A Novel Regression Approach: Analyzing Textual Data in Similarity Space

Ondřej Rozinek
University of Pardubice
Pardubice, Czech Republic
ondrej.rozinek@gmail.com

Monika Borkovcová
University of Pardubice
Pardubice, Czech Republic
monika.borkovcova@upce.cz

***Abstract*—The proliferation of textual data, notably in the form of database records, calls for innovative methods of analysis that go beyond traditional numerical techniques. While least squares regression has been a cornerstone in quantitative data analysis, its applicability to textual data remains largely unexplored. This study aims to bridge this gap by introducing a similarity-based least squares method tailored for textual data. Drawing on the principles of similarity measures in text, such as semantic and syntactic closeness, we propose an extension to the conventional least squares framework. Our approach incorporates word-based similarity metrics into the least squares objective function, enabling the analysis of textual data in a manner coherent with its qualitative nature. The developed methodology is rigorously evaluated using both synthetic and real-world database records, demonstrating its efficacy in uncovering intricate relationships within textual data. Our findings open new avenues for textual data analysis, blending the precision of classical statistical methods with the subtleties of text similarity.**

I. Introduction

In the vast realm of data analysis, the treatment and understanding of textual data, particularly database records, have become increasingly paramount. As the digital universe grows exponentially, with an estimated 2.5 quintillion bytes of data produced daily, a significant portion of this avalanche is textual data [1]. These are the records that detail transactions, logs, communications, and countless other human and machine interactions. Understanding the patterns, structures, and relationships within this textual data offers profound opportunities for knowledge discovery and decision support [2].

Traditional data analysis methodologies, prominently the least squares regression, have been foundational in the domain of quantitative data [3]. Rooted in the early 19th century and attributed to Legendre [4] and Gauss [5], the least squares method has been an invaluable tool in deducing relationships within data, finding applications from astronomy to economics. But can this time-tested method be adapted to the nuanced realm of textual data?

Textual data, unlike quantitative data, primarily relies on the notion of 'similarity' rather than 'magnitude' [6]. Two words or phrases may not exhibit a quantifiable difference, but they can show varying degrees of similarity based on their semantics, usage, or context [7]. Because of this property of textual data, there's a need for an analysis method that recognizes its qualitative aspect. This has led to the idea of modifying the least squares method to work within a similarity framework.

In this paper, we investigate the application of least squares regression in the context of textual data similarity. We introduce a method to adapt traditional regression techniques to work in a similarity space, with a primary focus on word similarities in database records. Our approach involves defining an appropriate similarity metric, reshaping the problem space, and ensuring results are both robust and interpretable [8].

With this research, we intend to enhance data analysis techniques, bridging the gap between the quantitative precision of traditional methods and the qualitative depth of textual data [2].

The primary areas of focus in this paper are regression analysis and similarity spaces, as we aim to adapt the regression analysis framework to function within similarity spaces.

II. Related Work

In the field of computer science, linear regression remains a fundamental technique for modeling the relationship between variables. Over the years, various objective functions and modifications have been proposed to enhance the performance and versatility of linear regression models. In this section, we provide an overview of the state-of-the-art objective functions and their popular modifications.

Ordinary Least Squares (OLS)

The most widely used objective function for linear regression is Ordinary Least Squares (OLS). OLS aims to minimize the sum of squared differences between predicted and actual target values [4], [5].

A. Regularized Regression

In the pursuit of addressing overfitting and enhancing model generalization, regularized linear regression methods have gained substantial popularity. These methods augment the traditional Ordinary Least Squares (OLS) objective function by incorporating penalty terms that encourage specific properties in the model. Here, we discuss three widely used regularized linear regression techniques:

1) Ridge Regression (L2 Regularization): Ridge regression, introduced by Tikhonov [9], extends the OLS framework by adding an L2 regularization term to the objective function.

This regularization term penalizes the magnitude of the coefficients, thereby encouraging smaller coefficients. The Ridge Regression objective function is defined as:

$$\hat{\theta} = \arg\min_{\theta} J_{\mathrm{R}}(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \quad (1)$$

Here, θ denotes the model parameters (coefficients), $\hat{y}_i$ is the predicted value, y_i is the actual target value, n is the number of features, and λ is the regularization parameter.

Ridge regression encourages smaller coefficient values, effectively reducing overfitting and improving model generalization.

2) Lasso Regression (L1 Regularization): Lasso regression, introduced by Tibshirani [10], promotes sparsity within the model by incorporating an L1 penalty term in the objective function. The Lasso Regression objective function is defined as:

$$\hat{\theta} = \arg\min_{\theta} J_{\mathrm{L}}(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} |\theta_j| \quad (2)$$

Similar to Ridge regression, the model parameters θ are adjusted during training, but Lasso encourages some coefficients to be exactly zero, effectively performing feature selection.

