COMP4432 Group Project

Wang Yuqi, [redacted]
Liu Chen, [redacted]

1. Introduction

Our project presents a comprehensive exploration into the application of statistical learning and deep learning methodologies on the *A Million News Headlines* dataset from Kaggle, which comprises an extensive collection of unlabeled news headlines.

The study is structured into **three main sections**: 1) Section Section 2 presents our exploratory data analysis, detailing the iterative refinement process that informed our methodological decisions, 2) Section Section 3 provides an in-depth examination of our final architectural design including any implementation details, 3) Finally, Section 4 offers a comparative quantitative evaluation of the various approaches investigated, along with a detailed analysis of our innovative contributions that extend beyond baseline methodologies.

2. Exploratory Data Analysis

This section presents a detailed **chronological exposition** of our investigative process, documenting the evolutionary development of our methodologies and models, as well as the empirical justification for each new architectural decision.

2.1. Statistical Learning

Overview

Dimensionality Reduction:
- Latent Semantic Analysis (LSA),
- Latent Dirichlet Allocation (LDA),
- Non-negative Matrix Factorization (NMF)

Clustering Algorithm:
- k-means (w/ k-means++ initialization)

Visualization:
- t-SNE

Adhering to the principle of Occam's razor, our methodology followed a progressive complexity approach, starting with simple statistical learning baseline models and incrementally introducing sophisticated components before eventually reaching Deep Learning territory. However, this section primarily focuses on the statistical learning methods used.

2.1.1. Features (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is utilized as our initial feature matrix. TF-IDF quantifies term significance within documents while accounting for term prevalence across the entire corpus, thereby providing a balanced representation of word importance.

Directly applying k-means on the raw TF-IDF matrix yielded suboptimal results. As illustrated in Figure 1, the resulting clusters exhibit severe imbalance, with the pink cluster encompassing over 90% of the data points, while smaller clusters appear sporadically distributed.

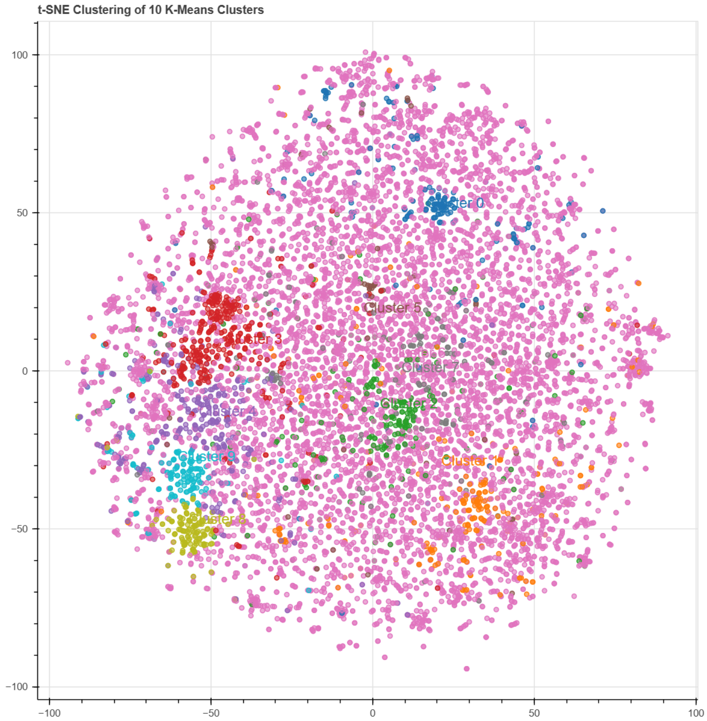


Figure 1: k-means clustering directly applied to TF-IDF matrix

Analysis of headlines within the pink cluster revealed no discernible patterns. While the smaller clusters demonstrated some patterns, these were primarily based on simplistic keyword matching without semantic understanding. Table 1 illustrates this limitation: headlines containing “us” were clustered together despite their completely different contextual meanings (e.g., “United States” versus pronoun usage).

Index	**Headline**
149184	us official says iran syria against all of us
1135986	the big dry see us hear us help us
1152296	what now for the us and china
1220409	us election
1194497	the two of us

Table 1: Top 5 headlines closest to cluster centroid

The clustering results gave us **two crucial insights**:

1. The imbalance of cluster size indicates that k-means alone is only able to distinguish between headlines of extreme similarity (e.g., all containing a term with high IDF), whereas headlines with less pronounced distinguishment are grouped into a gigantic nonsensical cluster.
2. Current method lacks contextual understanding and treats the same words with complete different contextual meaning as similar.

