ISSN 2231-3478 (Print)
JUSPS-B Vol. 37(4), 29-39 (2025). Periodicity-Monthly
Section B
(Online) ISSN 2319-8052

JOURNAL OF ULTRA SCIENTIST OF PHYSICAL SCIENCES
An International Open Free Access Peer Reviewed Research Journal of Physical Sciences
website:- www.ultrascientist.org
Estd. 1989

# Applications for AI and ML in the analysis of unstructured data across various sectors

**FARHA KHAN[1*], PRATIMA OJHA[1] and GHIZAL F. ANSARI[2]**

[1]Department of Mathematics, Madhyanchal Professional University, Bhopal-462001 (INDIA)
[2]Department of Physics, Madhyanchal Professional University, Bhopal-462001 (INDIA)
*Corresponding Author Email :farhakhan50705@gmail.com

http://dx.doi.org/10.22147/jusps-B/370401

**Acceptance Date 06th September 2025 Online Publication date 12th September 2025**

**Abstract**

The main focus of the topic is the process of transforming a collection of unstructured text documents into structured information based on mathematical and statistical principles. To begin, we'll look at document models via the lens of the Bernoulli method, where the existence or absence of tokens the fundamental building elements of documents forms the foundation. Multinomial document model is the center of attention in an additional issue. It resembles the Bernoulli model in many ways, but instead of using the presence flag, it uses the frequentist approach, which considers how often the tokens appear in the text. To get latent topical structure across text sources and to fine-tune with the use of machine learning, we move onto researching unsupervised topic modeling strategies in the following challenge. Finally, using unstructured data analysis, we provide a model for predicting users' moods and actions on social media. A model that may capture user behavior and mood on social media is the Behavior Dirichlet Probability Model (BDPM).

**Key words :** Latent Semantic Indexing, social media, unstructured text, Machine learning, AI

This is an open access article under the CC BY-NC-SA license (https://creativecommons.org/licenses/by-nc-sa/4.0)

## Introduction

Machine learning (ML) is a branch of AI that studies how computers may "learn" new skills by applying predefined sets of instructions called algorithms to existing datasets. Two main approaches to machine learning exist: supervised learning, which makes use of labels to teach algorithms how to predict output from input, and unsupervised learning, which makes use of labels to teach algorithms how to construct and organize output data. The purpose of this review was not to investigate the many kinds of algorithms that may be used for prediction. However, optimistic bootstrapping (a validation approach) might lead to exaggerated accuracy rates when used with particular algorithms, as shown in a recent work by Jacobucci et al. We examined publications for these algorithm/validation technique combinations to ensure consistency and deleted three of them to reduce bias.

The AUROC, or area under the receiver operating characteristic curve, is a popular metric for evaluating algorithm performance. In classification research, the accuracy results are informed by a confusion matrix that is constructed using the model's performance in categorizing the dataset's true positive, false negative, and false negative outcomes. Psychologists and medical professionals use area under the curve (AUC) as a general performance measure to compare case and control participants to determine how well diagnostic tests work. An improved model's ability to foretell a result, such suicidal behavior, is proportional to its area under the curve (AUC). In a range from 0 to 1, AUC values below 0.5 indicate a level of chance, >0.5 indicates a level of chance, >0.6 indicates poor predictive ability, >0.7 indicates fair predictive ability, >0.8 indicates strong predictive ability, and >0.9 represents exceptional predictive ability.

***Literature Review :***

Oza *et.al*.[2] Unstructured data stores a wealth of information, but software struggles to make use of it because it lacks basic, identifiable organization. Since it lacks a predetermined model or structure, relational databases are not a good fit for storing it. Machine- or human-generated unstructured data is saved in its original format with all relevant information. Modern tools like machine learning and artificial intelligence are required to process unstructured data. Metadata, Data Visualization, Image Analysis, and Natural Language Processing (NLP) are all important aspects to think about while studying unstructured data.

