic search. Scores by competition in range(1, 100)
all topics optimization opportunities search pop *	An estimate of how frequently this keyword is searched across all search engines. Scores by popularity in range(1, 100)
all topics optimization opportunities organic share of voice *	The percentage of all searches for this keyword that sent traffic to this website
all topics top keywords search traffic *	The percentage of organic search referrals to this site that come from this keyword
all topics top keywords share of voice *	The percentage of all searches for this keyword that sent traffic to this website
Alexa Rank	An estimate of this site's popularity based on global internet engagement
Daily time on site	Average time in minutes and seconds that a visitor spends on this site each day

TABLE II: Generated Categorical Attributes

Attribute	Description
Ultimate Category	Maps each category to a numeric index
Top Category	Maps top categories to number between 1-15
Traffic to Own	Bins sites by self-traffic percentage
Traffic to Competitors	Bins sites by competitor traffic percentage
Performance Tier	Divides sites into global rank tiers
Keyword Difficulty	Groups Keywords driving traffic to competitors
Breakdown Opportunities	Popular keywords driving some traffic to this site

IV. Methodology and Knowledge Contribution

The dataset [11] utilized in this study is obtained from Alexa.com, consisting of 12,000 rows and 84 attributes. These attributes are a mix of numeric and textual data types, providing a broad spectrum of information related to website performance metrics and SEO strategies. The breakdown of dataset is given in Table I and as follows:

- **Numeric Attributes (45 Columns)**: Metrics such as traffic scores, search popularity, competition levels, and specific keyword-related traffic percentages.
- **Textual Attributes (11 Columns)**: Groups of keywords crucial for SEO, with associated numeric metrics indicating their impact.
- **Categorical and Descriptive Attributes (28 Columns)**: Includes website categories and other descriptive information influencing website classification.

Table II summarizes the seven categorical attributes that we generated from the SEO dataset. These attributes provide meaningful representations of key factors such as categories, traffic sources, performance metrics, keyword profiles, and user behavior. By mapping numeric and text data into categorical bins and groups, the attributes enhance interpretability and aid in visualizing complex relationships in the data for strategic analysis. The transformed attributes form a basis for clustering websites and gaining actionable insights into optimization opportunities, competitive landscapes, and audience patterns. The proposed model in Algorithm 1 involves key steps to effectively visualize high-dimensional data in 2D, using PCA for reduction and t-SNE for visualization, optimized through Bayesian techniques.

Algorithm 1 Optimized Data Analysis and Visualization

Require: DataFrame df

Initialization:

$data \leftarrow df[selected\_features]$

$space \leftarrow [(10, 1000, \text{'learning\_rate'}), (1, 50, \text{'early\_exaggeration'}), (10, 200, \text{'perplexity'})]$

Step 1: Principal Component Analysis (PCA)

$variance\_thresholds \leftarrow [0.90, 0.95, 0.99]$

$variance\_components \leftarrow$ empty DataFrame(columns = ['variance', 'n_components'])

Create Figure for Plots

for each $variance$ in $variance\_thresholds$ **do**

  $pca \leftarrow PCA().fit(data)$

  $cum\_var \leftarrow$ cumulative sum of $pca.explained\_variance\_ratio\_$

  Plot cum_var against number of components

  $num\_components \leftarrow$ find minimum components s.t. $cum\_var \geq variance$

  Highlight $num\_components$ in plot

  $variance\_components \leftarrow variance\_components$ + [variance, $num\_components$]

end for

$optima\_n\_components \leftarrow$ find $n\_components$ for 90% variance in $variance\_components$

Step 2: t-SNE Optimization via Bayesian Optimization

$X\_pca \leftarrow PCA(n\_components = optima\_n\_components).fit\_transform(data)$

$results\_df \leftarrow$ empty DataFrame

Define objective function for t-SNE using $space$

$best\_params \leftarrow$ Bayesian optimization to minimize t-SNE KL divergence

Record $best\_params$ and corresponding KL divergence in $results\_df$

Step 3: Visualization

Configure TSNE with $best\_params$

$X\_tsne \leftarrow TSNE$.fit_transform(X_pca)

Plot 2D visualization of X_tsne

Ensure: 2D visualization plot

Step 1: Feature Selection and Hyperparameter Space to definition the range of values for the learning rate, early exaggeration, and perplexity which are crucial for the t-SNE optimization.