3) Elastic Net Regression (L1 + L2 Regularization): Elastic Net Regression, proposed by Zou and Hastie [11], strikes a balance between Ridge and Lasso regression by combining both L1 and L2 regularization terms. The Elastic Net objective function is defined as:

$$\hat{\theta} = \arg\min_{\theta} J_{\mathrm{EN}}(\theta) \quad (3)$$

$$= \arg\min_{\theta} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + \lambda_1 \sum_{j=1}^{n} |\theta_j| + \lambda_2 \sum_{j=1}^{n} \theta_j^2 \quad (4)$$

Here, λ_1 and λ_2 are regularization parameters that control the strength of L1 and L2 regularization, respectively.

Elastic Net provides a versatile regularization approach, allowing users to balance feature selection and regularization according to their specific needs.

4) Dropout: Dropout is a regularization technique introduced by Srivastava et al. [12] and is commonly used in neural networks. It involves randomly setting a fraction of neuron activations to zero during each training iteration, effectively creating an ensemble of subnetworks. The dropout technique helps prevent overfitting and encourages robustness in deep learning models.

These regularized linear regression techniques extend the classical OLS method by incorporating penalty terms that encourage certain properties in the learned models. The choice between these techniques depends on the data characteristics and the desired characteristics of the regression model.

B. Robust Regression

In scenarios where the data may be contaminated with outliers, robust regression techniques have been developed to provide more resilient modeling. These methods aim to minimize the impact of outliers on the model while still capturing the underlying trends in the majority of the data. One notable approach is Huber loss, introduced by Huber [13].

1) Huber Loss: Huber loss combines the benefits of both mean squared error (MSE) and mean absolute error (MAE) and is designed to handle data with outliers more effectively. The Huber Loss objective function is defined as follows:

$$\hat{\theta} = \arg\min_{\theta} J_{\mathrm{H}}(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} L_{\delta}(\hat{y}_i - y_i) \quad (5)$$

In this equation, n represents the number of data points, θ denotes the model parameters (coefficients), $\hat{y}_i$ is the predicted value, and y_i is the actual target value. The function L_{δ} is a piecewise loss function that combines the properties of MSE for small errors and MAE for large errors:

$$L_{\delta}(z) = \begin{cases} \frac{1}{2} z^2, & \text{if } |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta), & \text{if } |z| > \delta \end{cases}$$

Here, δ is a tuning parameter that controls the threshold for switching between the quadratic and linear regions. For small errors ($|z| \leq \delta$), the loss is quadratic, similar to MSE. For large errors ($|z| > \delta$), the loss is linear, similar to MAE.

Huber loss offers a robust alternative to traditional OLS by providing a balanced approach to handle outliers while still maintaining the benefits of squared loss for small errors. It is widely used in regression tasks where data quality and robustness to outliers are critical considerations.

III. Similarity Space

Similarity and dissimilarity functions are fundamental in several research areas, including information retrieval, machine learning, cluster analysis, and specific applications such as database searching and protein sequence comparisons. While dissimilarity functions are well-defined within metric spaces, similarity functions often lack a universally accepted definition, leading to potential ambiguities and inconsistencies.

In this study, we propose a structured framework for defining similarity spaces as counterparts to metric spaces. This framework provides a foundation for analyzing similarity functions with a clear mathematical perspective, setting the stage for future research and applications.

We start by investigating monotonically decreasing convex mappings of metric spaces. Although these mappings form the basis of our proposed similarity space, they do not preserve the original metric. Instead, we present an axiomatic system that defines the properties of this similarity space. Using this system, similarity functions can be precisely defined and examined. Additionally, a metric space can be derived by mapping back from the similarity space. Importantly, our research shows that certain measures, like the Jaccard index

and Tanimoto coefficient, fit well within the defined similarity space, a position they lacked as per [14].

While the theory of metric spaces has been established for over a century, the concept of similarity spaces is a more recent development [14]. Linking metric and similarity spaces is not straightforward. The former arises from spatial definitions, while the latter is derived from the comparison of shared and distinct attributes.

Definition III.1 (Similarity Space [14]–[18])**.** *A similarity on nonempty set X is a function $s\colon X \times X \to \mathbb{R}^+$ such that for all elements $x, y, z \in X$:*

(S1) $s(x, y) = s(y, x)$ (symmetry),
(S2) $s(x, z)+s(y, y) \geq s(x, y)+s(y, z)$ (triangle inequality),
(S3) $s(x, x) = s(x, y) = s(y, y) \iff x = y$ (identity of indiscernibles),
(S4) $s(x, y) \geq 0$ (non-negativity),
(S5) $s(x, y) \leq \min\{s(x, x), s(y, y)\}$ (bounded self-similarity).