Ritika *et.al*.[1] We acknowledge in this study that companies aiming to preserve data consistency and quality have a tremendous difficulty in today's digital world due to the explosion of data from varied sources. Effective data management is typically impeded by the limitations of traditional Master Data Management (MDM) systems with respect to breadth, flexibility, and efficiency. More and more businesses are turning to cloud-based MDM solutions that use AI and ML to combat these problems. A new age of more efficient corporate information management is upon us, and these solutions provide a way to improve data correctness, consistency, and comprehensiveness.

Mahadevkar *et.al*. (2024) Unstructured data continues to be a major problem in today's industrial environment, causing millions of dollars in yearly financial losses for many different industries. It is possible to significantly improve operational efficiency by making good use of this data. Artificial intelligence (AI) technologies might provide a better alternative to traditional techniques of data extraction, which have their limits. A thorough assessment of AI-driven methods for information

extraction from unstructured text is clearly lacking in the academic literature. In order to find, evaluate, and discuss potential future research directions in the area of unstructured document information extraction, this literature review is arranged in a methodical manner. When confronted with complicated document structures often seen in real-world contexts, like medical records, existing extraction approaches that rely on static patterns or rules frequently fail. There is a lack of high-quality, purpose-built datasets accessible to the public at this time. This highlights the critical requirement of creating fresh datasets that mirror real-world complicated challenges. According to the research, there is potential for AI-based algorithms to autonomously extract data from various unstructured documents, including printed and handwritten text. However, working with different document layouts has its own share of challenges. This paper proposes a framework for automated information extraction from unstructured texts using hybrid AI-based techniques. The methodology assumes processing a high-quality dataset. Furthermore, it stresses the need of scholars and companies working together to tackle the many difficulties of unstructured data processing.

Bundidth Dangsawang *et.al*. (2024) Unauthorized persons sell goods and services via social media, causing governments to lose taxes and customs charges. In order to identify such infractions in social media's unstructured data, this research suggests a methodology named SHIELD. In three stages, we will gather 2,373,570 records of marketed items from social media sites like Facebook and Twitter. The first step in text categorization is the collection of labeling keywords. The findings are categorized into three groups: Inspect for things that cannot be recognized from the text and need additional examination; Green Line for non-commercial goods; and Red Line for smuggled, unpaid duty, forbidden, and restricted goods. The second and third phases utilize a combination of keywords and three algorithms—Logistic Regression (LR), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM)—to identify smugglers in unstructured social media data and categorize illicit goods that have been imported. With an average F1 score of 90.55% and an accuracy of 99.44%, the LSTM method clearly came out on top in every test. Machine learning and natural language processing have the ability to identify unlawful actions and promote economic security using algorithms and methods like LR, GRU, and LSTM.

Olumide Johnson Ikumapayi *et.al*. (2025) Innovations in machine learning and artificial intelligence (AI) are changing the forensic accounting game by providing cutting-edge tools for detecting and preventing fraud in real time. Even when they work, traditional forensic accounting approaches can't always handle the amount and complexity of digital economy financial transactions. A dynamic and adaptable approach to recognizing and reducing fraudulent actions may be provided by AI-powered systems, especially those that use machine learning techniques. By using these technologies, forensic accountants may do real-time analyses of massive datasets, unearth previously unseen patterns, and spot irregularities that might indicate financial misdeeds like embezzlement, money laundering, or financial statement fraud. This study delves into the use of artificial intelligence (AI) in forensic accounting, specifically examining how supervised and unsupervised learning as well as natural language processing (NLP) approaches may significantly improve fraud detection skills. To make accurate predictions about future fraud episodes, supervised learning models are trained on previous data. These models include decision trees and support vector machines. On the other hand, clustering and anomaly detection are examples of unsupervised learning methods that may find financial data anomalies without categorizing them beforehand, which makes them useful for finding new fraud

schemes. By examining emails and financial reports, which include unstructured data, natural language processing (NLP) enhances these models even more, allowing them to spot misleading language or hidden hazards. Concerns about algorithmic bias, data privacy, and the risk of becoming too reliant on automated systems are among the ethical considerations and difficulties brought up by the use of AI in forensic accounting. This article examines case studies and real-world applications to demonstrate how AI may revolutionize forensic accounting processes. It aims to make these methods more efficient, accurate, and proactive, which will strengthen financial integrity and corporate governance.

