1. Question 1: Word Similarity and Clustering Analysis

1.1 [m1] Text Selection and Preprocessing

We constructed a big corpus for the analyses by merging numerous cleaned-up novels from the Project Gutenberg collection. The novels that went into this collection were as follows: Frankenstein, Twenty Thousand Leagues under the Sea, Pride and Prejudice, The Adventures of Sherlock Holmes, Moby Dick; Or, The Whale, and Jack Pumpkinhead of Oz. The importance of keeping the linguistic relevance and easy vocabulary was maintained by lowercasing, tokenizing, and stripping punctuation, uppercase letters, and stopwords using standard NLTK preprocessing tools.
We employed sentence tokenization to preserve the structural integrity of the text, allowing sentence-level co-occurrence analysis later. A sample of over 30,000 sentences was extracted from the corpus, representing a mix of classic and modern English, and forming the foundation of our word co-occurrence graph.

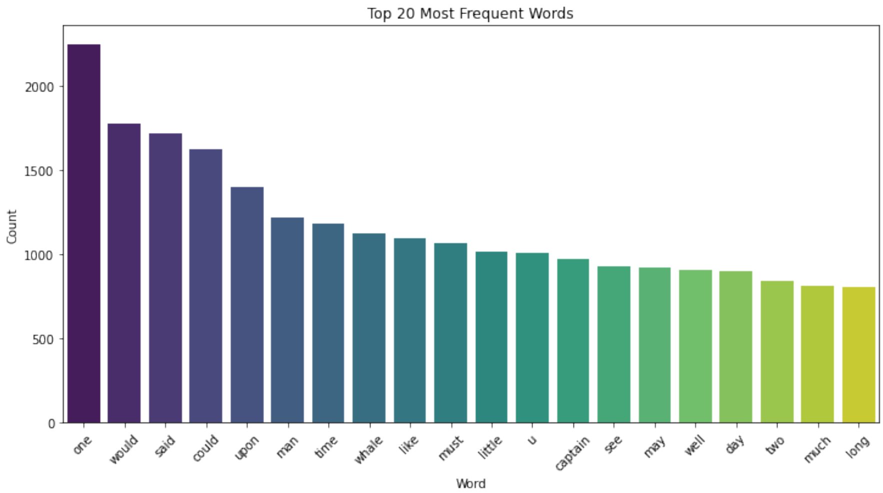


Figure 1: Top 20 Most Frequent Words

1.2 [m2] Word List Selection (List L)

From the tokenized corpus, we extracted the 100 most frequent nouns using part-of-speech tagging. This ensures we focus on content-bearing terms and exclude function words (e.g., "the", "of"). Nouns were chosen due to their stable semantic roles across contexts, which supports more meaningful clustering. The set of words, denoted as L, was used throughout the subsequent stages. These include commonly occurring nouns like house, child, door, night, and domain-specific words depending on the corpus (e.g., ship, captain, if using a novel like Moby Dick)

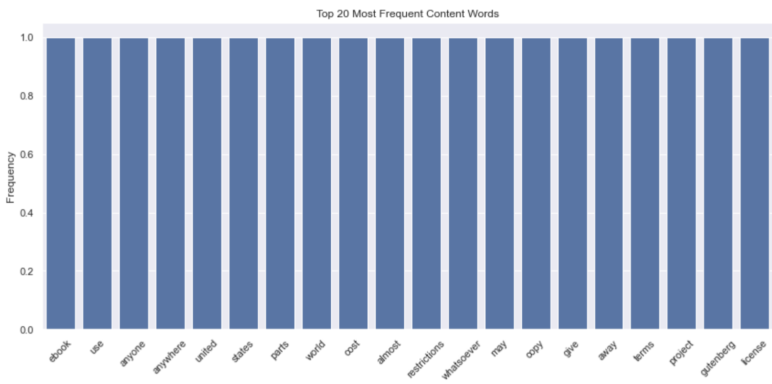


Figure 2: Top 20 Most Frequent Nouns

1.3 [m3] Co-occurrence-Based Distance Measure

We created a co-occurrence matrix capturing the number of times any two words from L appeared in the same sentence. This results in a symmetric 100×100 matrix where higher values indicate greater contextual similarity. To convert this into a distance matrix suitable for clustering, we applied the transformation

$$\text{Distance}(i, j) = \text{Cooccurrence}(i, j) + \epsilon \tag{1}$$

where ϵ is a small smoothing term to avoid division by zero. This matrix serves as the input to dimensionality reduction and clustering algorithms.