A similarity space is an ordered pair (X, s) such that X is nonempty set and s is similarity on X.

There are several points to clarify. The term 'similarity metric' has been introduced earlier. When referring to it as a 'metric', it pertains to the context of a monotonously decreasing convex transformation of either a partial metric or a distance metric. In this article, we'll predominantly use the term 'similarity' to sidestep potential ambiguities. It's noteworthy to mention that based on our definition, an item can exhibit a positive self-similarity, $s(x, x) > 0$, and the self-similarities among different items can vary, i.e., $s(x, x) \neq s(y, y)$. However, if x is identical to y, $s(x, y)$ might not be zero. [14]

A basic example of similarity space is the ordered pair $(\mathbb{R}^+, s)$ defined as follows

$$x \cap y = \min\{x, y\} = \frac{x + y - |x - y|}{2} \tag{6}$$

for all x, y in $\mathbb{R}^+$. Other examples of similarity spaces that are interesting in terms of broad practical application such as Jaccard index, Tanimoto coefficient, Generalized Rozinek similarity, Levenshtein similarity, longest common subsequence may be found in [14]. The similarity space has many other applications, notably in fixed-point theory and in the exploration of the existence and uniqueness of solutions to differential equations [19].

IV. Methodology

Regression analysis, traditionally applied to quantitative data, is based on understanding relationships between variables. As data analysis techniques evolve, there arises a need to adapt these methods to a more qualitative nature of data, especially when the underlying relationship is based on similarity rather than magnitude. [20]

Traditional regression techniques, such as the least squares method, focus on minimizing the distance between data points and the model's predictions. Given a data set X and a model prediction y, this objective can be formally expressed as:

$$\arg\min_{y \in Y} \sum_{x \in X} d(x, y)^2 \tag{7}$$

The squared norm (or squared distance) is given by $\|x\|^2$ (or $[d(x, y)^2$). In certain contexts, especially in optimization problems and machine learning, using the squared distance can simplify computations due to its differentiable properties. However, it's crucial to understand that the squared norm does not, in general, define a metric, because it doesn't satisfy the triangle inequality.

Theorem IV.1 (Induced Elementary Metric)**.** *If $s(x, y)$ is a similarity on X, then the function $d^s\colon X \times X \to \mathbb{R}^+$ given by*

$$d^s(x, y) = s(x, x) + s(y, y) - 2s(x, y) \tag{8}$$

is induced elementary metric on X.

Proof. Consider $x, y \in X$. Then $d^s(x, y) = s(x, x)+s(y, y)-2s(x, y)$ is always non-negative by the bounded self-similarity (S5) because $s(x, y) \leq \min\{s(x, x), s(y, y)\}$ holds. Moreover, if $d^s(x, y) = d^s(y, x) = 0$ we get $x = y$ because $s(x, x) = s(x, y) = s(y, y)$. Furthermore, the triangular inequality holds

$$\begin{aligned} d^s(x, y) &= s(x, x) + s(y, y) - 2s(x, y) \\ &\leq s(x, x) + s(y, y) - 2[s(x, z) + s(y, z) - s(z, z)] \\ &= [s(x, x) + s(z, z) - 2s(x, z)] \\ &+ [s(y, y) + s(z, z) - 2s(y, z))] \\ &= d^s(x, z) + d^s(y, z). \end{aligned}$$

□

Considering our definition of distance in terms of similarity, we have $d(x, y)^2 = (s(x, x) + s(y, y) - 2s(x, y))^2$.

By substituting this expression, the objective becomes:

$$\arg\min_{y \in Y} \sum_{x \in X} [s(x, x) + s(y, y) - 2s(x, y)]^2 \tag{9}$$

In the context of optimization, finding the argument x which minimizes $f(x)$ is equivalently expressed as finding the argument that maximizes $-f(x)$. This relationship can be denoted as $\arg\min\ f(x) = \arg\max\ (-f(x))$.

