American Journal of Information Science and Technology
2026, Vol. 10, No. 1, pp. 44–53
https://doi.org/10.11648/j.ajist.20261001.16

SciencePG
Science Publishing Group

Research Article

Unsupervised Semantic Discovery in Customer Feedback: A Technical and Economic Analysis of Label-free Clustering

Partha Majumdar*

Department of Computer Science, Swiss School of Business and Management (SSBM), Geneva, Switzerland

Abstract

This report offers a comprehensive technical and economic analysis of unsupervised semantic clustering as a cost-effective alternative to supervised machine learning for extracting insights from unstructured customer feedback. The main argument is that the high costs and long delays associated with manual data labelling—the basis of supervised models—create a significant bottleneck for businesses leveraging Voice of Customer (VoC) data. An economic analysis shows that labelling a moderate dataset can cost tens of thousands of dollars and take months, potentially leaving insights outdated. This "labelling tax" is not a one-time cost but a recurring expense due to concept drift, increasing the financial burden. As a solution, the analysis evaluates an unsupervised pipeline that transforms raw text into geometric representations to uncover latent semantic structures without human labels. The process involves multiple stages: first, text is vectorised using TF-IDF, amplifying sentiment-rich terms. Next, the high-dimensional sparse matrix is compressed with Truncated SVD, reducing dimensionality and grouping synonyms into latent concepts. Finally, K-Means partitions the data into clusters. Empirical validation on the IMDb movie review dataset shows the pipeline's effectiveness; rigorous preprocessing, including removing named entities to avoid topic bias, enabled the algorithm to identify two clusters that, when visualised with t-SNE, aligned strongly with positive and negative sentiment labels. While the accuracy (~70%) is lower than that of supervised models (>95%), the report argues that this performance is adequate for many strategic uses, such as trend analysis and customer segmentation. The approach's advantages—scalability, domain independence, and the discovery of emergent "unknown unknowns"—position unsupervised clustering as a high-ROI, scalable tool for organisations seeking to quickly and affordably unlock their text data.

Keywords

Unsupervised Semantic Clustering, Cost of Data Annotation, Sentiment Analysis, Cost-accuracy Trade-off, Voice of Customer (VoC) Data, AI, ML

1. Introduction: The Unstructured Data Paradigm

In the contemporary digital economy, the volume of unstructured textual data generated by consumer interactions has reached unprecedented levels. It is estimated that more than 80% of enterprise data is in unstructured formats, primarily

*Correspondence: Partha Majumdar (partha.majumdar@hotmail.com)

Received: 28 January 2026; **Accepted:** 21 February 2026; **Published:** 9 March 2026

text, including customer support tickets, product reviews, social media discourse, and open-ended survey responses. This data represents a reservoir of latent strategic value, offering granular insights into customer sentiment, product performance, and emerging market trends. However, the operational reality of extracting intelligence from this data is constrained by a significant bottleneck: the dependency of traditional machine learning frameworks on supervised training.

Supervised learning, particularly in Natural Language Processing (NLP), has long been the gold standard for sentiment analysis and text classification. State-of-the-art models, such as those based on Transformer architectures (e.g., BERT, RoBERTa), have achieved remarkable accuracy, often exceeding 90% on standard datasets like IMDb movie reviews. [12] Yet these models are predicated on the availability of large-scale, high-quality labelled datasets—ground-truth annotations provided by human experts. The creation of such datasets is a resource-intensive process that entails substantial costs in time, capital, and human labour. [1].

This report presents a rigorous evaluation of an alternative paradigm: Unsupervised Semantic Clustering. By analysing the experimental study "Finding Hidden Customer Sentiments Without Labels" [9] and a comprehensive review of the current literature and economic data, we test the hypothesis that meaningful customer segmentation can be achieved without the prohibitive costs of manual labelling. The analysis demonstrates that by leveraging classical vector space models, combined with dimensionality reduction and geometric clustering, organisations can recover latent sentiment structures that closely align with human-perceived categories, thereby unlocking the value of Voice of Customer (VoC) data at a fraction of the traditional cost.

The contribution of this study is therefore not the introduction of a novel clustering algorithm, but rather an integrative, decision-oriented evaluation that combines established NLP techniques with a structured economic analysis of annotation costs, latency, and operational sustainability.

2. The Economic Imperative: Quantifying the Supervision Bottleneck

To fully appreciate the utility of unsupervised clustering, one must first conduct a forensic accounting of the costs associated with the supervised alternative. The assertion that unsupervised learning saves "enormous time and expense" is not merely anecdotal; it is supported by a robust analysis of the data annotation market and labour dynamics.