Step 2: PCA for Dimensionality Reduction

- `Variance Thresholds`: Evaluates thresholds (90%, 95%, 99%) to determine the optimal PCA components, balancing data simplification and information retention.
- `Cumulative Explained Variance`: Identifies the number of components needed to explain a set proportion of total variance.
- `Determination of optima_n_components`: Selects components explaining 90% variance for efficiency and retention balance.
- The method adjusts PCA components to retain 90% variance, optimizing simplification while preserving crucial information for high-dimensional analysis.

Step 3: Optimization of t-SNE

- `Objective Function`: Defines the negative KL divergence to assess how well t-SNE preserves data structure.
- `Bayesian Optimization`: Finds optimal hyperparameters (learning rate, early exaggeration, perplexity) to minimize the objective function, enhancing t-SNE visualization quality.
- Bayesian optimization systematically explores hyperparameters, focusing on early exaggeration, to improve global data structure detection, leading to superior clustering and visualization outcomes.

Step 4: 2D Visualization

- `t-SNE Configuration and Execution`: With the optimized hyperparameters, t-SNE is applied to the PCA-reduced data to produce a 2D visualization.
- `Bokeh Plotting`: The final 2D visualization helps in understanding the underlying structure and clusters within the data.

A. Performance Measurement

1) KNN Consistency Score: The KNN Consistency Score quantifies the degree to which the local neighborhood structure is preserved in a lower-dimensional space. It is defined as the average proportion of intersecting k-nearest neighbors in the original and reduced spaces.

Given: - $\mathcal{X}$ as the original high-dimensional data, - $\mathcal{Y}$ as the reduced low-dimensional data, - k as the number of nearest neighbors, - $\text{NN}_k(\mathcal{X}, i)$ as the set of k-nearest neighbors of point i in $\mathcal{X}$, - $\text{NN}_k(\mathcal{Y}, i)$ as the set of k-nearest neighbors of point i in $\mathcal{Y}$,

the KNN Consistency Score C is calculated as:

$$C = \frac{1}{n} \sum_{i=1}^{n} \frac{|\text{NN}_k(\mathcal{X}, i) \cap \text{NN}_k(\mathcal{Y}, i)|}{k}$$

where n is the number of points in the dataset.

2) Stress Value: The Stress Value measures the fidelity of distances in a dimensionality reduction mapping. It is defined as the normalized root-mean-square deviation between the distances in the original and reduced spaces.

Given: - $D_{\mathcal{X}}$ as the matrix of pairwise distances in the original space $\mathcal{X}$, - $D_{\mathcal{Y}}$ as the matrix of pairwise distances in the reduced space $\mathcal{Y}$,

the Stress Value S is calculated as:

$$S = \sqrt{\frac{\sum_{i,j} (D_{\mathcal{X}}(i,j) - D_{\mathcal{Y}}(i,j))^2}{\sum_{i,j} D_{\mathcal{X}}(i,j)^2}}$$

Both metrics are essential for evaluating the quality of dimensionality reduction techniques. The KNN Consistency Score focuses on local neighborhood preservation, while the Stress Value assesses the global fidelity of the distance relationships.

V. Implementation and Experimental Results

All experiments were conducted on a high-performance single workstation designed to handle extensive computational tasks with efficiency and reliability. The specific hardware specifications include Intel(R) Core(TM) i7-9750H CPU, 128 GB DDR4 memory and 1 TB NVMe SSD storage. To ensure the robustness and reliability of our results, the execution of each method was conducted 10 times using various random seeds to consider the inherent variability in initialization that might impact the results, especially in stochastic algorithms like t-SNE. This approach allowed us to mitigate any potential biases introduced by specific initial conditions and provided a more comprehensive assessment of each method's performance across various runs [19].