The ineffectiveness of the raw TF-IDF matrix stems from its susceptibility to the curse of dimensionality, characterized by high dimensionality and sparsity. This renders direct application of k-means clustering ineffective (solved with dimensionality reduction in the upcoming sections). Additionally, TF-IDF’s inability to capture contextual information presents a challenge that requires future attention (solved in Section 2.2).

2.1.2. LSA (Latent Semantic Analysis)

To address the primary challenges identified in Section 2.1.1 , we implemented dimensionality reduction, with Latent Semantic Analysis (LSA) as our initial approach.

LSA represents a well-established dimensionality reduction technique in NLP, designed to reveal underlying semantic structures within text data. The method operates by applying Singular Value Decomposition (SVD) to the TF-IDF matrix to generate a topic matrix.

Cluster	Top 5 Terms	# Headlines
1	**police man charged court crash**	**727921**
2	man charged murder court jailed	3375
3	police investigate probe search hunt	21243
4	abc news rural national business	32406
5	interview extended michael nrl john	43310
6	new abc man police news	33030
7	rural national news nsw health	19145
8	crash killed car us dies	72984
9	us court australia day says	232906
10	court fire crash accused face	57864

Table 2: Top 5 terms of each cluster after LSA + k-means

The LSA implementation demonstrates marked improvement over the results presented in Section 2.1.1 , particularly in terms of cluster size distribution. Nevertheless, significant challenges persist, notably in cluster 1, which exhibits poor cluster separability. This is evidenced by the substantial overlap between its high-frequency terms and those of other clusters as shown in Table 2. The visualization in Figure 2 further substantiates this observation, where the blue cluster shows considerable overlap with adjacent clusters in the reduced dimensional space.

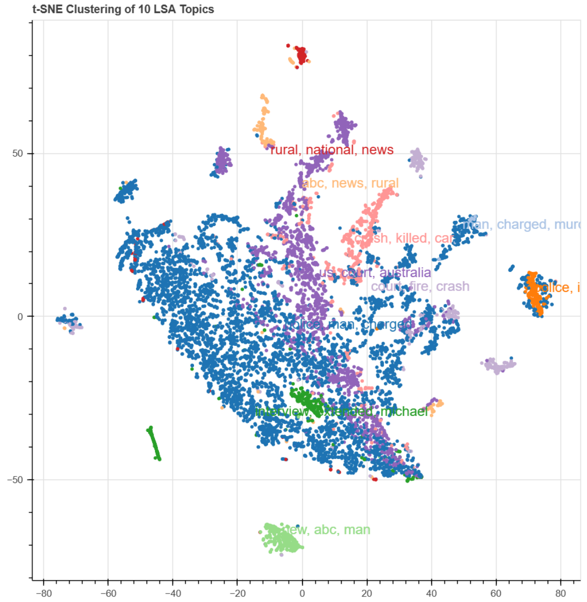


Figure 2: k-means clustering on LSA

2.1.3. NMF (Non-negative Matrix Factorization)

NMF is a matrix factorization technique used for topic modeling. It factorizes a non-negative document-term matrix A into two lower-dimensional non-negative matrices: W (document-topic matrix) and H (topic-word matrix). The goal is to minimize the reconstruction error: A ≈ W × H, where W captures the topic distribution for each document, and H represents the word distribution for each topic. NMF gives the result with all samples:

Cluster	Top 5 Terms	# Headlines
1	police investigate probe missing search	51826
2	**govt council nsw says plan**	**478157**
3	man charged murder missing jailed	49232
4	rural news national nsw exchange	27534
5	interview extended michael nrl john	36952
6	new zealand laws year cases	48540
7	abc weather business sport news	24018
8	crash car killed fatal dies	82429
9	australia us day south australian	359853
10	court accused face murder charges	85643

The application of NMF demonstrates notable improvements over LSA in terms of and clustering. Specifically, NMF reduces the dominance of large clusters seen in LSA, as evidenced by the more balanced cluster sizes in the results. This improvement can be attributed to NMF's non-negativity constraint, which ensures that the resulting topic distributions are more interpretable and focused, effectively capturing distinct semantic patterns in the data.