In contrast, if our aim is to maximize similarity, the objective becomes centric to the mutual similarity $s(x, y)$. The objective to achieve this can be expressed as:

$$\arg\max_{y \in Y} \sum_{x \in X} s(x, y)^2 \tag{10}$$

Using the elementary similarity as an objective function, we obtain for simple linear regression model

$$\arg\max_{\theta_0, \theta_1} \sum_{x \in X} (y \cap \hat{y})^2 = \arg\max_{\theta_0, \theta_1} \sum_{x \in X} \min\{y, \hat{y}\}^2 \tag{11}$$

where the simple linear regression model has the form:

$$\hat{y_i} = \theta_0 + \theta_1 x_i + \epsilon_i. \tag{12}$$

Here:

- $\hat{y_i}$ is the predicted value of the dependent variable for the i^{th} observation.
- θ_0 is the y-intercept, representing the predicted value of y when $x = 0$.
- θ_1 is the slope of the regression line, signifying the change in y for a unit change in x.
- ϵ_i is the random error term for the i^{th} observation, capturing the unexplained variation in y.

In regression analysis, the primary goal is often to minimize the discrepancy between observed values and model predictions. However, an alternative approach is to frame this as a similarity maximization problem. When viewing the task through the lens of similarity, we aim to maximize the elementary similarity between observed and predicted values.

Given the data set X and the observed values y, the elementary similarity objective function for a simple linear regression model can be defined as:

$$J(\theta_0, \theta_1) = \sum_{i=1}^{n} (\min\{y_i, \hat{y_i}\})^2 \tag{13}$$

where $\hat{y_i}$ denotes the predicted value of the i^{th} observation. The goal is then to find parameters θ_0 and θ_1 that maximize this similarity:

$$\arg\max_{\theta_0, \theta_1} J(\theta_0, \theta_1). \tag{14}$$

The parameters θ_0 and θ_1 are estimated using the data in such a way that they maximize the elementary similarity between observed and predicted values.

To maximize the elementary similarity between observed and predicted values, we must take the derivatives with respect to θ_0 and θ_1 and set them equal to zero.

Given our objective function θ_0, θ_1 where $\hat{y_i} = \theta_0 + \theta_1 x_i$, we need to determine when this function reaches its maximum.

Taking the partial derivative with respect to θ_0:

$$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} 2(\min\{y_i, \theta_0 + \theta_1 x_i\})(1) = 0 \tag{15}$$

Similarly, the partial derivative with respect to θ_1 is:

$$\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{n} 2(\min\{y_i, \theta_0 + \theta_1 x_i\})(x_i) = 0 \tag{16}$$

For the maximum similarity, we set the above partial derivatives equal to zero. This gives us a system of nonlinear equations, which can be solved using iterative methods or optimization techniques.

It's important to note that, unlike the traditional least squares method, the elementary similarity approach doesn't yield a closed-form solution for θ_0 and θ_1 due to the non-linear nature of the objective function. Advanced optimization techniques, such as gradient ascent (since we are maximizing) or specialized algorithms, are necessary to determine the optimal values of θ_0 and θ_1 that maximize our similarity measure.

The objective function $J(\theta_0, \theta_1)$ is formulated based on the elementary similarity between the observed and predicted values. It aims to maximize the squared similarity.

For optimization purposes, and especially when using techniques like gradient ascent or descent, derivatives come into play. These derivatives, or gradients, indicate the rate of change of the objective function with respect to the parameters.

To obtain the partial derivatives of the objective function J with respect to θ_0 and θ_1, we will differentiate the function. When we take the derivative of the minimum value, we must consider two potential cases for each data point:

1) When y_i is less than or equal to $\hat{y_i}$
2) When y_i is greater than $\hat{y_i}$

However, the form of our derivatives suggests we only need to consider the values for which y_i is less than or equal to $\hat{y_i}$.

Let's explore the partial derivatives more rigorously:

1) *Partial Derivative with respect to θ_0:*
   Given:
   $$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} 2(\min\{y_i, \theta_0 + \theta_1 x_i\})(1)$$
   When $y_i \leq \hat{y_i}$, this simplifies to:
   $$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} 2y_i$$
   However, when $y_i > \hat{y_i}$, the contribution to the sum is 0.
   For maximization, this should be set to zero:
   $$\sum_{i=1}^{n} 2y_i = 0$$
   But this equation is not informative in its current form.
2) *Partial Derivative with respect to θ_1:*
   Given:
   $$\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{n} 2(\min\{y_i, \theta_0 + \theta_1 x_i\})(x_i)$$
   When $y_i \leq \hat{y_i}$, this becomes:
   $$\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{n} 2y_i x_i$$
   However, when $y_i > \hat{y_i}$, the contribution is again 0.
   For maximization:
   $$\sum_{i=1}^{n} 2y_i x_i = 0$$
   But this equation, too, is not immediately informative for estimating θ_1.

It is essential to note that the provided equations represent the conditions under which the objective function is maximized with respect to θ_0 and θ_1. However, this does not mean that solutions for θ_0 and θ_1 can be directly extracted from them.