2.1. The Financial Architecture of Data Annotation

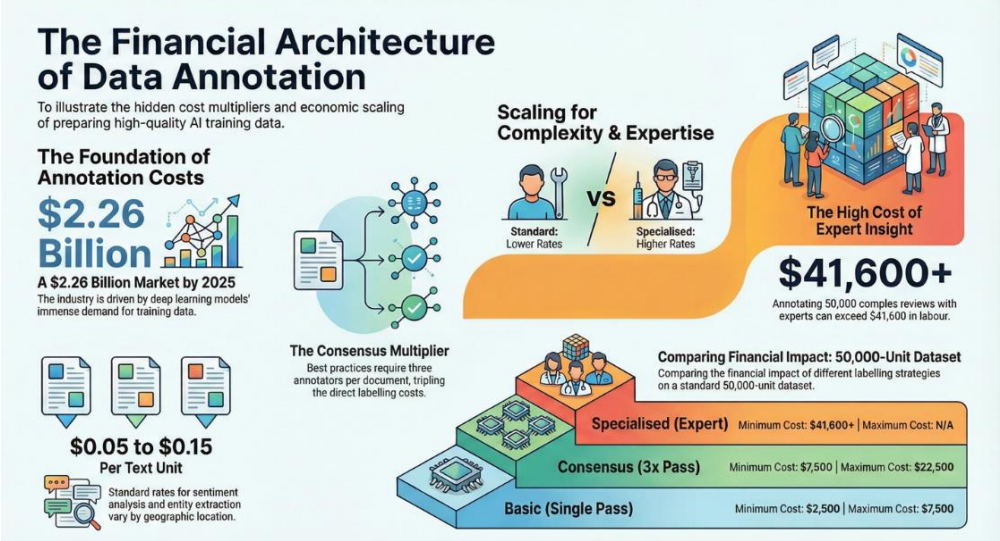

Figure 1. *The hidden economics of data annotation—how scale, expertise, and consensus multiply the true cost of AI training data.*

Data annotation is a burgeoning industry, projected to grow from $1.7 billion in 2024 to over $2.26 billion by 2025. [8] This growth is driven by deep learning models' voracious appetite for training data. However, for an individual enterprise seeking to build a custom sentiment model, the unit economics are daunting.

The cost of manual text labelling is influenced by task complexity, the required domain expertise, and annotators' geographic location. For standard sentiment analysis and entity extraction, industry rates typically range between $0.05 and $0.15 per text unit. [8] While this may appear negligible in isolation, the scale of modern datasets acts as a multiplier.

Consider the IMDb Large Movie Review Dataset, which

contains 50,000 reviews. [9] A basic calculation of direct labelling costs reveals the financial scale:

1) Scenario A (Crowdsourced/Offshore): At the lower bound of $0.05 per label, a single pass over 50,000 reviews costs $2,500.
2) Scenario B (Managed Service/Expert): At the upper bound of $0.15 per label, the cost rises to $7,500.

However, these figures represent a naive baseline. Best practices in data science indicate that a single label per document is insufficient to ensure ground truth, particularly for subjective tasks such as sentiment analysis, where human ambiguity is high. [16] To establish Inter-Annotator Agreement (IAA) and ensure high-quality training data, it is standard to have each document reviewed by at least three independent annotators. This creates a consensus mechanism but triples the direct cost:

1) Consensus Labelling Cost: $7,500 to $22,500 for a single static dataset.

Furthermore, specialised domains—such as legal, medical, or technical finance—require annotators with subject matter expertise. In these cases, costs do not scale linearly but exponentially. Hourly rates for specialised data annotation can range from $20 to $50+ per hour, depending on the requisite skill level. If an expert can annotate 60 complex reviews per hour, the labour cost alone for 50,000 reviews approaches $41,600 (at $50/hr), excluding platform fees and project management overhead.

2.2. The Temporal Latency of Manual Labelling

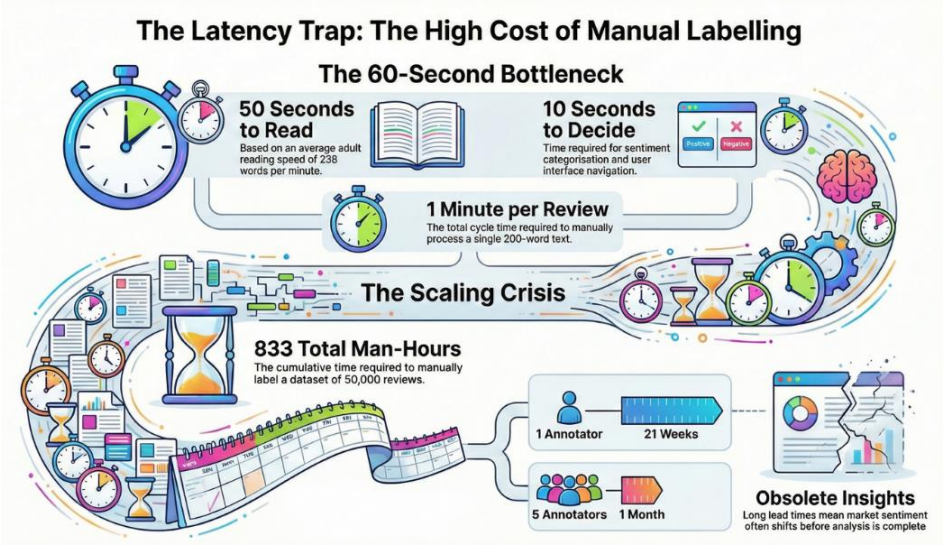

Figure 2. *Manual labelling's latency trap—how minutes per task compound into months of delay and stale insights at scale.*

Beyond financial capital, time is a critical resource in competitive markets. The latency introduced by manual labelling creates a "time-to-insight" gap that can render analysis obsolete before it is even completed.