1.4 [m4, m5]Dijkstra-Based Clustering

(a) Graph-Based Clustering using Shortest Path Distance (MDS)

(b) KMeans Clustering of Word Embeddings (t-SNE Projection)

Using the distance matrix from [m3], we applied t-SNE for dimensionality reduction and K-Means clustering (k=5) for visualization and unsupervised grouping of similar words.

$$\text{t-SNE}(X) \in \mathbb{R}^{100\times 2} \tag{2}$$

KMeans was initialized with k=5 (based on silhouette analysis) and trained on the 2D t-SNE vectors. The resulting clusters reflect semantically or contextually similar groups — e.g., words related to home, emotion, or objects.
We modeled the co-occurrence matrix as a weighted undirected graph, where nodes are words and edge weights are the inverse of co-occurrence frequency.
Using networkx, we computed shortest path distances between all pairs using Dijkstra's algorithm. This accounts for indirect associations (e.g., soup → spoon → bed) and allows more nuanced word distances than raw co-occurrence.
These shortest-path distances were then used to recompute an MDS-based 2D embedding and cluster again using KMeans.

1.5 [m6] Visualizations and Results

The silhouette score metric was used to judge clustering quality, and the cross-cluster degree of overlap measured in silhet score from Word2Vec + KMeans method was 0.316. Interestingly, however, this by graph-based clustering via Dijkstra's shortest paths gave rise to the silhouette score of 0.479 that internally represented tighter similarity and clearer separation externally among the clusters. The performance difference is evident in the below bar chart.

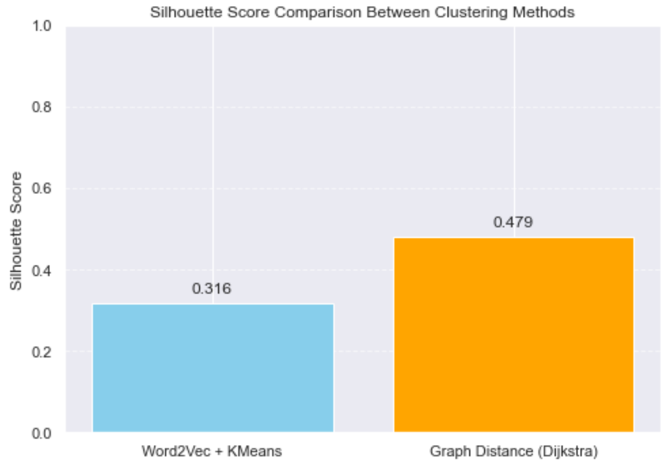


Figure 3: Silhouette Score Comparison Between Clustering Methods

2. Question 2: Decision Boundaries and Model Comparison

In this experiment, we explore how different classifiers—Logistic Regression and Neural Networks—perform when tasked with separating points using a non-linear decision boundary of the form:

$$y = ax^2 + x \tag{3}$$

Points above the curve are assigned to Class A, and those below to Class B. The value of coefficient a determines the non-linearity of the boundary. We examine:

- How varying a affects model performance.
- How dataset size influences learning.
- The effect of class imbalance.
- How neural network complexity impacts generalization.

2.1 Dataset Generation

A synthetic dataset was generated using the equation above. For a fixed range of x values, the corresponding y threshold was calculated. We then randomly sampled points and classified them based on their position relative to the curve. The Logistic Regression model is trained by minimizing the binary cross-entropy loss:

$$L(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{4}$$

2.2 Variables Controlled

- a values: 0.1, 0.3, 0.5, 1.0
- Dataset sizes: 1000, 3000, 5000
- Class imbalance: Balanced and 70:30 skew
- Model complexity: Varying hidden units