The challenge is that, because of the use of the "min" function, the equations are non-linear and may not have an analytical solution. Therefore, numerical optimization methods, like the Newton-Raphson method, will likely be the best approach to find the values of θ_0 and θ_1 that maximize J.

This similarity-based approach for simple linear regression offers an alternative perspective to the conventional least squares method. While it introduces complexities in terms of parameter estimation, the focus on maximizing similarity between actual and predicted values might provide unique insights and potential applications in specific domains where such a perspective is beneficial.

By redefining our objective in terms of similarity rather than the traditional least squares method, we present a novel perspective on regression analysis. This approach could offer unique insights and potentially lead to alternative regression techniques that emphasize maximizing agreement or overlap between actual and predicted values.

A. Nonlinear Optimization: The Newton-Raphson Method

Given the non-linear nature of our objective function, traditional linear regression techniques do not provide a direct solution. One of the most prominent techniques to find the maximum of a function like ours is the Newton-Raphson method.

The Newton-Raphson method is an iterative procedure used to find successively better approximations to the roots (or zeros) of a real-valued function. The method can be generalized for multidimensional problems, making it apt for our regression task.

Given a current estimate $\boldsymbol{\theta}$, the update equation for the Newton-Raphson method is given by:

$$\boldsymbol{\theta}^{(new)} = \boldsymbol{\theta}^{(old)} - [\boldsymbol{H}^{(old)}]^{-1}\boldsymbol{g}^{(old)} \tag{17}$$

where $\boldsymbol{g}$ is the gradient vector (first derivative) and $\boldsymbol{H}$ is the Hessian matrix (second derivative).

For our objective function:

$$J(\theta_0, \theta_1) = \sum_{i=1}^{n} (\min\{y_i, \hat{y}_i\})^2 \tag{18}$$

The gradient $\boldsymbol{g}$ will contain our earlier derived partial derivatives with respect to θ_0 and θ_1. The Hessian matrix $\boldsymbol{H}$ will be a 2x2 matrix, with entries being the second partial derivatives of J with respect to θ_0 and θ_1.

The Newton-Raphson iterations continue until the change in $\boldsymbol{\theta}$ between successive steps is below a predetermined small threshold or until a maximum number of iterations is reached.

While the Newton-Raphson method provides rapid convergence, it's worth noting that the method might converge to a local maximum, saddle point, or even diverge if not initialized properly. Therefore, careful initialization and possibly multiple starting points might be necessary to ensure convergence to the global maximum.

By redefining our objective in terms of similarity, we present a novel approach to regression analysis. While the computational complexity of estimating parameters increases due to the non-linearity of the objective function, methods like the Newton-Raphson provide efficient tools for such tasks. Emphasizing similarity over the traditional error minimization could open avenues to alternative regression techniques that prioritize agreement or overlap between observed and predicted values in specific application areas.

B. Objective Formulation in Similarity Spaces

In the realm of similarity-based optimization, we often seek to maximize mutual similarities while keeping into account self-similarities which can act as regularizations. This paradigm gives rise to an objective function of the following form:

$$\arg\max_{y \in Y} \sum_{x \in X} [s(x, y) - \alpha s(x, x) - \theta s(y, y)] \tag{19}$$

Each term in this objective function has a distinct interpretation:

- $s(x, y)$: This term represents the mutual similarity between x and y. Intuitively, within the optimization process, this term pushes for the selection of a y that exhibits high similarity to each element x in the set X.
- $-\alpha s(x, x)$: This is a regularization term based on the self-similarity of x. The coefficient α scales its impact. An elevated value of α emphasizes scenarios where x's self-similarity is minimal, steering the optimization away from entities that are too self-reliant or idiosyncratic.
- $-\theta s(y, y)$: Analogous to the preceding term, this component regularizes based on the self-similarity of y. A large θ value prioritizes generalized solutions for y, discouraging selections that are too narrowly tailored or self-focused.

C. Balancing Mutual and Self-Similarity

The coefficients α and θ in the objective function offer a nuanced control over the balance between mutual similarity and self-similarity. While mutual similarity pushes for closeness between entities, self-similarity acts as a counterbalance, ensuring that solutions don't gravitate towards overly specific or unique representations. The optimal balance, dictated by the values of α and θ, would typically be grounded in domain expertise or determined empirically based on specific application goals.

D. Objective Function Using Dot Product

A vector space equipped with an inner product $\langle \cdot, \cdot \rangle$ is termed a Hilbert space if it is complete with respect to the norm induced by that inner product, i.e., $\|x\| = \sqrt{\langle x, x \rangle}$. Under certain conditions if we measure similarity based on the Lebesgue measure we can see

$$\langle x, y \rangle = \mu(x \cap y) = s(x, y)$$

where μ denotes the Lebesgue measure. This measure satisfies certain axioms as given in Definition III.1.