To model this, we must look at human reading speeds and cognitive processing times. Research indicates that the average silent reading rate for adults in English is approximately 238 words per minute (wpm) for non-fiction text. [3] The average IMDb movie review, while variable, often exceeds 200 words, with many reviewers writing extensive critiques.

Assuming a conservative average review length of 200 words:

1) Reading Time: ~50 seconds per review.
2) Decision & Annotation Time: ~10 seconds (selecting the sentiment category, navigating the UI).
3) Total Cycle Time: 60 seconds (1 minute) per review.

For a dataset of 50,000 reviews:

$$\text{Total Man-Hours} = \frac{50{,}000\ reviews * 1\ minute/review}{60\ minutes\ /hour} \approx 833 \text{ hours}$$

This equates to nearly 21 weeks of full-time work for a single human annotator working a standard 40-hour week. Even with a team of five annotators working concurrently, the process requires a full month of dedicated effort. In dynamic environments—such as tracking customer reaction to a product launch or a PR crisis—a four-week lead time for data preparation is unacceptable. The sentiment will have shifted, and the opportunity for intervention will have passed.

2.3. The Recurring Cost of Maintenance (Label Drift)

The costs calculated above assume a one-time static dataset.

However, language and market context are fluid. A model trained on movie reviews from 2010 may fail to accurately interpret the slang, cultural references, or emerging topics of 2025. This phenomenon, known as concept drift or label drift, necessitates periodic retraining of supervised models.

Each retraining cycle requires a fresh injection of labelled data to capture the new distribution. Consequently, the $20,000+ investment in labelling is not a capital expenditure (CapEx) but an operational expenditure (OpEx), creating a permanent drain on resources. Unsupervised learning, by contrast, incurs near-zero marginal cost for new data; the algorithm simply re-clusters the new inputs based on their inherent statistical properties.

3. Theoretical Framework: Mechanisms of Unsupervised Semantic Discovery

Having established the economic necessity of unsupervised methods, we must now examine the theoretical mechanisms that underpin their viability. The experiment relies on a pipeline that transforms linguistic symbols into geometric entities. This transformation rests on the Distributional Hypothesis, which posits that words that occur in similar contexts tend to have similar meanings.

3.1. The Vector Space Model (VSM) and TF-IDF

The fundamental challenge in NLP is converting variable-length text strings into fixed-length numerical vectors that algorithms can process. The experiment employs the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme, a cornerstone of information retrieval [9].

In a standard Bag-of-Words (BoW) model, a document is represented by the frequency of its constituent words. However, this approach suffers from the dominance of high-frequency functional words (e.g., "the," "and," "is") which carry little semantic content. TF-IDF corrects this by introducing a penalty term for ubiquity.

The TF-IDF weight $w_{t,d}$ for a term t in document d is defined as:

$$w_{t,d} = tf(t,d) * \log(\frac{N}{df(t)})$$

Where:

1) N is the total number of documents in the corpus.
2) df(t) is the number of documents containing term t.

Implications for Sentiment clustering: In the context of customer feedback, sentiment-bearing adjectives like "atrocious," "sublime," "refund," or "masterpiece" tend to have relatively low document frequencies df(t) compared to neutral nouns like "movie" or "product." Consequently, the $\log(\frac{N}{df(t)})$ term—the Inverse Document Frequency—boosts the vector weights of these sentiment-rich words. This mathematical transformation effectively highlights the features that distinguish positive reviews from negative ones, pushing them into distinct regions of the vector space without any explicit instructions [15].

3.2. Dimensionality Reduction: The Role of Truncated SVD

Text data is inherently high-dimensional. A vocabulary of 50,000 unique words results in a vector space of 50,000 dimensions. However, a single review typically contains only a few hundred words, resulting in a sparse matrix in which over 99% of the values are zero.

High dimensionality presents two problems:

1) *Computational Cost:* Clustering algorithms such as K-Means compute distances between points. Computing the Euclidean distance between 50,000 points in 50,000 dimensions is computationally expensive.
2) *The Curse of Dimensionality:* As dimensions increase, the volume of the space increases so fast that the available data becomes sparse. In a very high-dimensional space, data points tend to become equispaced, rendering distance-based clustering ineffective.

To mitigate this, the experiment applies Truncated Singular Value Decomposition (SVD) to the TF-IDF matrix. This technique, often referred to in NLP contexts as Latent Semantic Analysis (LSA), factorises the matrix to identify the directions of greatest variance. By retaining only the top 300 components (as done in the experiment), the method compresses the data while preserving its semantic structure [9].