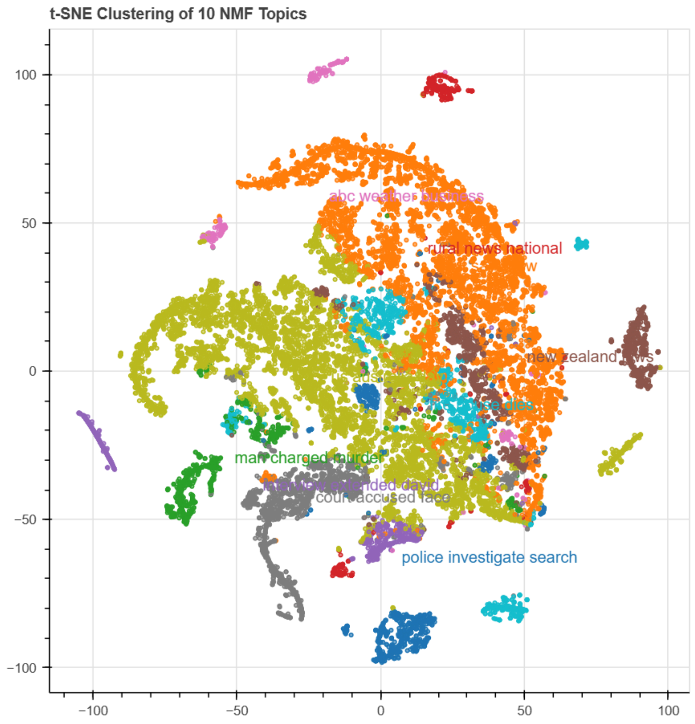


Figure 3:

However, challenges remain. While NMF mitigates some of the issues with overly large clusters, certain clusters (e.g., Cluster 2 still contain a disproportionately large number of headlines, indicating that some topics are overly broad or lack semantic coherence. This suggests that while NMF provides a step forward, further refinement is needed to improve topic separability and minimize overlap between clusters.

2.1.4. LDA (Latent Dirichlet Allocation)

LDA is yet another dimensionality reduction technique based on generative probabilistic model. It assumes that documents are mixtures of latent topics, and each topic is a distribution over words. The core idea is to uncover these hidden topics by estimating the probability of topics in each document and the probability of words in each topic.LDA uses a process of iteratively assigning words to topics and refining the distributions using Bayesian inference until convergence.

Cluster	Top 5 Terms	# Headlines
1	new crash qld sa year	12252
2	says water government **cup**	12553
3	court hospital sydney nt	12238
4	nsw govt interview plan	13005
5	man police charged home	12706
6	minister national farmers funding	12205
7	australia day report world	12808
8	police australian wa death	12233

Table 4: k-means clustering on LDA features

The LDA model achieved the strongest cluster balance with robust clustering quality (Table 4, Figure 4), but still faces the fundamental statistical learning limitation: it **relies solely on keyword frequency matching without contextual understanding**, resulting in semantically unrelated headlines being grouped together due to sharing surface-level vocabulary.

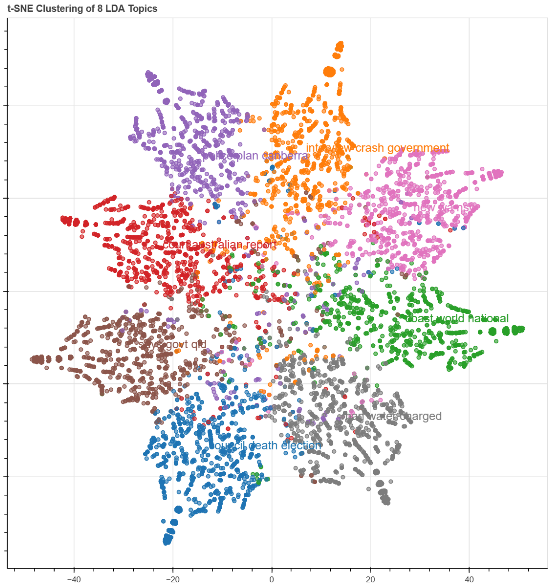


Figure 4: 2D t-SNE projection of k-means applied on top of LDA

Table 5 illustrates the semantic limitations of LDA's keyword-based clustering. The model groups headlines containing the word **cup** into Cluster 2, disregarding the term's contextual meaning across different **sports** and **law enforcement events**. For instance, headlines about World Cup finals coexist with those about law enforcement at Melbourne.