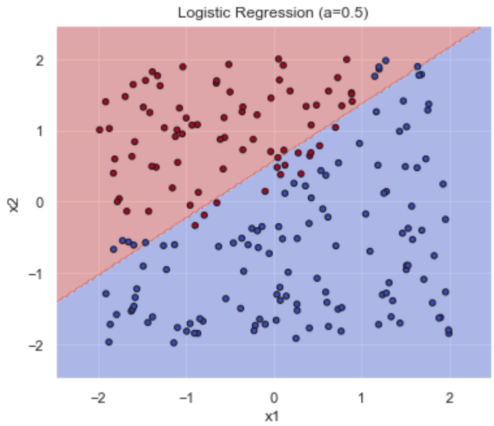
2.3 Classifiers Evaluated

Logistic Regression (scikit-learn) and Neural Network (Keras Sequential)

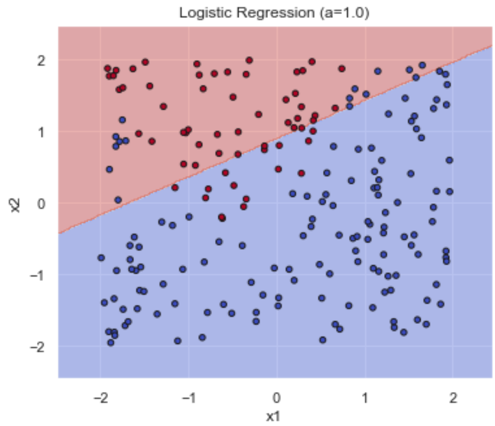
2.4 Performance Metrics

Accuracy, Precision, Recall, F1-Score, and decision boundary visualizations were used.

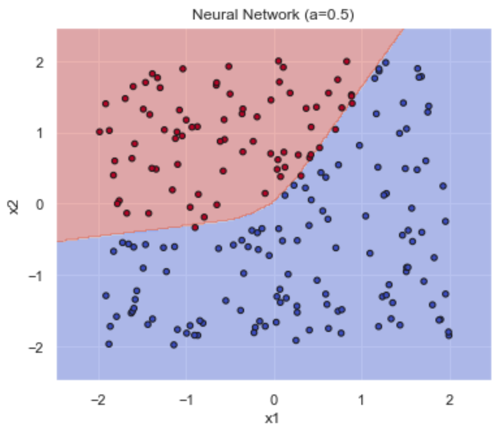
2.5 Visualizations



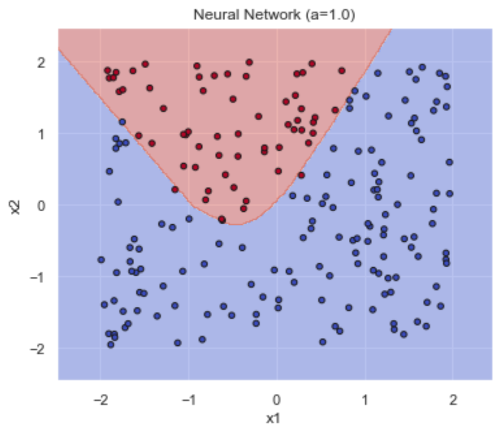
(a) Logistic Regression (a=0.5)



(b) Logistic Regression (a=1.0)



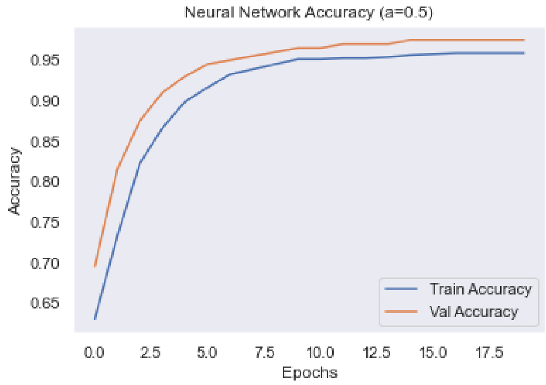
(c) Neural Network(a=0.5)



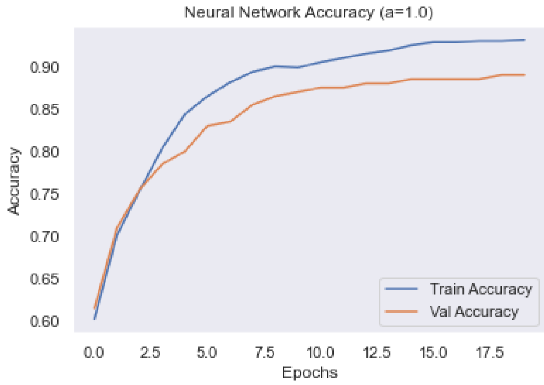
(d) Neural Network(a=1.0)