Semantic Coalescence: Crucially, SVD helps resolve synonymy. If Review A uses "scary" and Review B uses "terrifying," a raw TF-IDF model sees these as orthogonal dimensions. However, because these words co-occur in similar contexts across the corpus, SVD collapses them into the same latent semantic component. This allows the clustering algorithm to group Review A and Review B together, recognising they share a "horror/fear" sentiment despite having no overlapping vocabulary.

3.3. Geometric Clustering: K-Means

The final stage of the pipeline is K-Means clustering. This algorithm partitions the dataset into K clusters by minimising the within-cluster variance. It assumes that valid groupings (sentiments) exist as dense, spherical clouds of points in the vector space.

The algorithm iteratively assigns each data point to the nearest centroid (cluster centre) and then recalculates the centroid based on the new members.

$$\arg min_{S=\{S_1,S_2,\ldots,S_k\}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

In the experiment, K is set to 2, corresponding to the binary nature of the sentiment (Positive vs. Negative). While the algorithm is "blind" to the concepts of positivity or negativity, it detects that the data naturally separates into two dominant groups—one characterised by vocabulary such as "wonderful, great, best" — and the other by "bad, waste, worst" [9].

4. Case Study: The IMDb Unsupervised Experiment

The experiment provides practical validation of the theoretical framework described above. Applying these techniques to the IMDb Large Movie Review Dataset provides empirical evidence of the method's efficacy.

4.1. Methodology and Preprocessing Pipeline

The study utilised a subset of 5,000 reviews for rapid prototyping and the full 50,000-review dataset for final validation. The dataset is balanced, with an equal number of positive and negative reviews. The preprocessing pipeline implemented is rigorous and highly specific to the domain of sentiment analysis:

1) Lowercasing & Normalisation: Standardising text to reduce vocabulary size.
2) Contraction Expansion: Using the contractions library to convert terms like "isn't" to "is not" or "won't" to "will not" [9].

Insight: This step is critical for sentiment analysis. The token "not" is a sentiment inverter. In a standard tokeniser, "isn't" might be treated as a unique token or split in a way that obscures the negation. Explicitly expanding it ensures that the "not" token can interact with adjectives (e.g., "not good"), preserving negative polarity.

3) Punctuation and Whitespace Removal: Cleaning noise that does not contribute to semantic meaning.
4) Stopword Removal: Eliminating high-frequency, low-information words using NLTK [9].

Lemmatisation (POS-Aware): Converting words to their base forms (e.g., "running" → "run", "better" → "good"). The experiment explicitly uses Part-of-Speech tagging to ensure correct reduction, differentiating between "meeting" (noun) and "meeting" (verb).

5) Named Entity Removal: A sophisticated step using SpaCy's transformer model (en_core_web_trf) to identify and remove person names [9].

Strategic Importance: This is a vital bias-reduction technique. In movie reviews, specific actors or directors (e.g., "Adam Sandler" or "Christopher Nolan") might be strongly correlated with specific sentiments due to their fan bases or typical movie quality. If names are retained, the clustering algorithm may group reviews by *topic* (e.g., "Movies starring Tom Cruise") rather than by *sentiment*. By removing names, the algorithm is forced to cluster based solely on descriptive emotional language.

4.2. Feature Extraction and Dimensionality Reduction

Following preprocessing, the textual data was transformed into a numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) vectorisation scheme, with an n-gram range of (1, 3). [9] This configuration allows the model to capture not only individual words (unigrams) but also contiguous word sequences (bigrams and trigrams). The inclusion of higher-order n-grams is a deliberate and impactful feature engineering decision for sentiment analysis tasks. While unigrams are effective at representing general topics or themes within a document, sentiment is frequently expressed through short phrases rather than isolated terms. Expressions such as *"not good," "waste of time,"* or *"highly recommend"* encode polarity that can be ambiguous or even misleading when their constituent words are considered in isolation.

The resulting TF-IDF representation produces a very high-dimensional, sparse document-term matrix, reflecting the large vocabulary introduced by the inclusion of multi-word expressions. To address both computational inefficiency and statistical challenges associated with such high-dimensional data, the matrix was subsequently subjected to a Truncated Singular Value Decomposition (SVD). This technique projects the original feature space onto a lower-dimensional latent semantic space, reducing the dimensionality to 300 components.

The report includes a cumulative explained variance plot to assess the extent to which the original variance in the data is preserved after dimensionality reduction. As expected, 300 components do not capture the full statistical variance in the original TF-IDF matrix. However, this outcome is neither unexpected nor undesirable. A substantial portion of the remaining variance lies in the long tail of infrequent terms, including rare words, misspellings, idiosyncratic expressions, and other forms of lexical noise that contribute little to semantic discrimination. Retaining these components can, in fact, degrade clustering performance by amplifying noise and diluting meaningful structure.