*Concept of Latent Semantic Indexing (Lsi) :*

According to Deerwester, Dumais, and Harshman (1990), LSI is a method that involves transferring user-supplied text into a latent-or hidden-dimensional semantic space. What makes this method special is that it can create a latent semantic space where texts with various sets of keywords or words may still be connected (with high cosine similarity). The created latent semantic space is lower dimensional than the original text's starting space.

*Identification Of Topics Using Latent Semantic Indexing (Lsi) :*

To find these subjects, we need to use latent semantic indexing (LSI), which finds latent themes in texts and lets us capture the notion of entities that define them; these entities or words connect and vary among themselves according to some main latent textual dimension. Because of this, we may examine the diversity within a corpus of documents within the context of a "semantic field". Essentially, semantic fields are descriptions of collections of related concepts that provide insight on the common understanding of a certain concept or view. The semantic field perceives these associated concepts as "hypernyms" because they are assumed to have shared semantic properties. On the surface, hypernyms seem to be a kind of <is-a> connection that exists in a semantic context.

**An Svd-Based Example of Lsi**

To demonstrate LSI with SVD, Table 1 provides text data.

Table 1. Sample Data

| Documents | Text |
|---|---|
| C1 | Today's lunch has apples, oranges and mangoes for fruits |
| C2 | Alpha courier company delivered mangoes and bananas to my apartment at 12 o clock. |
| C3 | Oranges and bananas are fresh from Alpha. |
| C4 | Bananas and Apples were ordered but bananas were never delivered. |
| C5 | Alpha makes deliveries to my apartment mostly at 12 o clock. |
| M1 | The printing of all the related diagrams. |
| M2 | The intersection water of paths in diagrams. |
| M3 | Water falls diagrams give a clear view of the process. |
| M4 | The water harvesting process used by our company is best. |

By plotting the aforementioned information in a Term Document matrix, we get (4.2).

| terms | C1 | C2 | C3 | C4 | C5 | M1 | M2 | M3 | M4 |
|---|---|---|---|---|---|---|---|---|---|
| apples → | [1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0] |
| oranges → | [1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0] |
| mangoes → | [1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0] |
| alpha → | [0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0] |
| bananas → | [0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0] |
| apartment → | [0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0] |
| clock → | [0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0] |
| fresh → | [0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0] |
| company → | [0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0] |
| diagrams → | [0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0] |
| water → | [0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1] |
| process → | [0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1] |

Documents C and M are distinguished by the fact that C deals with fruits and M with processes. (4.2) is amenable to matrix representation as

$$X = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

If we follow the steps in (4.1), we can now break down X into its component matrices. Results from breaking down X are (4.3), (4.4), and (4.5).

$$T_0 = \begin{bmatrix} 0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\ 0.20 & -0.07 & 0.14 & -0.55 & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\ 0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\ 0.40 & 0.06 & -0.34 & 0.10 & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\ 0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\ 0.30 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\ 0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\ 0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\ 0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & -0.68 & 0.23 \\ 0.03 & 0.45 & 0.14 & -0.01 & -0.30 & 0.28 & 0.34 & 0.68 & 0.18 \end{bmatrix}$$

$$S_0 = \begin{bmatrix} 3.34 & & & & & & & & \\ & 2.54 & & & & & & & \\ & & 2.35 & & & & & & \\ & & & 1.64 & & & & & \\ & & & & 1.50 & & & & \\ & & & & & 1.31 & & & \\ & & & & & & 0.85 & & \\ & & & & & & & 0.56 & \\ & & & & & & & & 0.36 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} 0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\ 0.61 & 0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & 0.24 \\ 0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & 0.02 \\ 0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & -0.08 \\ 0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & -0.26 \\ 0.00 & 0.19 & 0.10 & 0.02 & 0.39 & -0.30 & -0.34 & 0.45 & -0.62 \\ 0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & 0.02 \\ 0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0.00 & 0.25 & 0.45 & 0.52 \\ 0.08 & 0.53 & 0.08 & -0.03 & -0.60 & 0.36 & 0.04 & -0.07 & -0.45 \end{bmatrix}$$