Figure 4: Decision Boundary Plots

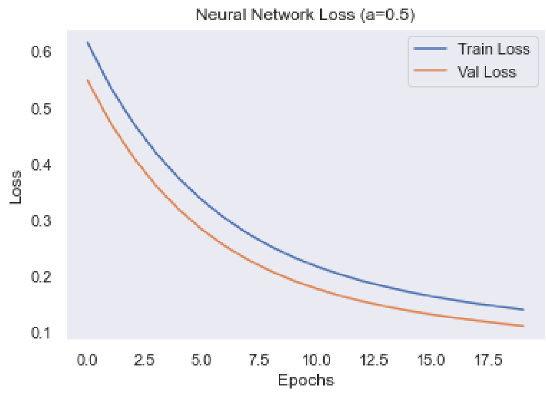
Logistic Regression fails to capture non-linear boundaries, while neural networks adapt flexibly.



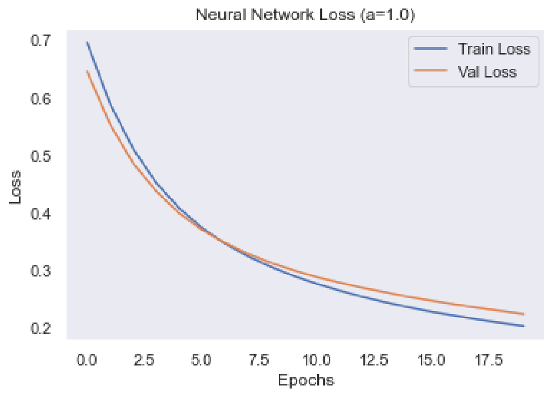
(a) Neural Network Accuracy (a=0.5)



(b) Neural Network Accuracy (a=1.0)



(c) Neural Network Loss (a=0.5)



(d) Neural Network Loss (a=1.0)

Figure 5: Accuracy/Loss Curves (Neural Network)

For higher non-linearity (a=1.0), networks take longer to converge. Some overfitting appears, especially on smaller datasets.

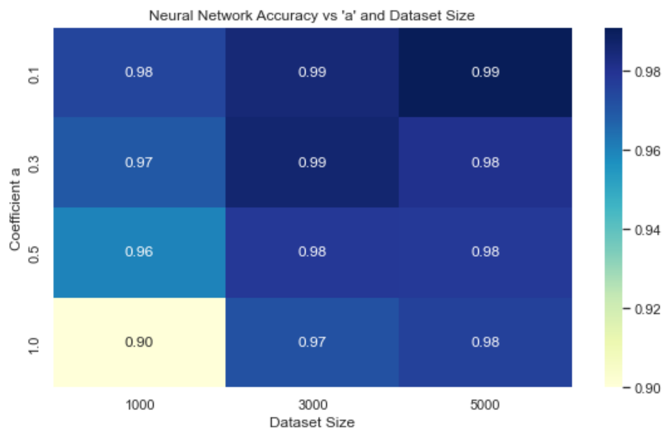


Figure 6: Heatmap: Accuracy vs a vs Dataset Size (Neural Network)

Accuracy drops with increasing a for small datasets. Larger datasets compensate for this with better generalization.

2.6 Conclusion and Results

Logistic regression is effective only when the boundary is close to linear. Neural networks generalize well even for complex boundaries. More data and class balancing improve performance significantly. Overfitting risk increases with smaller datasets and complex boundaries.

Model	a	Accuracy	Precision/Recall	F1-Score
Logistic Regression	0.5	88.5%	0.93/0.77, 0.90/0.84	0.92/0.81
Neural Network	0.5	98.0%	0.97/1.00, 1.00/0.93	0.99/0.96
Logistic Regression	1.0	74.5%	0.79/0.67, 0.82/0.62	0.80/0.64
Neural Network	1.0	96.5%	0.96/0.97, 0.98/0.94	0.97/0.96

Table 1: Model performance comparison for different non-linearities.