By truncating the SVD to 300 components, the model effectively filters out low-variance components and focuses on the dominant latent dimensions that encode shared semantic patterns across documents. This process allows conceptually similar reviews—despite differing surface-level vocabulary—to be positioned closer together in the reduced vector space. Consequently, the clustering algorithm operates on a representation that emphasises the dataset's semantic core rather than superficial lexical variation, thereby improving robustness and interpretability in the subsequent unsupervised analysis [9].

The number of SVD components was selected based on the cumulative explained variance curve, with the aim of balancing semantic retention with computational efficiency. Empirical inspection showed that the marginal gain in explained variance diminished substantially beyond approximately 300

components, indicating that additional dimensions primarily captured low-variance lexical noise rather than meaningful semantic structure. Retaining 300 components, therefore, preserved the dominant latent dimensions while ensuring tractable computation for large-scale datasets. The number of clusters (K = 2) was chosen based on the known binary polarity structure of the IMDb dataset (positive versus negative sentiment). Since the objective of this experiment was to evaluate whether unsupervised clustering could recover the underlying sentiment manifold, selecting K = 2 provided a direct test of the alignment between emergent clusters and ground-truth polarity labels.

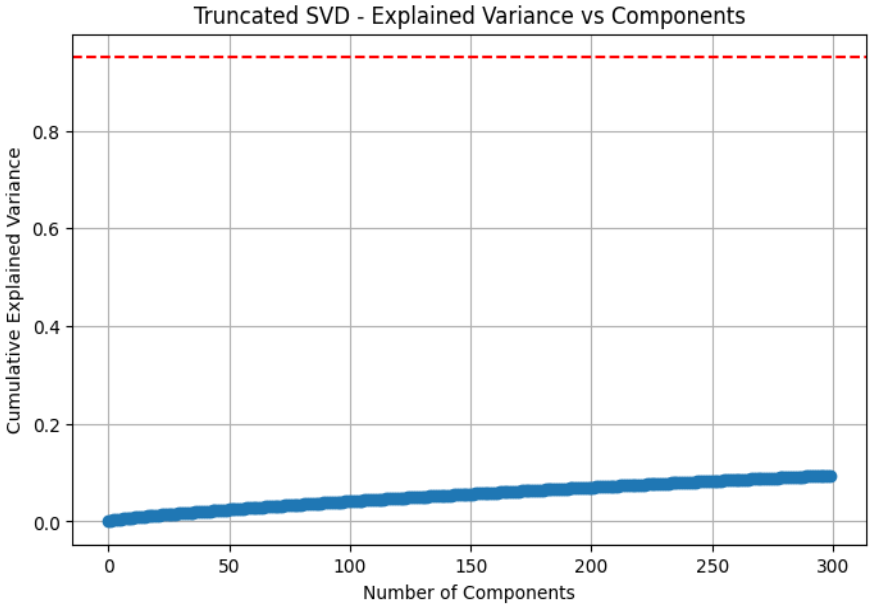

Figure 3. *Cumulative explained variance from Truncated SVD applied to TF-IDF features, showing that 300 components capture the semantic core of the dataset while discarding low-variance noise from the long tail of rare terms.*

4.3. Evaluation and Results

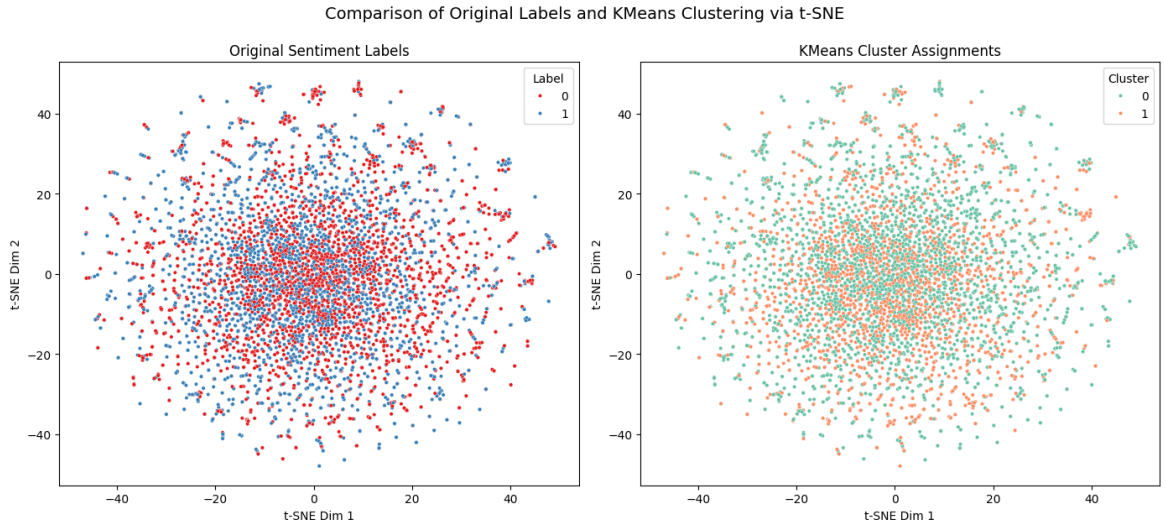

Figure 4. *t-SNE visualisation comparing ground-truth sentiment labels (left) with K-Means cluster assignments (right), showing that the unsupervised clusters closely align with the underlying Positive and Negative sentiment structure of the IMDb dataset.*

Since the IMDb dataset contains ground-truth labels (Positive/Negative), the experiment could objectively evaluate the quality of the unsupervised clusters by mapping them back to these labels.