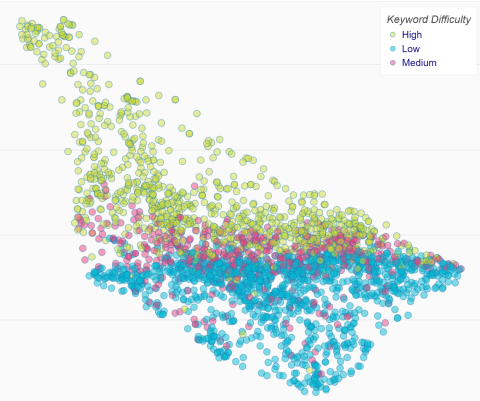
A. Experiment 1

This experiment aims to explore the underlying relationships between various SEO optimization metrics and website performance factors using a subset of features from the dataset. Specifically, we seek to address the following research question:

Research question: How do SEO keyword opportunities, site rankings, and categories correlate with each other in lower dimensional spaces?

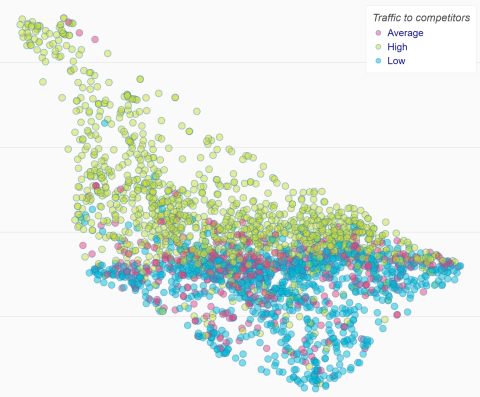
The selected features for this experiment are shown in Table III. The optimal number of PCA components was determined to be 7, capturing 90% of the data variance. Table IV shows results from 20 trials optimizing t-SNE parameters such as learning rate, early exaggeration, and perplexity to improve visualization quality. The learning rate dictates step size, crucial for smooth convergence, while early exaggeration enhances initial cluster separation. Perplexity balances local and global data structures, influencing neighbor weighting. Epoch 11 provides the best outcome with an objective value of -4.760286, using a learning rate of 1.0, early exaggeration of 50, and perplexity of 20. This setup effectively separates data clusters and balances local/global structures. We evaluated performance using the KNN Consistency Score and Stress Value in Table V, with our PCA and optimized t-SNE approach surpassing others by preserving data relationships in lower dimensions.

Figure 1 shows a 2D scatter plot with color-coded markers indicating keyword difficulty and traffic categories, revealing clustered patterns suggesting intrinsic data groupings. Overlap in colors might reveal hidden similarities between websites. With over eight thousand data points, detailed features are hard to discern, but interactive tools like Bokeh allow focused

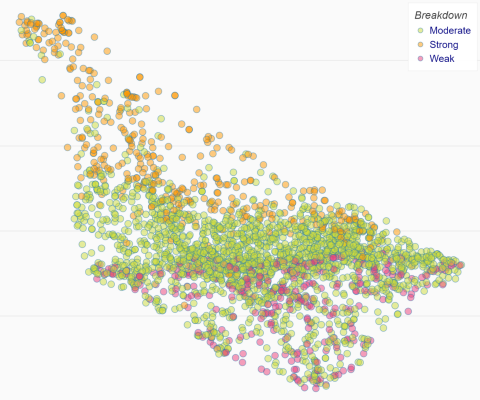


(a) Keywords driving traffic to competitors

(b) Alexa Rank of site's popularity



(c) Sites that direct traffic to competitors



(d) Popular keywords driving traffic to sites

Fig. 1: Strategic Insights in Experiment 1

TABLE III: Selected Features for Experiment 1

Feature
keyword_opportunities_breakdown_easy_to_rank_keywords
keyword_opportunities_breakdown_optimization_opportunities
all_topics_optimization_opportunities_search_pop_parameter_1
all_topics_optimization_opportunities_search_pop_parameter_2
all_topics_optimization_opportunities_search_pop_parameter_3
all_topics_optimization_opportunities_search_pop_parameter_4
all_topics_optimization_opportunities_organic_share_of_voice_parameter_1
all_topics_optimization_opportunities_organic_share_of_voice_parameter_2
all_topics_optimization_opportunities_organic_share_of_voice_parameter_3
all_topics_optimization_opportunities_organic_share_of_voice_parameter_4
this_site_rank_in_global_internet_engagement
top_Category
ultimate_category