Given a dataset with observed values y and input values x, our goal is to find parameters that maximize the projection of

Algorithm 1 Optimization using Dot Product Maximization with Regularization

Input: Data points $(x_1, y_1), \ldots, (x_n, y_n)$, Regularization coefficient λ, Learning rate α, Tolerance ϵ, Maximum iterations $max\_iterations$
Output: Parameters θ_0, θ_1
Initialize θ_0, θ_1 to some starting values
Initialize $prev\_cost$ to a large value
$iteration \leftarrow 0$
while $iteration < max\_iterations$ **do**
  $J \leftarrow \sum_{i=1}^{n} y_i(\theta_0 + \theta_1 x_i) - \lambda \sum_{i=1}^{n} (\theta_0 + \theta_1 x_i)^2$
  if $|J - prev\_cost| < \epsilon$ **then break**
  $prev\_cost \leftarrow J$
  $gradient\_0 \leftarrow \sum_{i=1}^{n} y_i - 2\lambda(\theta_0 + \theta_1 x_i)$
  $gradient\_1 \leftarrow \sum_{i=1}^{n} y_i x_i - 2\lambda x_i(\theta_0 + \theta_1 x_i)$
  $\theta_0 \leftarrow \theta_0 + \alpha gradient\_0$
  $\theta_1 \leftarrow \theta_1 + \alpha gradient\_1$
  $iteration \leftarrow iteration + 1$
return θ_0, θ_1

y onto its predicted values $\hat{y} = \theta_0 + \theta_1 x$. This can be quantified using the dot product of these vectors:

$$\hat{\theta_0}, \hat{\theta_1} = \arg\max_{\theta_0, \theta_1} \left\{ \sum_{i=1}^{n} y_i \hat{y_i} - \lambda \sum_{i=1}^{n} {\hat{y_i}}^2 \right\} \tag{20}$$

$$= \arg\max_{\theta_0, \theta_1} \left\{ \sum_{i=1}^{n} y_i(\theta_0 + \theta_1 x_i) - \lambda \sum_{i=1}^{n} (\theta_0 + \theta_1 x_i)^2 \right\} \tag{21}$$

where λ is a regularization coefficient that controls the balance between maximizing the projection and minimizing the magnitude of the parameters.

E. Deriving the Gradients

To find the values of θ_0 and θ_1 that maximize our objective function, we compute the gradient and set it to zero.

For θ_0:

$$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} y_i - 2\lambda \sum_{i=1}^{n} (\theta_0 + \theta_1 x_i) \tag{22}$$

$$= n\overline{y} - 2\lambda(n\overline{\theta_0} + \overline{\theta_1 x}) \tag{23}$$

where $\overline{y}$ is the mean of the observed values, and $\overline{\theta_0}$ and $\overline{\theta_1 x}$ are the means of the predicted values.

For θ_1:

$$\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{n} y_i x_i - 2\lambda \sum_{i=1}^{n} x_i(\theta_0 + \theta_1 x_i) \tag{24}$$

$$= \sum_{i=1}^{n} y_i x_i - 2\lambda(\sum_{i=1}^{n} x_i \theta_0 + \sum_{i=1}^{n} \theta_1 x_i^2) \tag{25}$$

Setting these gradients to zero yields the conditions for the optimal parameters θ_0 and θ_1. The gradient ascent updates for the parameters at each iteration are given by:

$$\theta_0 \leftarrow \theta_0 + \alpha \frac{\partial J}{\partial \theta_0} \tag{26}$$

$$\theta_1 \leftarrow \theta_1 + \alpha \frac{\partial J}{\partial \theta_1} \tag{27}$$

where α is the learning rate, which controls the size of the steps taken in the direction of the gradient. The optimization continues until a maximum number of iterations is reached or the change in the objective function value between successive iterations is less than a specified tolerance ϵ. The detailed steps are presented in Algorithm 1.

1) Robustness to Outliers: Consider two datasets: one without outliers D and one with an outlier D'. Let the objective values for these datasets be represented as $J(D)$ and $J(D')$.

Without the regularization term, the difference in objectives due to an outlier is:

$$\Delta J_{\text{no-reg}} = J(D') - J(D) \tag{28}$$

$$= \sum_{i \in D'} y_i(\theta_0 + \theta_1 x_i) - \sum_{i \in D} y_i(\theta_0 + \theta_1 x_i) \tag{29}$$

Given the influence of an outlier, this difference could be significantly large.