While these headlines share the keyword "cup," they represent fundamentally different semantic contexts. This semantic disparity, despite balanced cluster sizes, demonstrates statistical learning's core weakness: its reliance on surface-level word co-occurrence patterns prevents it from capturing the nuanced topical relationships that human readers naturally understand.

Cluster	Headline Example
2	mickelson prepares ryder **cup** day
2	serbia hopman **cup** final
2	15 minutes madness decided cricket world **cup** final
2	police crack the whip on melbourne **cup** drink
2	aussies brink davis **cup** win

Table 5: Headline from Cluster 2 showing semantic inconsistency despite shared keyword

2.2. Deep Learning

Overview

Dimensionality Reduction
- Symmetric Autoencoder
- Asymmetric Autoencoder (w/ Linear Output)

Clustering Algorithms
- k-means (w/ k-means++ initialization)

Embeddings
- Llama3.1-8B (4096 dimension)
- Llama3.2-1B (2048 dimension)

Visualization
- t-SNE

To address the contextual limitations of statistical learning methods discussed in Section 2.1 , we extended our approach to incorporate deep learning models.

2.2.1. Word Embeddings

Recent advances in Auto-regressive Large Language Models (AR-LLMs) have demonstrated unprecedented capabilities in context understanding and general knowledge representation. This motivated our approach to leverage state-of-the-art models for generating context-rich embeddings for each headline. We take the last embedding of the output layer, as in AR-LLMs, the final embedding encapsulates contextual information from all preceding context tokens.

Based on a careful evaluation of computational requirements and embedding quality, we selected two open-source models from the Llama family: Llama3.2-1B and Llama3.1-8B. We deployed these models locally using Ollama, an open-source platform for running large language models. The embedding generation process was implemented in batches of 10,000 rows using the ollama.embeddings API, with the implementation details available in `dl/embeddings.py`.

To overcome the lack of context awareness in the statistical learning methods as discussed in Section 2.1 , we further step-up the complexity and introduce deep learning models.

2.2.2. Dimensionality Reduction

As encountered in Section 2.1.1 , the high dimensionality of raw word embeddings poses challenges for clustering algorithms such as k-means, necessitating effective dimensionality reduction techniques.

Our approach leverages various autoencoder architectures, which comprise an encoder-decoder structure with an intentional bottleneck layer. The fundamental principle of autoencoders lies in their ability to learn efficient low-dimensional representations by encouraging the network to reconstruct high-dimensional input through a compressed intermediate representation. This compression-reconstruction mechanism, combined with neural networks' non-linear capabilities, enables non-linear dimensionality reduction of the word embeddings.

2.2.3. Overparameterization

One might notice that, our architectures discussed below are significantly overparameterized, with the Llama3.2 autoencoder comprising over 7.5 million trainable parameters and the Llama3.1 autoencoder exceeding 12 million parameters, both substantially surpassing the dataset size. Counterintuitively, these overparameterized models demonstrated superior performance, achieving lower validation losses compared to their underparameterized counterparts.

This seemingly paradoxical behavior aligns with a well-documented phenomenon in deep learning known as *Double Descent*, which challenges the classical bias-variance trade-off paradigm, showing that beyond a certain threshold of model complexity, test performance can improve with increasing model size due to neural networks' inherent self-regularization properties.

2.2.4. Baseline Autoencoder

Our initial implementation employed a conventional autoencoder configuration featuring **symmetric** four-hidden-layer architectures for both encoder and decoder components. For embeddings from Llama3.2-1B (2048 dimensions), the encoder architecture follows the dimensional progression: $(2048) \to 1024 \to 1024 \to 512 \to 256 \to (30)$, with the decoder mirroring this structure in reverse: $(30) \to 256 \to 512 \to 1024 \to 1024 \to (2048)$. For the Llama3.1-7B model's 4096-dimensional embeddings, we maintained a similar architectural pattern, adjusting only the input and output dimensions: $(4096) \to 1024 \to 1024 \to 512 \to 256 \to (30)$.

The training process optimizes mean-squared error (MSE) loss between input embeddings and their reconstructions, incorporating a dropout rate of 0.2 for regularization.