Therefore, in order to approximate X' (a two-dimensional function) from X,

$$X' = \begin{bmatrix} 0.22 & -0.11 \\ 0.20 & -0.07 \\ 0.24 & 0.04 \\ 0.40 & 0.06 \\ 0.64 & -0.17 \\ 0.27 & 0.11 \\ 0.27 & 0.11 \\ 0.30 & -0.14 \\ 0.21 & 0.27 \\ 0.01 & 0.49 \\ 0.04 & 0.62 \\ 0.03 & 0.45 \end{bmatrix} \times \begin{bmatrix} 3.34 & \\ & 2.54 \end{bmatrix} \times \begin{bmatrix} 0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\ 0.61 & 0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & 0.24 \end{bmatrix}$$

$$= \begin{bmatrix} 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ 0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\ -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\ -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{bmatrix}$$

*Identification Of Subnets Using K-Means Algorithm :*

Produced semantic space is highly dependent on the social linguistics structural form. There are distinct semantic structures for each social classification or subnet in the whole dataset. By identifying these structures independently for each subnet, the thematic material may be fine-tuned. Here, we'll investigate the possibility of using k-means to partition the dataset into smaller subnets.

*K-Means Clustering :*

In order to classify 'n' realizations of the relevant random variable into k clusters that are distinct from one another yet similar within themselves, the k-means method is used.

The architype of a cluster is the set of observations that are all assigned to it based on its nearest mean value. The result is a set of Voronoi cells formed from the partitions produced by applying the algorithm to the space obtained by the observations.

With respect to an existing collection of insights *(x1, x2, ..., xn)*, After mapping each realization to the real line, which is a vector with d dimensions, the algorithm tries to split the n realizations into *k* ($\leq n$) classes $S = \{S1, S2, ..., Sk\}$ make an effort to reduce the variation within each cluster as much as possible. The stated goal is to determine what:

$$\arg_s \min \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \arg_s \min \sum_{i=1}^{k} | S_i | \operatorname{Var} S_i$$

where $\mu i$ denotes the meaning of realizations that are contained in *Si* minimization of cluster variance is denoted by argmin. An alternative formulation of the issue might be the minimizing of the squared deviations of all pairs of points inside the same *Si*, that is

$$\arg_s \min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} \|x-y\|^2$$

You may get the same kind of information from the following:

$$\sum_{x \in S_i} \| x - \mu_i \|^2 = \sum_{x \neq y \in S_i} (x - \mu_i)(\mu_i - y)$$

This is equivalent, as a result of the law of variances, to maximizing the sum of squared deviations between points in different clusters, or inter-cluster variances, because the overall sum of squares or variation remains constant.

It should be mentioned that the algorithm's convergence to the global maximum or minimum is not guaranteed. The selection of starting clusters has a significant impact on the result. However, because of how fast it is, the technique is usually performed many times with different starting parameters. However, even when considering only a few dimensions, the worst-case performance may be comparable to exponential time convergence.

The basic premise is based on spherical clusters that are spaced apart in a way that converges close to the cluster's center. Since associating to the closest center would thus be error-proof, it is implicitly expected that the clusters are of equal size.

*Determining Optimal K For K-Means :*

Even if *k* isn't always the best choice, the *k*-means method is still somewhat innocent in that

it will divide the input dataset into *k* divisions. Finding the best value of *k* is thus an essential first step before running the algorithm.

The elbow technique is one approach that checks the value of *k*. Elbow is based on the idea of running k-means on the input data over a range of k and then determining the sum of squared errors (SSE) for each *k*. Finally, an arm-shaped line graph representing the calculated SSE for all possible values of k would be interesting to see. You should choose *k* equal to the location or value of the elbow. It seems to reason that as *k* increases, the SSE would fall until it reaches zero. The goal of this approach is to find a value of *k* at which the rate of reduction of SSE becomes insignificant.