Table 1. Cluster Alignment with Ground Truth [9].

Original Label	Count
Positive (1)	2,504
Negative (0)	2,496

Table 2. K-Means Assigned Cluster Counts [9].

Assigned Cluster	Count
Cluster 0	2,991
Cluster 1	2,009

While the split is not perfectly 50/50, the algorithm successfully identified two distinct, massive groupings. To verify that these groupings corresponded to sentiment, the experiment utilised t-SNE (t-Distributed Stochastic Neighbour Embedding) for visualisation.

The t-SNE plots provide strong evidence of success. The projection coloured by Original Labels shows two overlapping but distinct lobes of data points. The projection coloured by K-Means Cluster Assignments shows a nearly identical structure. This visual correlation confirms that the K-Means algorithm effectively recovered the underlying sentiment manifold of the data. It separated the reviews into two groups that closely map to the human-defined concepts of "Positive" and "Negative," based solely on the mathematical proximity of their TF-IDF vectors [9].

It is important to recognise that t-SNE is a non-deterministic, exploratory visualisation tool meant for qualitative analysis of high-dimensional data. Although it provides compelling visual evidence of cluster organisation, it is not a formal metric for assessing clustering and should be considered alongside quantitative measures of alignment.

5. Comparative Analysis: Unsupervised vs. Supervised Performance

To provide a comprehensive technical assessment, we must contextualise the findings within the broader landscape of sentiment analysis benchmarks. How much accuracy is sacrificed by abandoning supervision?

5.1. Benchmark Accuracy Metrics

Research replicating this specific pipeline (TF-IDF + K-Means on IMDb) typically reports clustering accuracy (purity) in the range of 60%-73% [7].

In this context, "accuracy" refers to cluster purity, measured by comparing unsupervised cluster assignments against the actual sentiment labels in the benchmark dataset. Unlike supervised classification accuracy, clustering effectiveness is often assessed using metrics such as the Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), or the silhouette score. Since IMDb provides labelled ground truth, purity offers a straightforward and meaningful way to evaluate how well the formed clusters correspond to the known sentiment categories.

1) *Interpretation:* This means that roughly 7 out of 10 reviews are correctly grouped into their dominant sentiment category without any human training.
2) *Baseline:* Random guessing would yield 50% accuracy on a balanced dataset. The unsupervised approach provides a significant 20%+ lift over chance.

In contrast, supervised models trained on the full 25,000-sample training set achieve higher performance:

1) Naive Bayes / SVM: 85% - 89% accuracy [10].
2) Deep Learning (LSTM/CNN): 88% - 90% accuracy [14].
3) State-of-the-Art (BERT/RoBERTa): 94% - 96% accuracy [12].

Table 3. The Cost-Accuracy Trade-off.

Methodology	Labelling Cost	Computational Cost	Setup Time	Approx. Accuracy
Unsupervised (K-Means)	$0	Very Low	Days	~70%
Supervised (SVM)	High ($5k+)	Low	Weeks	~88%
Supervised (BERT)	High ($5k+)	High (GPU)	Weeks	~95%

5.2. The "Good Enough" Threshold

The discrepancy in accuracy (70% vs. 95%) raises a strategic question: Is 70% accuracy sufficient?

For automated decision-making (e.g., automatically refunding a customer in response to a negative review), 70% is likely insufficient; the risk of error is too high. However, for aggregate market intelligence, trend analysis, and customer segmentation, 70% is often sufficient. If a business wants to know

if sentiment is trending up or down after a product update, or if they want to group customers into "Promoters" and "Detractors" for a marketing campaign, the broad strokes provided by unsupervised clustering offer actionable utility.

The unsupervised model effectively filters noise and organises the data, allowing human analysts to focus on representative samples from each cluster rather than reading thousands of raw documents.

6. Strategic Implementation: From Clustering to Customer Intelligence

Having quantified the financial and temporal burden of supervised pipelines earlier, this section focuses specifically on the operational implications of adopting an unsupervised alternative.

The ability to cluster customers based on feedback without manual labelling opens new strategic avenues for businesses. It transforms the VoC process from a reactive, resource-constrained activity into a proactive, scalable capability.

6.1. Scalability and Real-time Analysis

The most profound advantage of the unsupervised approach is scalability. The computational cost of K-Means scales linearly with the number of data points (O(n)). This means the system can process 50,000 reviews almost as easily as 5,000. In contrast, supervised learning has a linear *labelling* cost—analysing 10x more data requires 10x more human hours to create the training set (or risks model drift if the training set is not updated) [16].