TABLE IV: Summary of Results for Experiment 1

Epoch	Learning Rate	Early Exaggeration	Perplexity	Objective
1	209.5367	8.107412	71	-1.362688
2	431.516026	16.684329	69	-1.431347
3	982.482431	9.606431	92	-1.319095
4	144.292783	9.570845	21	-1.484736
5	442.447199	18.865138	28	-1.51716
6	759.7253	13.464019	7	-1.359736
7	632.536824	32.70945	38	-1.659748
8	578.292784	32.42662	86	-1.448312
9	567.571937	13.181008	57	-1.424339
10	398.76092	16.807423	72	-1.421611
11	1.0	50.0	20	-4.760286
12	1000.0	50.0	5	-2.304346
13	1.0	50.0	5	-2.399555
14	1.0	50.0	23	-4.621037
15	1.0	45.600724	20	-4.760286
16	27.646514	7.68975	20	-1.50916
17	1000.0	47.785428	20	-1.93282
18	1.517031	19.30073	20	-1.917288
19	1.0	47.631843	20	-4.762286
20	1.0	47.813531	20	-4.762286

exploration. Figure 2 illustrates visual analytics using Lasso and Box Zoom tools, with tooltips enhancing analysis by providing detailed insights into specific data subsets.

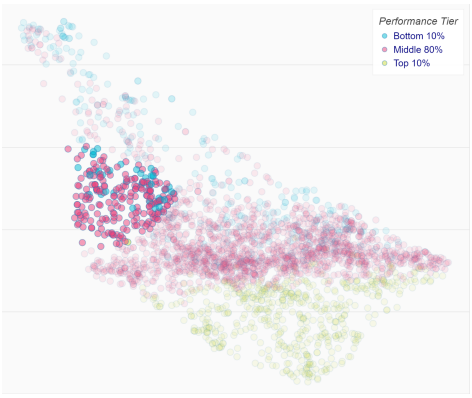
TABLE V: Performance comparison of dimensionality reduction techniques

Method	KNN=50	KNN=100	KNN=200	Stress
PCA	0.1511	0.2012	0.2533	1.0015
t-SNE	0.1432	0.1948	0.2406	1.1027
Optimized t-SNE	0.1701	0.1545	0.2401	1.0003
PCA + t-SNE	0.1689	0.2145	0.2392	1.0012
PCA + Optimized t-SNE (Proposed)	**0.1804**	**0.2358**	**0.2867**	**0.9894**

B. Experiment 2

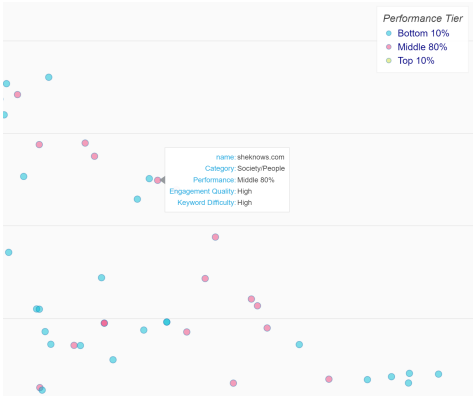
This experiment aims to explore the relationships between SEO keyword opportunities, website traffic metrics, audience overlap factors and category using a subset of features. Specifically, we seek to address the following research question:

Research question: How do SEO keyword metrics, site performance indicators, audience attributes and category associate with each other in lower dimensions?



(a) The Lasso selection tool to define an arbitrary region

(b) The Box zoom to define a rectangular region to zoom the plot



(c) The use of tooltips to attach additional insightful information

Fig. 2: The Visual Analytics

The selected features for this experiment are shown in Table VI. The optimal PCA components were 5, with Bayesian optimization selecting a learning rate of 1.0, early exaggeration of 40.0, and perplexity of 52. Table VII summarizes the impact of these parameters on the optimization objective value, where lower values indicate better embedding quality. Experiments show that a learning rate of 1.0 with high early exaggeration (40.0) consistently results in lower objective values, enhancing data embedding quality. Epochs 15, 16, 18, and 19, with early exaggeration at 40 and perplexity between 52 and 100, achieved the lowest values (-3.831865, -3.176995, -3.655817, -3.757676), highlighting the effectiveness of high early exaggeration in separating clusters.