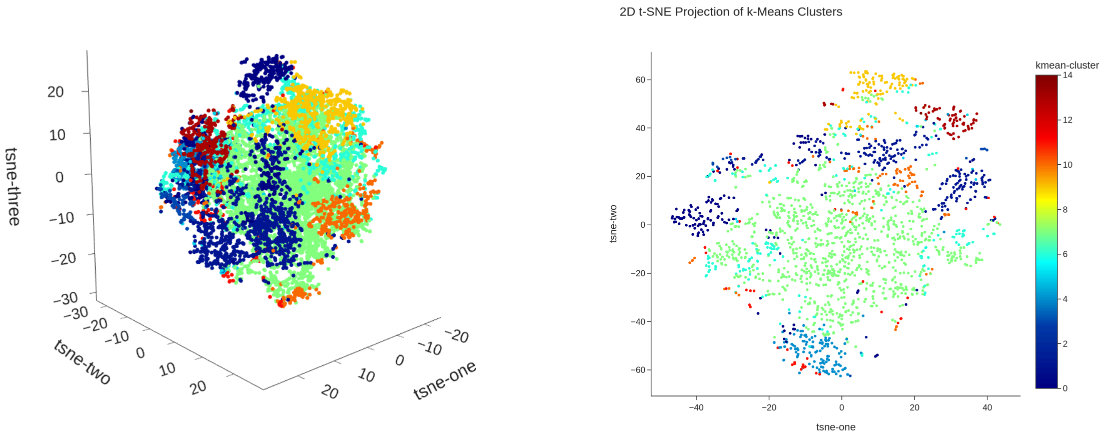
2.2.5. Limitation of the Baseline Autoencoder

Key Points

Insufficient **discriminative power**, shown by:

- vague cluster boundaries in t-SNE plot
- decreasing SS and CHI scores as k increase

While the baseline autoencoder successfully leverages enhanced contextual understanding, its latent representations demonstrate **insufficient discriminative power** for effective cluster formation. Figure 5 illustrates this limitation, showing poorly delineated cluster boundaries in both 2D and 3D visualizations.



,

Figure 5: 2D (right) and 3D (left) t-SNE projection of k-means clustering of AE latent

This clustering inefficacy is quantitatively validated by the deteriorating trends in both Silhouette Score (SS) and Calinski–Harabasz Index (CHI) as the number of clusters k increases, as shown in Figure 6. The Silhouette Score (SS), defined as:

$$s(i) = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

provides crucial insights into cluster quality. The score approaches 0 when data points exhibit equal distances to multiple cluster centroids ($b(i) \approx a(i)$), where $a(i)$ represents intra-cluster cohesion and $b(i)$ measures inter-cluster separation. The declining SS trend with increasing k indicates that the latent space represents data points in a uniformly distributed manner, lacking natural cluster boundaries. This uniform distribution is evidenced by the fact that increasing k fails to improve cohesion ($a(i)$) while deteriorating separation ($b(i)$) due to more and more equally distant cluster centroids.

Similarly, the diminishing CHI as the number of clusters k increases suggests that the relative separation between clusters diminishes while the compactness within clusters fails to improve proportionally. This phenomenon can be heuristically attributed to the latent space's inability to encapsulate sufficiently distinct semantic boundaries among the word embeddings.

Given these limitations, it becomes imperative to enhance the AE's capacity to generate more discriminative latent representations. The following sections introduce two architectural improvements designed to address these shortcomings and achieve more effective semantic clustering.

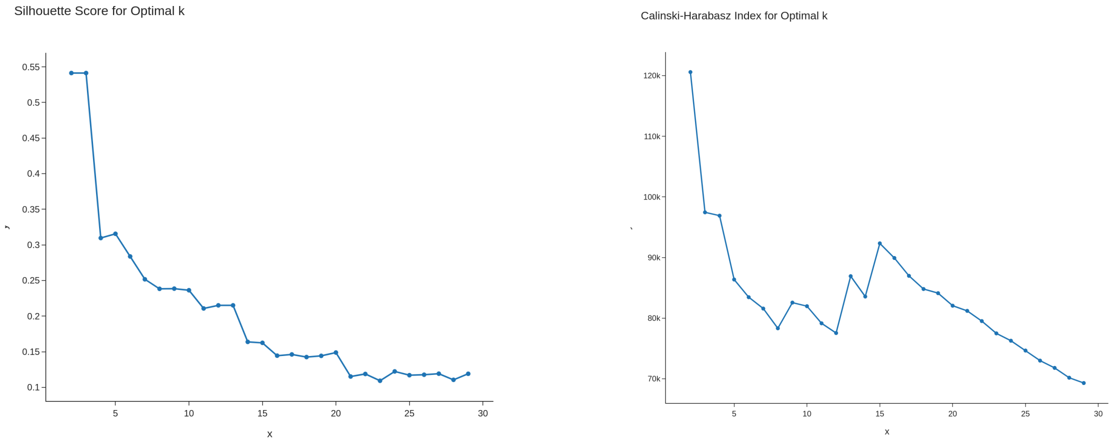


Figure 6: SS plot against k (left graph), CHI plotted against k (right graph)