Furthermore, unsupervised pipelines can be deployed in real-time. As new reviews arrive via the API, they can be projected into the existing SVD vector space and immediately assigned to the nearest cluster centroid. This enables the creation of real-time dashboards that track the "pulse" of customer sentiment without the lag associated with human annotation cycles [11].

6.2. Domain Agnosticism

Supervised models are notoriously brittle when moving between domains. A model trained on movie reviews will perform poorly on restaurant reviews because the sentiment lexicons differ (e.g., "delicious" vs "thrilling"). [2] Transferring a supervised model requires a technique known as domain adaptation, which often necessitates labelling a new dataset for the target domain.

Unsupervised clustering is domain agnostic. The pipeline described in the experiment (TF-IDF + SVD + K-Means) relies only on the statistical distribution of words within the *current* corpus. If applied to restaurant reviews, it will automatically identify that "tasteless" and "cold" cluster together (Negative) while "tasty" and "fresh" cluster together (Positive), without requiring any new labels. This flexibility reduces substantial costs for conglomerates operating across multiple product lines or verticals.

6.3. Discovery of "Unknown Unknowns"

While the IMDb experiment focused on binary sentiment (Positive/Negative), K-Means can be configured with $K > 2$ to discover more granular structures. In a retail context, setting $K=10$ might reveal clusters corresponding not just to sentiment, but to specific topics:

1) Cluster A: "Shipping delays / Lost packages"
2) Cluster B: "Product quality / Material issues"
3) Cluster C: "Customer service / Rude staff"

Supervised models can only classify text into categories predefined by the taxonomy designers. Unsupervised clustering can reveal emergent categories—issues or trends that the business was previously unaware of and thus had not labelled.

7. Limitations and Advanced Mitigations

While powerful, the TF-IDF/K-Means pipeline has limitations. K-Means assumes that clusters are spherical and of roughly equal size, which may not always hold true for complex textual manifolds. Additionally, TF-IDF is a "Bag of Words" model; it struggles with complex syntactic structures, sarcasm, and context-dependent meaning (e.g., "The movie was not bad" might be clustered with "bad" if negation is not handled correctly).

7.1. Advancing to Embeddings (BERT/Word2Vec)

Future iterations of this work can enhance accuracy by replacing TF-IDF with Word Embeddings (Word2Vec, GloVe) or Contextual Embeddings (BERT). [13] Unlike TF-IDF, embeddings represent words as dense vectors where semantic similarity is encoded as geometric proximity. "Terrible" and "Awful" are close together in embedding space, even if they never co-occur in the same document [6].

The present study intentionally employs TF-IDF with Truncated SVD as a computationally lightweight and economically efficient baseline. While embedding-based clustering may yield improved semantic representations, such approaches often introduce higher computational complexity, thereby altering the cost-efficiency trade-off that forms the central focus of this analysis.

Recent research on Deep Embedded Clustering (DEC) combines the feature-extraction capabilities of deep neural networks with clustering objectives, achieving significantly higher purity than standard K-Means while remaining fully unsupervised [17].

7.2. Hybrid Approaches: Active Learning

To bridge the gap between the low cost of unsupervised learning and the high accuracy of supervised learning, organisations can employ Active Learning [5] or Semi-Supervised Learning.

In this workflow:

1) Run the Unsupervised Clustering pipeline to group the data.
2) Identify the centroids (most representative samples) and the boundary cases (most ambiguous samples).
3) Manually label only these critical samples (e.g., 500 documents instead of 50,000).
4) Propagate these labels to the rest of the cluster or use them to train a lightweight classifier.

This "Human-in-the-Loop" approach can reduce labelling costs by 90% while achieving accuracy comparable to fully supervised models [4].

8. Conclusion

The experimental evidence, supported by broader industry research, confirms that unsupervised clustering is a highly effective, cost-efficient strategy for customer segmentation. By leveraging the mathematical properties of language—specifically, the tendency of similar sentiments to share lexical patterns—algorithms such as K-Means can recover the latent structure of customer feedback without requiring costly human supervision.

While it may not yet match the precision of state-of-the-art supervised models for granular classification tasks, the unsupervised approach offers a superior Return on Investment (ROI) for exploratory analysis and broad-spectrum sentiment tracking. It eliminates the "labelling tax" that stifles AI innovation, allowing businesses to turn their massive archives of unstructured text into an active asset.

For organisations seeking to "listen" to their customers at scale, the path forward is clear: start with unsupervised discovery. It offers the fastest route to insight, the lowest barrier to entry, and the flexibility to adapt to an ever-changing marketplace. The enormous time and expense of manual labelling need no longer be the price of entry for understanding the voice of the customer.

9. Related Work

Majumdar, P. (2025, July 2). Finding hidden customer sentiments without labels: how semantic AI and clustering can automatically detect patterns in movie reviews.