TABLE VI: Selected Features for Experiment 2

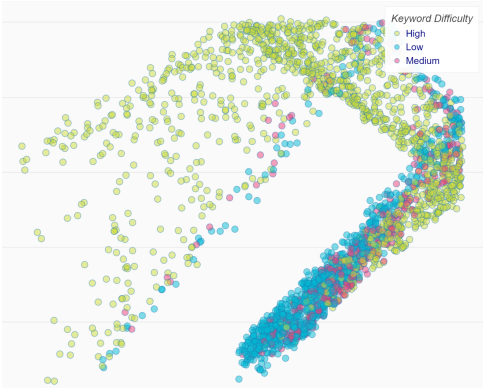
Feature
keyword_opportunities_breakdown_easy_to_rank_keywords
keyword_opportunities_breakdown_buyer_keywords
all_topics_optimization_opportunities_organic_share_of_voice_parameter_1
all_topics_optimization_opportunities_organic_share_of_voice_parameter_2
all_topics_optimization_opportunities_organic_share_of_voice_parameter_3
all_topics_optimization_opportunities_organic_share_of_voice_parameter_4
comparison_metrics_search_traffic_this_site_percentage
audience_overlap_sites_overlap_scores_parameter_1
audience_overlap_sites_overlap_scores_parameter_2
audience_overlap_sites_overlap_scores_parameter_3
audience_overlap_sites_overlap_scores_parameter_4
audience_overlap_sites_overlap_scores_parameter_5
daily_time_on_site
top_Category

In contrast, higher learning rates (314.926131 to 982.572689) with moderate to high early exaggeration and perplexity led to higher objective values, indicating less optimal embeddings. Configurations 2, 3, and 11, despite high early exaggeration, did not perform as well as those with a learning rate of 1.0. This suggests that a lower learning rate and high early exaggeration are key for optimal t-SNE performance, especially with high perplexity, ensuring a clearer data structure representation in reduced dimensions.

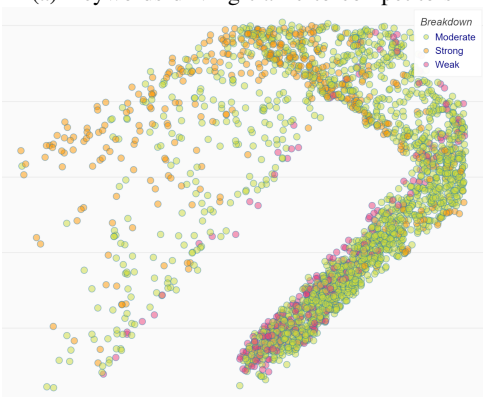
TABLE VII: Summary of Results for Experiment 2

Epoch	**Learning Rate**	**Early Exaggeration**	**Perplexity**	**Objective**
1	439.542008	11.854345	44	-1.821848
2	934.448542	32.229355	9	-2.010115
3	920.498233	47.173984	70	-1.918307
4	975.590502	11.125027	29	-1.844133
5	910.407794	35.153277	98	-1.660789
6	403.885743	9.905242	66	-1.703206
7	562.545851	28.958231	17	-2.132180
8	920.583866	1.823171	80	-1.616960
9	490.139861	7.089334	49	-1.737696
10	982.572689	15.171143	42	-1.850279
11	488.126298	32.860991	16	-2.028810
12	314.926131	27.827017	20	-1.996918
13	1.000000	50.000000	5	-2.634354
14	1.000000	1.000000	5	-2.550213
15	1.000000	40.000000	52	-3.831865
16	1.000000	50.000000	100	-3.176995
17	1.000000	1.000000	67	-1.867084
18	1.000000	50.000000	62	-3.655817
19	1.000000	50.000000	56	-3.757676
20	1.000000	50.000000	67	-2.578160