2.2.6. Improved Autoencoders

To address the limitations identified in Section 2.2.5 , we propose three changes:

1. **Bottleneck Dimension Reduction**: Analysis of the baseline AE's encodings revealed significant sparsity, with an average of 10 out of 30 dimensions having zero activation. This observation suggests that the original bottleneck dimension may be unnecessarily large, potentially contributing to the suboptimal clustering performance. We therefore propose reducing the bottleneck dimensionality to encourage more compact and meaningful representations.

2. **Asymmetric Encoder-Decoder**: We also hypothesize that the poor discriminative power of the baseline autoencoder's latent representations may stem from the overly deep decoder architecture. When the decoder possesses excessive non-linear transformation capacity, the encoder is not sufficiently incentivized to produce linearly separable latent embeddings, instead potentially over-relying on the decoder's non-linear capabilities. To address this, we propose simplifying the decoder to a single linear layer that directly maps latent embeddings to reconstructed outputs.

3. **Prompted Word Embeddings**: Recognizing that the final token embedding in AR-LLMs is enriched by all preceding tokens and that AR-LLMs possess an exceptional ability to follow human instructions, we incorporated a custom prompt preceding the headlines before they are processed by Llama 3 for word embedding generation. This approach leverages the contextual understanding capabilities of AR-LLMs to enhance the quality of the generated embeddings.

Evaluation of Improvement 1: Reduced Bottleneck Dimension

Reducing the bottleneck dimension from 30 → 10 yielded two significant improvements:

1. Elimination of the sparsity observed in the baseline encodings
2. Enhanced cluster separability, as evidenced by the t-SNE visualizations (maximum perplexity used, indicating improved global structure preservation)

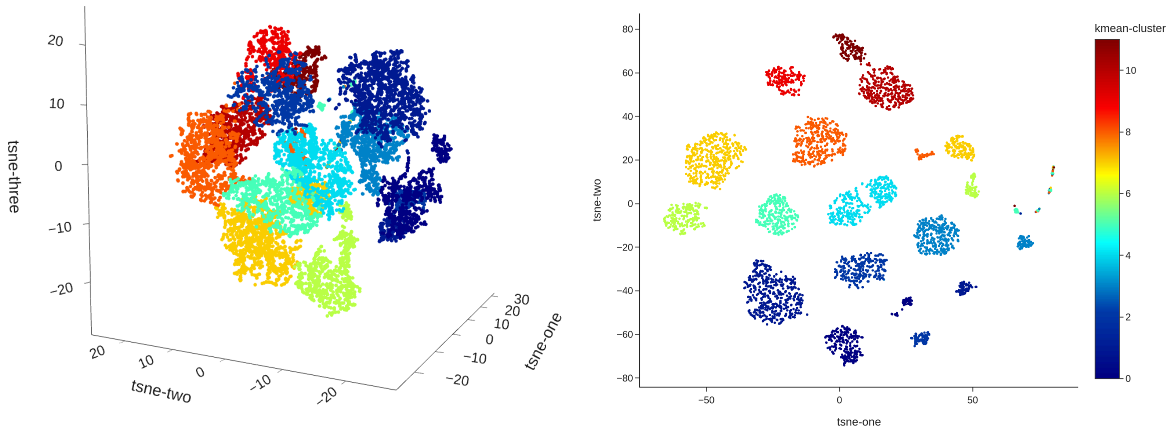


Figure 7: 3D (left) 2D (right) t-SNE projection of k-means on reduced bottleneck AE latent

The SS trend in Figure 8 reveals distinct peaks at k=7 and k=12, indicating that the reduced-dimension latent space supports natural clustering at multiple granularities. This multi-scale clustering capability suggests the latent space has captured meaningful semantic relationships.