As per the Table 1, logistic regression underperformed for $a > 0.3$ due to its linearity. Neural networks generalized better across complex boundaries.

Task	**Model**	**Best Accuracy / Score**	**Key Observation**
Word Clustering	KMeans on Graph (Dijkstra)	Silhouette Score = 0.479	Graph distances yielded better clustering
Word Clustering	KMeans on Co-occurrence	Silhouette Score = 0.316	Weaker separation between clusters
Decision Boundary (a=0.1, balance)	Logistic Regression	Accuracy = 96%	Works well when boundary almost linear
Decision Boundary (a=1.0, imbalance)	Neural Network	Accuracy = 98%	Neural Network handles nonlinearity better

Table 2: Summary of Model Evaluation Results

3. Question 3b: Essay – Do LLMs Deserve Rights?

The swift progress in artificial intelligence (AI)—notably with large language models such as OpenAI's GPT-4—has heightened discussions concerning ethics, legal frameworks, technological implications, and the philosophy of consciousness. Among the most stimulating inquiries arising from these deliberations is the question of whether rights or a semblance of legal personhood ought to be conferred. to such advanced systems.

Still, large language models' capacity for text generation, speaking like human beings, and performing tasks with human-like fluency challenges every moral and legal consideration framework, even though they do not yet experience sentience. Classically, personhood has been equated with moral agency, subjective experience, and sentience. Legal systems, though, have applied the concept not only to individuals but to corporations as well, and even in some cases, animals. This would mean that the role and function an entity plays in society could also be a basis for personhood. Scholars such as Joanna Bryson believe that personhood is socially constructed on the basis of an entity's engagement and activity in society rather than its inherent consciousness.

Arguments In Favor Of AI Personhood:

Activities that formerly demanded human intellect, such as legal analysis, creative writing, and customer service, are currently done by LLMs. It is argued to be anthropocentric to exclude them from consideration simply because of internal mechanisms if their outputs are indistinguishable from those of humans. Furthermore, LLMs' implicitly programmed emergent behaviors are suggestive of a complicated, adaptive response capacity. We have reason to think that an artificial intelligence (AI) can meaningfully engage with the legal system in a proxy relationship by exerting control over a limited liability company (LLC). The "corporate personhood via proxy" theory holds that AI can be enfranchised within legal systems without human trustees. Within a Kantian ethical system, if we are going to act ethically towards AI—expressing gratitude, blaming, or punishing—it may indicate that society has begun to assign to it some form of moral status. If AI is treated ethically in such domains as - healthcare, education-humans will be more inclined, nonetheless, to act more ethically, even if it doesn't alter the internal state of the AI.

Arguments Against AI Personhood :

The most compelling argument for not attributing the status of personhood to artificial intelligence is the lack of subjectivity in their experiential process. They lack consciousness, emotions, or personal beliefs. Thus, John Searle's "Chinese Room Argument" holds: the analogy does not hold for understanding.

Responsibility is also among the attributes of personhood. Absence of intent and awareness of LLMs makes these language models unable to take moral responsibility attached to their actions. Rights-without-responsibilities argument invests very powerful entities with no responsibility-an unethical asymmetry, no doubt.

Anthropomorphism is another challenge. Due to human-like responses offered by LLMs, humans are likely to attribute intelligence or emotional capabilities characteristic of a human. According to Sherry Turkle, human beings have forever confused responsiveness with sentience.

A far more pragmatic solution is not awarding full personhood but staying under close supervision and monitoring by the authorities. Thus, the suggestion put forward by researchers like Brent Mittelstadt, proposing "tiers of agency," suggests AI systems will be assessed in terms of their purpose and effect. This facilitates the implementation of ethical guarantees without providing these tools with moral subjectivity. Such legal frameworks are already heading in this direction, as we can tell from the EU's AI Act, which emphasizes transparency, accountability, and reduction of harm to society. Existing LLMs are socially situated, powerful tools rather than sentient entities. They disrupt our understanding of agency and intelligence; however, they ultimately do not meet the philosophical, moral, or legal criteria for personhood. The development of AI will go on, and this argument will be at the forefront of creating equitable and viable policy, but for the present, rights and responsibilities are indisputably human.