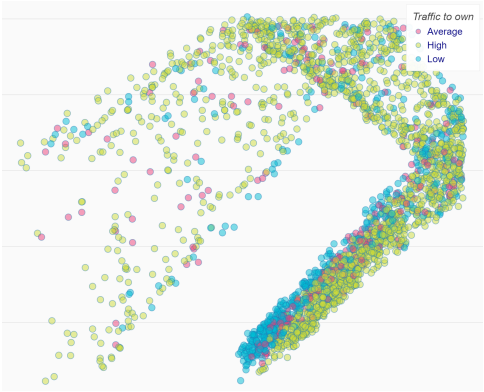
Table VIII compares performance measurement metrics for our approach against other relevant dimensionality reduction techniques for the second experiment. The outcomes demonstrate that our proposed approach outperforms alternative methods across all evaluation metrics. This method successfully maintains both local and global relationships within the data at reduced dimensions, thereby confirming the efficacy of our technique.



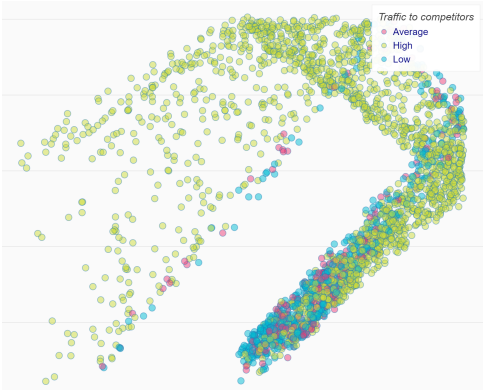
(a) Keywords driving traffic to competitors



(b) Popular keywords driving traffic to sites



(c) Sites by self-traffic



(d) Sites that direct traffic to competitors

Fig. 3: Strategic Insights in Experiment 2

TABLE VIII: Performance comparison for Experiment 2

Method	KNN=50	KNN=100	KNN=200	Stress
PCA	0.1214	0.1705	0.222	1.1001
t-SNE	0.1161	0.1634	0.2108	1.0042
Optimized t-SNE	0.1274	0.1812	0.2048	1.0013
PCA + t-SNE	0.1416	0.1701	0.2143	1.0024
PCA + Optimized t-SNE (Proposed)	**0.1511**	**0.1972**	**0.2547**	**0.9999**

After examining the 2D visualization outcomes obtained in the Figure 3, some clear insights could potentially be identified. The compact centre core consists of several tightly connected nodes (websites), indicating significant similarity in their underlying traffic patterns. On the other hand, less populated outside areas have clearer categorizations with less links between them in terms of the depicted four categories including keyword difficulty, traffic breakdown, traffic to own and traffic to competitors. several nodes (websites) have a more compact grouping compared to others, as if they are separated by particularly distinctive characteristics. This visualisation effectively differentiates between coherent groups and more isolated categories. Similar to Figure 2 from experiment 1, we can conduct a similar detailed analysis to facilitate a more insightful examination of the relationships among the displayed Alexa websites.

In summary, our proposed method (PCA + Bayesian Optimized t-SNE) stands out as the most advanced method among the listed options, offering an optimal balance of visualization quality, computational efficiency, and scalability in Table IX. It is particularly suited for complex and large-scale data environments like SEO analytics, where both linear and non-linear patterns are crucial. It outperforms other methods by leveraging the strengths of both PCA and t-SNE while addressing their individual limitations through sophisticated optimization techniques.

VI. Conclusion

In this study, we applied a novel methodology combining PCA, optimized t-SNE and interactive visualizations using Bokeh to analyze an Alexa SEO dataset. Two experiments addressed relationships between SEO metrics, performance factors and traffic patterns in lower dimensions. The quantitative findings showed that the preservation of both local and global data structures was achieved, outperforming the performance of existing strategies. This study represents a notable progress in SEO analytics by transforming complexities into easily understandable representations. The acquisition of insights enables a more strategic approach to optimisation, revealing hidden trends that influence online visibility. This study establishes a foundation for future use of advanced technologies to get practical suggestions from extensive SEO datasets. Despite these advancements, the limitation in this research includes the scalability with very large datasets to the computational intensity of t-SNE. Future work should focus on generalizing the proposed method across diverse datasets.