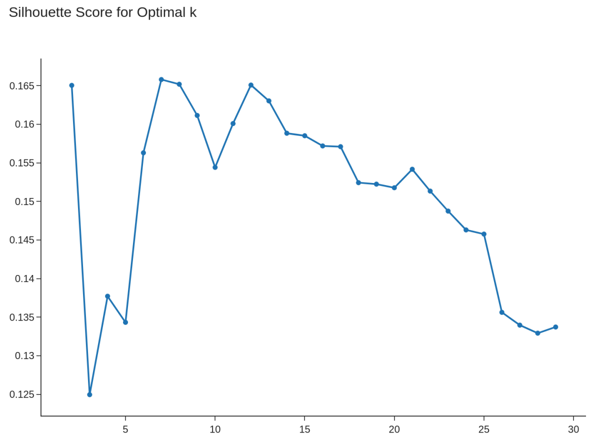


Figure 8: SS plotted against k, for k-means on reduced-bottleneck AE latent representation

Evaluation of Improvement 2: Asymmetric Architecture

The second improvement replace the multi-layer decoder with a single linear mapping (30 → 4096). This modification produced equally compelling results:

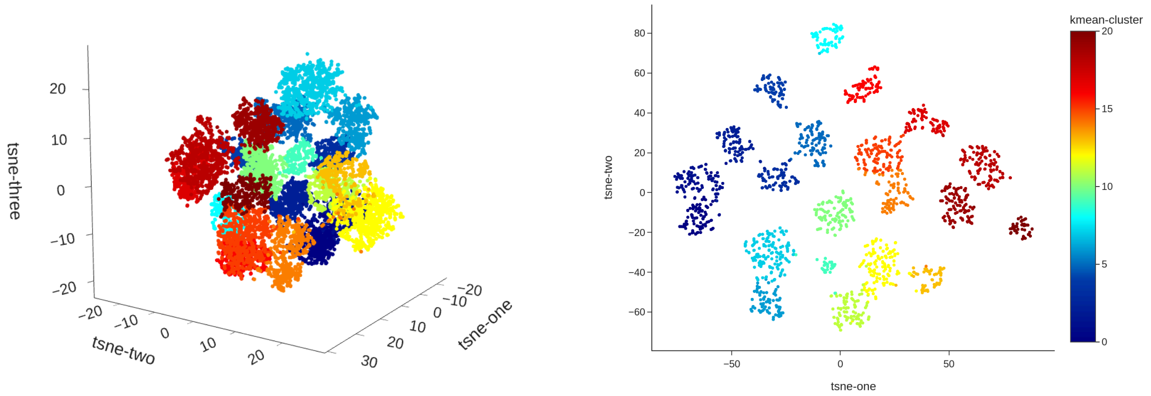


Figure 9: 3D (left) 2D (right) t-SNE projection of k-means on asymmetric AE latent

The t-SNE projections demonstrate clear cluster separation, with the increasing SS trend confirming the improved discriminative power of the latent representations.

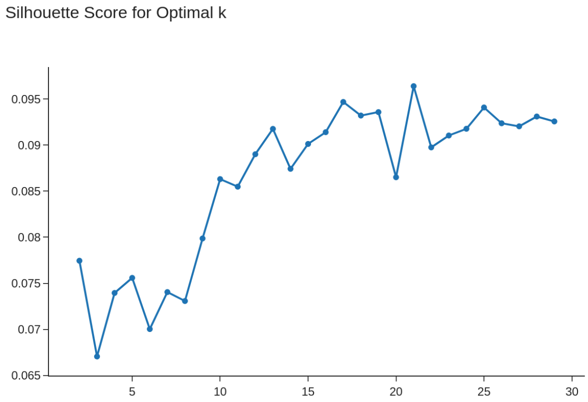


Figure 10: SS plotted against k, for k-means clustering on Asymmetric AE latent representation