Recent sentiment analysis research has primarily focused on supervised learning methods, including traditional approaches such as Naïve Bayes and Support Vector Machines, as well as advanced deep neural networks, such as LSTM, CNN, and Transformer models like BERT, which often achieve over 90% accuracy on benchmark datasets such as IMDb. Meanwhile, some studies have examined unsupervised and semi-supervised techniques such as TF-IDF with K-Means, embedding-based clustering, and Deep Embedded Clustering, achieving moderate yet valuable results without labelled data. There is also a growing focus on active and semi-supervised learning to lessen annotation efforts. Nonetheless, earlier research generally treated technical performance and annotation costs as separate issues. This study advances the field by combining a quantitative economic evaluation of data annotation costs with practical testing of an unsupervised semantic clustering method, positioning unsupervised sentiment analysis not only as an alternative modelling approach but also as a cost-effective, strategic solution for large-scale Voice of Customer insights.

Abbreviations

$	The United States of America Dollar
AI	Artificial Intelligence
API	Application Programming Interface
ARI	Adjusted Rand Index
BERT	Bidirectional Encoder Representations for Transformers
BoW	Bag of Words
CapEx	Capital Expenses
CNN	Convolutional Neural Network
DEC	Deep Embedded Clustering
GloVe	Global Vector of Word
hr	Hour
IAA	Inter-Annotator Agreement
IMDb	Internet Movie Database
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMI	Normalised Mutual Information
OpEx	Operational Expenses
POS	Part of Speech
PR	Public Relations
RoBERTa	Robustly Optimised BERT Pretraining Approach
ROI	Return on Investment
SVD	Singular Value Decomposition
t-SNE	t-distributed Stochastic Neighbour Embedding
TF-IDF	Term Frequency-Inverse Document Frequency
VoC	Voice of Customer
VSM	Vector Space Model
Word2Vec	Word to Vector

Author Contributions

Partha Majumdar: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Admon, W. (n.d.). How much do data annotation services cost? The complete guide 2025. Basic AI. https://www.basic.ai/blog-post/how-much-do-data-annotation-services-cost-complete-guide-2025

[2] Al-Moslmi, T., Omar, N., Abdullah, S., & AlBared, M. (2017, August 29). Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Explore. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7891035

[3] Brysbaert, M. (n.d.). How many words do we read per minute? A review and meta-analysis of reading rate. Ghent University.

[4] Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-supervised learning. MIT Press.

[5] Esuli, A., & Sebastiani, F. (n.d.). Active learning strategies for multi-label text classification. https://iris.cnr.it/retrieve/3602e6cb-5631-48b0-bfbc-5338f8e20529/prod_160970-doc_129465.pdf

[6] Inbarasu, P., Karthikeyan, R., & Naveen, V. (2025). Enhanced sentiment analysis of user reviews using BERT and K-means clustering. International Journal of Engineering and Techniques. https://ijetjournal.org/wp-content/uploads/IJET-V11I3P4.pdf

[7] Kumar, P., Kumar, A., & Malik, K. (2025). An empirical sentiment analysis of movie reviews by utilizing machine learning algorithms. National Research Journal of Information Technology & Information Science. https://www.nrjitis.in/images/paper_pdffiles/AN%20-69132c826b302.pdf

[8] LTS GDS (n.d.). Data annotation pricing: the complete cost guide for enterprise AI projects in 2025. https://www.gdsonline.tech/data-annotation-pricing/

[9] Majumdar, P. (2025, July 2). Finding hidden customer sentiments without labels: how semantic AI and clustering can automatically detect patterns in movie reviews. Zenodo. https://doi.org/10.5281/zenodo.15791416

[10] Naeem, M. Z., Rustam, F., Mehmood, A., Mui-zzud-din, Ashraf, I., & Choi, G. S. (2022, March 15). Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. PubMed Central. https://doi.org/10.7717/peerj-cs.914

[11] Pandhare, S., Mali, S., Sahani, K., & Shinde, A. H. (2025, October). A systematic review - sentiment analysis based on movie review. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/IJARSCT-29495

[12] Patten, S., Chen, P. Y., Schweikert, C., & Hsu, D. F. (2025, October 30). Enhancing sentiment classification with machine learning and combinatorial fusion. ArXiv. https://arxiv.org/html/2510.27014v1

[13] Ravi, J., & Kulkarni, S. (2023, February 7). Text embedding techniques for efficient clustering of twitter data. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC9904526/

[14] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021, February 25). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE. https://doi.org/10.1371/journal.pone.0245909

[15] Sampath, M., & Vignesh, S. (2024, July 10). Text clustering for topic identification: a TF-IDF and K-means approach applied to the 20 newsgroups dataset. EasyChair. https://easychair.org/publications/preprint/C6Ft

[16] Springbord (n.d.). Challenges of data labelling and how to overcome them. Data Labeling Services. https://www.springbord.com/blog/challenges-of-data-labeling/

[17] Zhang, Q., Li, Y., & Malik, M. S. A. (2025, August 19). Enhanced text clustering and sentiment analysis framework for online education: A BIF-DCN approach in computer education. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC12453793/