However, with the regularization term:

$$\Delta J_{\text{reg}} = J(D') - J(D) \tag{30}$$

$$= \left[\sum_{i \in D'} y_i(\theta_0 + \theta_1 x_i) - \lambda \sum_{i \in D'} (\theta_0 + \theta_1 x_i)^2 \right] \tag{31}$$

$$- \left[\sum_{i \in D} y_i(\theta_0 + \theta_1 x_i) - \lambda \sum_{i \in D} (\theta_0 + \theta_1 x_i)^2 \right] \tag{32}$$

The regularization term, $-\lambda \sum_i (\theta_0 + \theta_1 x_i)^2$, penalizes large values of the parameters, thereby limiting the magnitude of predictions.

To examine the impact of an outlier on this regularized objective, consider a single outlier point $(x_{\text{out}}, y_{\text{out}})$ such that y_{out} is much larger than other values.

The contribution of this outlier to the objective is:

$$\Delta J_{\text{outlier}} = y_{\text{out}}(\theta_0 + \theta_1 x_{\text{out}}) - \lambda(\theta_0 + \theta_1 x_{\text{out}})^2 \tag{33}$$

While the data term $y_{\text{out}}(\theta_0 + \theta_1 x_{\text{out}})$ tries to fit the outlier closely, the regularization term $-\lambda(\theta_0 + \theta_1 x_{\text{out}})^2$ prevents the model parameters from adapting too much to the outlier. Thus, a suitable choice of λ can limit the outlier's influence, ensuring robustness.

2) Convergence to Linear Regression Model: To demonstrate that the regularized objective function converges towards a linear regression solution, we need to examine the properties of the objective function and the gradient ascent update rules. Given that we are maximizing our function, we need it to be concave. This function is a dot product minus a sum of squared parameters (a regularization term). Under certain conditions, this function can be concave, especially if the dot product term dominates the behavior. The negative sum of squares (regularization term) is concave. For concave functions, gradient

ascent with a suitable learning rate guarantees convergence to a global maximum. And the potential concavity of our objective function, these updates will iteratively increase the value of J until it converges to its global maximum, fitting the model to the given data.

The regularization term penalizes extreme values of $\hat{\theta}$. This ensures that the algorithm doesn't diverge, promoting stability and convergence. Moreover, the regularization can be viewed as a form of penalty that keeps the parameter values bounded, ensuring that the gradient ascent does not lead to unbounded growth of the parameters.

F. Potential Applications and Implications

The presented objective provides a flexible framework for tasks that revolve around similarity considerations. By judiciously setting the parameters, one can tailor the objective to a variety of use cases, from information retrieval to cluster analysis. Further research could delve into understanding the impacts of different similarity metrics on this optimization paradigm.

V. Experiments

This article explores a novel approach to recommending pairs of articles to readers based on the similarity of their content. The central hypothesis is that the likelihood of a reader clicking on a recommended article (B) can be modeled using the similarity between the primary article (A) and the recommended one.

We construct a regression model where:

- The independent variable x is the normalized word-based similarity measure between all pairs of words from sentences A and B.
- The dependent variable Y is the click-through rate (CTR) for the recommended article B.

The normalized edit similarity between two sentences, $s(x, y)$, could be defined as:

$$s(x, y) = \frac{|x| + |y| - d(x, y)}{|x| + |y| + d(x, y)}$$

where $|x|$ and $|y|$ denote the lengths of sentences x and y respectively, and $d(x, y)$ is the Levenshtein distance between them.

For two sentences A and B, their average normalized word similarity can be calculated as [14]:

$$\text{Average Word Similarity} = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} s(a_i, b_j)$$

where n is the number of words in sentence A, m is the number of words in sentence B, a_i is the ith word in sentence A, and b_j is the jth word in sentence B.

VI. Dataset

We present a demonstrative example that illustrates the application of ordinary least squares (OLS) in a similarity space. This is achieved by maximizing the dot product, complemented with regularization. The resulting regression model is given by:

$$\text{CTR} = \theta_0 + \theta_1 \times \text{Average Word Similarity}$$

For the data presented in Table VI and as outlined in Algorithm 1, the estimated regression coefficients are:

$$\text{CTR} = 43.91 + 61.89 \times \text{Average Word Similarity}.$$

VII. Discussion

This methodology offers a unique take on regression by focusing on maximizing the similarity between observed and predicted values. The regularization term ensures model stability and avoids overfitting. Further work is needed to compare its efficacy to traditional regression techniques and to explore other potential applications.

The proposed model offers a unique approach to understanding user engagement based on content similarity. However, several other factors, such as article length, topic, and writing style, could influence the CTR. Future work could incorporate these aspects for a more comprehensive recommendation system.