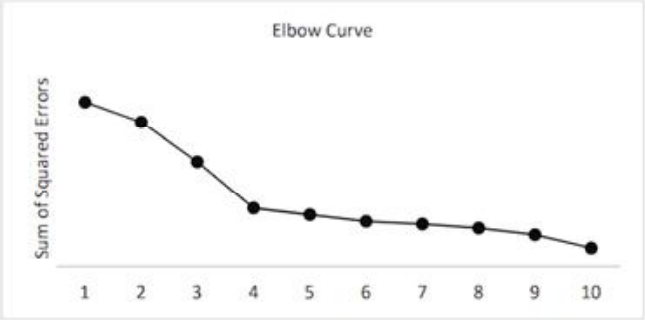
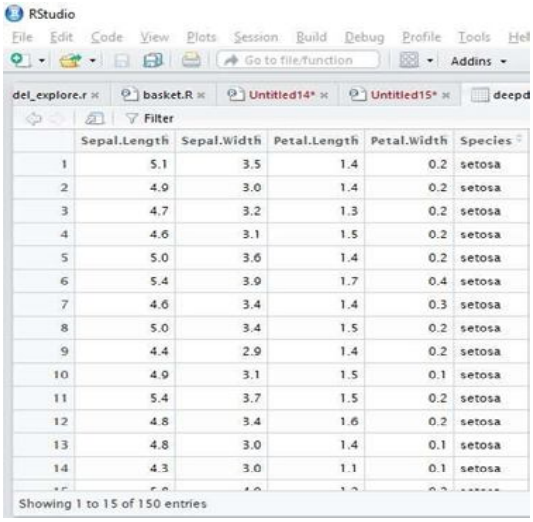



Fig 1 Elbow Curve for Determining Optimal k

As shown in Figure 1, the ideal value of k is 4.

**Implementation of K-Means In R**

Here is the procedure to follow in order to use the k-means algorithm in R: The 'iris' dataset, which is a R package, is used. To begin, we import the data into the R environment. Figure 2 shows a sample of the non-scale data.




| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 | setosa |

**Fig 2 Data for k-means**

To ensure that the clustering variables have similar variances of 0 and 1, we apply scale transformations to them.

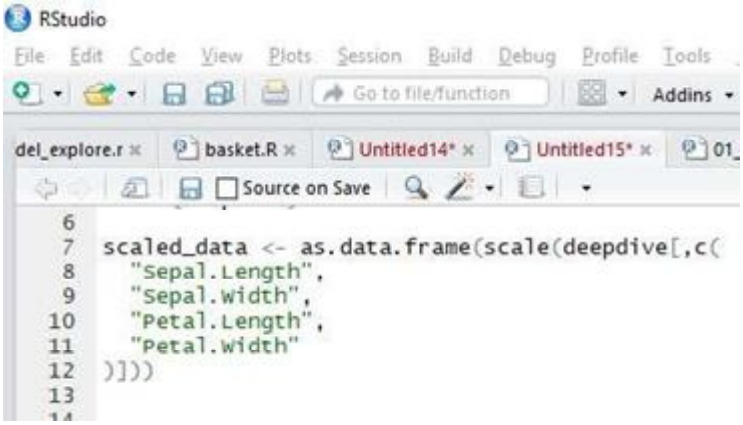
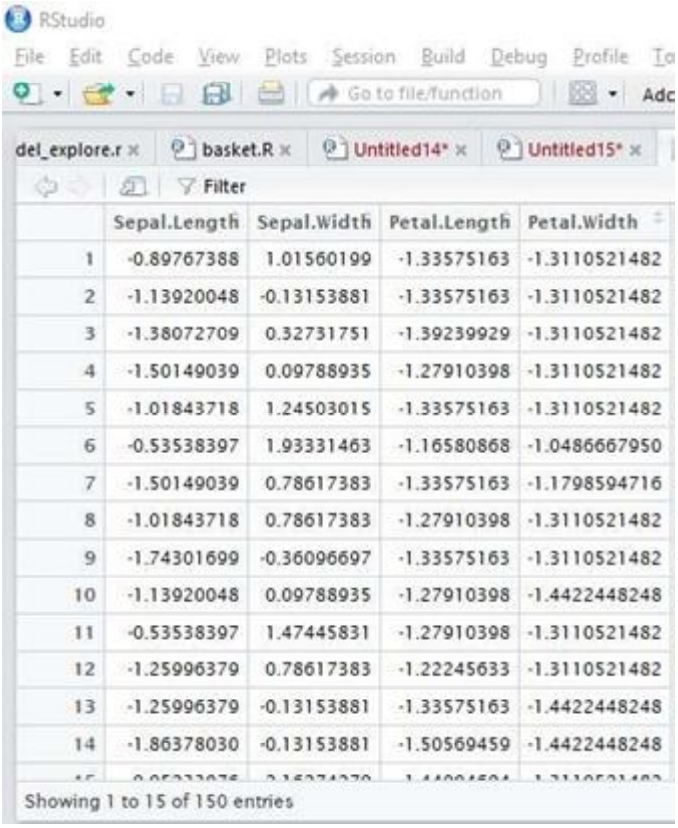