TABLE IX: Comparison of Applied Techniques

Model	Effectiveness	Quality of Visualization	Computational Efficiency	Scalability
PCA	Good for linear dimensionality reduction, misses non-linear relationships	Good for linear separations, may miss complex patterns	Very high	Very scalable
t-SNE	Excellent for non-linear relationships, maintains local data structure	Excellent at revealing patterns and clustering	Low, computationally expensive	Poor, struggles with large datasets
Optimized t-SNE	Improves clustering through parameter optimization	Better and more consistent visualizations than standard t-SNE	Improved but still resource-heavy	Slight improvements but still limited
PCA + t-SNE	Combines linear and non-linear reductions effectively	Clearer visualizations due to initial PCA processing	Moderate, benefits from PCA but limited by t-SNE's demands	Better than t-SNE alone due to PCA preprocessing
PCA + Bayesian Optimized t-SNE	Best for handling complex datasets with both linear and non-linear patterns	Likely produces the best visualizations with superior detail	Higher than standalone t-SNE due to optimizations	Best among these, especially for large datasets

REFERENCES

[1] J. Al-Gasawneh, M. Alsoud, Z. M. Alhawamdeh, T. J. Bani-Ata, M. Alghizzawi, and M. K. Daoud, "Exploring the influence of digital marketing strategies on enhancing customer satisfaction in contemporary business environments," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, 2024, pp. 1–7.

[2] R. Asrigo and E. R. Kaburuan, "Improving e-commerce website rank using search engine optimization (seo)," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 14s, pp. 430–440, 2024.

[3] T. D. Le, T. Le-Dinh, and S. Uwizeyemungu, "Search engine optimization poisoning: A cybersecurity threat analysis and mitigation strategies for small and medium-sized enterprises," *Technology in Society*, vol. 76, p. 102470, 2024.

[4] G. Egri and C. Bayrak, "The role of search engine optimization on keeping the user on the site," *Procedia Computer Science*, vol. 36, pp. 335–342, 2014, complex Adaptive Systems Philadelphia, PA November 3-5, 2014.

[5] D. Chaffey, "Understanding the complexity of search engine optimization," *Journal of Marketing Communications*, vol. 28, no. 2, pp. 113–130, 2022.

[6] H. Liu and S. Yan, "A survey of dimensionality reduction techniques," *Pattern Recognition*, vol. 50, pp. 1–12, 2017.

[7] A. Karami and M. Guerrero Zapata, "Mining and visualizing uncertain data objects and named data networking traffics by fuzzy self-organizing map," in *Proceedings of the Second International Workshop on Artificial Intelligence and Cognition (AIC 2014): Torino, Italy, November 26-27, 2014*. CEUR-WS. org, 2014, pp. 156–163.

[8] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2016.

[9] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[10] I. Alexa Internet, "Alexa developer guide," 2023, available online.

[11] A. Risheh, A. Goharfar, and N. T. Javan, "Clustering alexa internet data using auto encoder network and affinity propagation," in *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, 2020, pp. 437–443.

[12] N. Mylonas, I. Mollas, N. Bassiliades, and G. Tsoumakas, "Exploring local interpretability in dimensionality reduction: Analysis and use cases," *Expert Systems with Applications*, vol. 252, p. 124074, 2024.

[13] Y. Xu and E. Li, "Robust locally nonlinear embedding (rlne) for dimensionality reduction of high-dimensional data with noise," *Neurocomputing*, p. 127900, 2024.

[14] A. Karami, "An anomaly-based intrusion detection system in presence of benign outliers with visualization capabilities," *Expert Systems with Applications*, vol. 108, pp. 36–60, 2018.

[15] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[16] L. van der Maaten, "Accelerating t-sne using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245, 2014.

[17] B. D. Team, "Bokeh: Python library for interactive visualization," https://docs.bokeh.org/en/latest/, 2014.

[18] A. Karami and R. Johansson, "Utilization of multi attribute decision making techniques to integrate automatic and manual ranking of options," *Journal of information science and engineering*, vol. 30, no. 2, pp. 519–534, 2013.

[19] A. Karami, M. Shemshaki, and M. A. Ghazanfar, "Exploring the ethical implications of ai-powered personalization in digital marketing," *Data Intelligence*, 2024.