Given the potential concavity of our objective function and the properties of gradient ascent, our approach guarantees convergence to a solution that fits the model to the data while also being robust due to regularization.

VIII. Conclusion

By redefining our objective in terms of similarity, we present a novel approach to regression analysis. While the computational complexity of estimating parameters increases due to the non-linearity of the objective function, methods like the Newton-Raphson provide efficient tools for such tasks.

Emphasizing similarity over the traditional error minimization opens the door to alternative regression techniques that prioritize agreement or overlap between observed and predicted values. One practical application of this can be seen in content recommendation systems. By treating each article as a vector of words or concepts and computing the similarity between them, our regression-based similarity model can suggest pairs of articles that align closely in terms of content. This ensures that readers are not only provided with content tailored

TABLE I. Dataset of Sentence Pairs and their CTR

Sentence	SM	CTR
A:Machine learning methodologies are advancing rapidly. B:AI has greatly impacted technological evolution.	0.75	90
A:Computers are essential for modern work. B:Programming is a fundamental skill in the digital age.	0.65	85
A:Weather prediction relies on robust algorithms. B:Good data analytics improves forecast accuracy.	0.70	88
A:Birds are adapted for aerial locomotion. B:Fish have evolved to swim efficiently in water.	0.3	60
A:Digital marketing strategies are diversifying. B:Online advertisements drive significant business revenue.	0.55	80

** SM = Similarity Measure*

to their interests but also with additional related articles, enhancing their reading experience and engagement.

Furthermore, this approach could revolutionize the way recommendation engines work, especially in the digital publishing realm. Instead of merely suggesting individual articles based on users' past reads or popular trends, platforms could present pairs or clusters of articles that delve into similar themes or topics. This not only facilitates deeper immersion into a particular subject area but also encourages users to spend more time on the platform, benefiting both the user and the content provider.

Moving forward, more extensive studies can be conducted to validate the efficiency and effectiveness of our proposed method in real-world recommendation systems, further bridging the gap between traditional regression analysis and modern-day practical applications.

ACKNOWLEDGMENT

It was supported by the Erasmus+ project: Project number: 2022-1-SK01-KA220-HED-000089149, Project title: Including EVERyone in GREEN Data Analysis (EVERGREEN) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Slovak Academic Association for International Cooperation (SAAIC). Neither the European Union nor SAAIC can be held responsible for them.

REFERENCES

[1] S. Sagiroglu and D. Sinanc, "Big data: A review," *International Journal of Computer Sciences and Engineering*, vol. 1, no. 5, pp. 28–42, 2013. [Online]. Available: http://www.ijcse.net/docs/IJCSE13-01-05-070.pdf

[2] R. Feldman and J. Sanger, "The text mining handbook: Advanced approaches in analyzing unstructured data," *Cambridge University Press*, 2007.

[3] S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, 1986.

[4] A.-M. Legendre, "Nouvelles méthodes pour la détermination des orbites des comètes," *Imprimerie Impériale*, 1805.

[5] C. F. Gauss, "Theoria combinationis observationum erroribus minimis obnoxiae," *Werke*, vol. 5, pp. 1–52, 1821.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119, 2013.

[7] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, 2003, pp. 73–78.

[8] M. Zhang, J. Tang, and X. Zhang, "Text data processing and analysis: A brief survey," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 2065–2066.

[9] A. N. Tikhonov and V. Y. Arsenin, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics*, vol. 4, no. 3, pp. 1035–1038, 1963.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[14] O. Rozinek and J. Mareš, "The duality of similarity and metric spaces," *Applied Science*, vol. 11, p. 1910, 2021.

[15] B. Ma and K. Zhang, "The similarity metric and the distance metric," *Proceedings of the 6th Atlantic Symposium on Computational Biology and GenomeInformatics*, p. 1239–1242, 2005.

[16] S. Almazel, Q. H. Ansari, and M. A. Khamsi, *Topics in Fixed Point Theory*. Berlin: Springer-Verlag, 2014.

[17] C. C. H. Elzinga and M. M. Studer, "Normalization of distance and similarity in sequence analysis," *Sequence Analysis and Related Methods (LaCOSA II)*, p. 445, 2016.

[18] E. Alhajjar and C. Lefèvre, "On the similarity metric," *Mathematica Militaris*, vol. 24, no. 1, p. 4, 2019.

[19] B. M. Rozinek O, "Theorems for boyd–wong contraction mappings on similarity spaces," *Mathematics*, vol. 11, 2023.

[20] M. B. L. Badri and D. St-Yves, "Supporting predictive change impact analysis: A control call graph based technique," *12th Asia-Pacific Software Engineering Conference*, p. 167–175, 2005.