**Fig 3 Scaling of Clustering Variables**

Figure 4 shows a snapshot of the data that has been scaled.




| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | -0.89767388 | 1.01560199 | -1.33575163 | -1.3110521482 |
| 2 | -1.13920048 | -0.13153881 | -1.33575163 | -1.3110521482 |
| 3 | -1.38072709 | 0.32731751 | -1.39239929 | -1.3110521482 |
| 4 | -1.50149039 | 0.09788935 | -1.27910398 | -1.3110521482 |
| 5 | -1.01843718 | 1.24503015 | -1.33575163 | -1.3110521482 |
| 6 | -0.53538397 | 1.93331463 | -1.16580868 | -1.0486667950 |
| 7 | -1.50149039 | 0.78617383 | -1.33575163 | -1.1798594716 |
| 8 | -1.01843718 | 0.78617383 | -1.27910398 | -1.3110521482 |
| 9 | -1.74301699 | -0.36096697 | -1.33575163 | -1.3110521482 |
| 10 | -1.13920048 | 0.09788935 | -1.27910398 | -1.4422448248 |
| 11 | -0.53538397 | 1.47445831 | -1.27910398 | -1.3110521482 |
| 12 | -1.25996379 | 0.78617383 | -1.22245633 | -1.3110521482 |
| 13 | -1.25996379 | -0.13153881 | -1.33575163 | -1.4422448248 |
| 14 | -1.86378030 | -0.13153881 | -1.50569459 | -1.4422448248 |

Fig 4 Scaled Data for Subnet

After that, we'll build an elbow curve to find the best value for k.

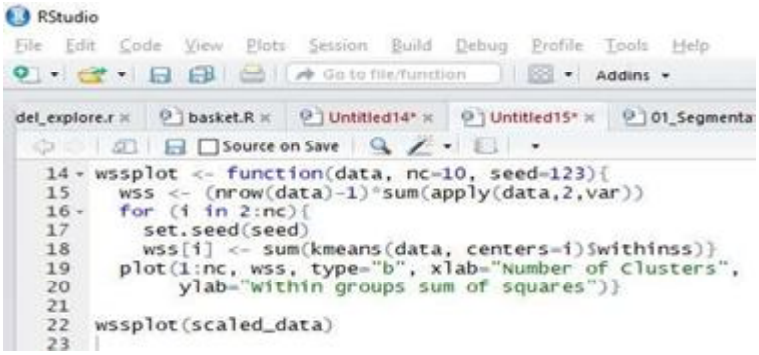



**Fig 5 Code to Plot Elbow Curve**

Figure 6 displays the elbow curve that was produced.

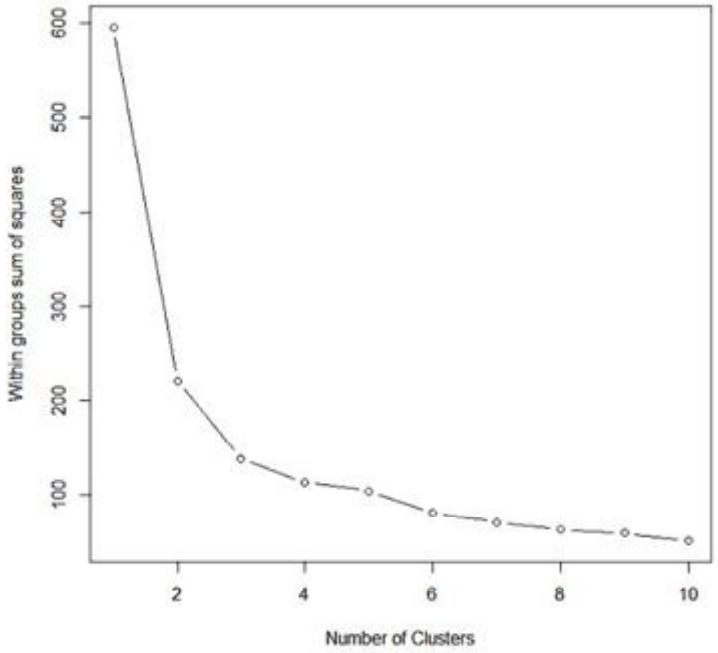



Fig. 7 Elbow Curve on Scaled Data

A decent match for k is shown in Figure 7 with a number of clusters equal to 3. Therefore, in the subnet k-means implementation, we choose k = 3.

## Conclusion

To begin, the idea and need of a method that can capture semantic links have been brought

up, with the work of Deerwester (1990) on Latent Semantic Indexing (LSI) serving as an inspiration. The fundamental mathematical ideas that control LSI are laid out in great detail. Steps in the suggested approach include identifying subnets, tokenizing N-grams, determining the document term matrix using inverse document frequency, topic modeling on each subnet to find the top N-grams, and lastly, determining the polarity and relevance of N-grams within the topics. By carrying out each step in turn, we are able to extract the positive and negative polarity of the main drivers from the textual data.

**References**

1. Ritika, "Investigating the Role of Artificial Intelligence and Machine Learning in Cloud-Based Master Data Management,' International Journal of Research Publication and Reviews, Vol 4, no 11, pp 1424-1430 November 2023 (2023).
2. Oza, Ravi & Punjani, Dipti & Domadiya, Dr. Dipti. ANALYSIS OF UNSTRUCTURED DATA USING ARTIFICIAL INTELLIGENCE. 2320-2882 (2023).
3. Okeleke, Patrick & Ajiga, Daniel & Folorunsho, Samuel & Ezeigweneme, Chinedu. Predictive analytics for market trends using AI: A study in consumer behavior. 10.53430/ijeru.2024.7.1.0032 (2024).
4. Sharma, Himanshu. (2024). The Role of Artificial Intelligence and Machine Learning in Strengthening Cloud Security: A Comprehensive Review and Analysis. IJARCCE. 13. 10.17148/IJARCCE.2024.13808.
5. Paramesha, Mallikarjuna & Rane, Nitin & Rane, Jayesh., Big Data Analytics, Artificial Intelligence, Machine Learning, Internet of Things, and Blockchain for Enhanced Business Intelligence. 1. 110-133. 10.5281/zenodo.12827323 (2024).
6. Unalp, Aynur., AI-Driven Predictive Analytics: Shaping the Future of Strategic Decision Making. 10.13140/RG.2.2.22309.41447 (2024).
7. Yuandi Wu, "Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring," Expert Systems with Applications Volume 255, Part C, 1 December 2024, 124678 (2024).
8. Daniel, Samuel., AI and Data-Driven Strategies: Transforming Competitive Intelligence in Market Analysis. 10.13140/RG.2.2.12664.15362 (2024).
9. Asimiyu, Zainab., Leveraging Artificial Intelligence in Data Analytics: Strategic Insights and Innovations (2024).
10. Shah, Sayed & Soomro, Assadullah & Rahu, Mushtaque & Hussain, Kashif & Karim, Sarang. Innovative Machine Learning Solutions: Investigating Algorithms and Applications. Journal of Applied Engineering & Technology (JAET). 8. 32-51. 10.55447/jaet.08.01.146 (2024).