Evaluation of Improvement 3: Prompted Word Embeddings

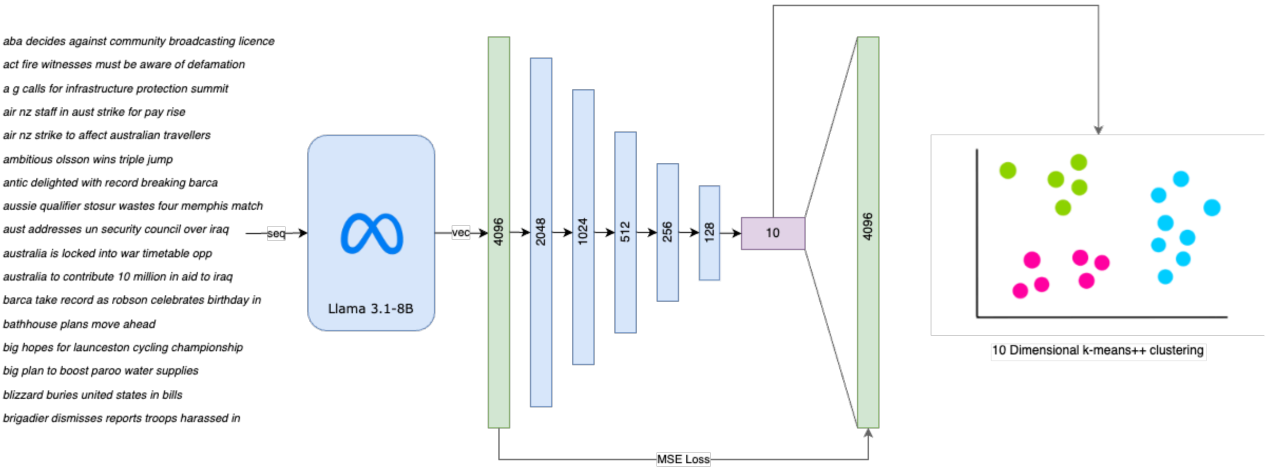
Custom Prompt Used:

- Represent various key aspects of this news headline for semantic clustering: “{headline}”

Quantitative evaluation of the enhancements introduced by the custom prompt is challenging. Nonetheless, we observe a marked increase in thematic coherence within each cluster, indicating improved cohesion. Previously, clusters contained several “noisy headlines” that did not clearly belong to the group. This issue has been substantially mitigated through the implementation of the custom prompt. A more detailed discussion of this qualitative improvement is provided in Section 4.1 .

(See Next Page)

3. Final Design


aba decides against community broadcasting licence
act fire witnesses must be aware of defamation
a g calls for infrastructure protection summit
air nz staff in aust strike for pay rise
air nz strike to affect australian travellers
ambitious olsson wins triple jump
antic delighted with record breaking barca
aussie qualifier stosur wastes four memphis match
aust addresses un security council over iraq
australia is locked into war timetable opp
australia to contribute 10 million in aid to iraq
barca take record as robson celebrates birthday in
bathhouse plans move ahead
big hopes for launceston cycling championship
big plan to boost paroo water supplies
blizzard buries united states in bills
brigadier dismisses reports troops harassed in
seq
Llama 3.1-8B
vec
4096
2048
1024
512
256
128
10
4096
MSE Loss
10 Dimensional k-means++ clustering

3.1. Model Card

3.1.1. Data

Cleaned Dataset: 1213003 rows
Transformed Dataset:

- Llama 3.1-8B Embeddings: 1213003 rows × 4096 cols
- Llama 3.2-1B Embeddings: 1213003 rows × 2048 cols

We eventually chose Llama 3.1-8B as the embedding for our final implementation, as we found its clustering results to be slightly superior to Llama 3.2-1B based on human evaluation.

3.1.2. Model

Type: Asymmetric Reduced-bottleneck Autoencoder (See Section 2.2.6)

Hyperparameters:

Encoder:

- Input Layer: 4096
- Hidden Layer: 2048, 1024, 256, 128
- Output Layer: 10

Decoder:

- Input Layer: 10
- Hidden Layer: *None*
- Output Layer: 4096

Dropout Rate: 0.2
Computation: 100 Epochs
Activation Function: LeakyReLU ($\alpha = 0.01$)

3.2. Training Procedure

1. Remove duplicates and nonsensical headlines from the dataset
2. Convert the dataset into word embeddings with Llama 3.1-8B
3. Train **Asymmetric Reduced-Bottleneck Autoencoder** for 100 Epoch
4. Convert the 4096 dimensional word embeddings into 10 dimensional
5. Train k-means++ clustering, select best k using SS and CHI

4. Results and Discussion

4.1. Qualitative Evaluation

Cluster 1

1. local captures the bushfire near peregian springs
2. grampians fire flares up overnight
3. more water released from wivenhoe somerset dams
4. rain damages broken hill buildings
5. out of control blze in victoria's high country
6. catastrophic fire risk closes wa schools
7. back burning fires started around bendigo

Cluster 2:

1. indigo shire determined to change council boundary
2. report urges council to accept vasse hospital site
3. kilkivan shire moves to greater transparency
4. shoalhaven council to go ahead with rezoning
5. mayor warns against gambling on goldfields future
6. doors set to close on eden state forests office
7. residents to get caloundra south briefing

Cluster 3:

1. figures show average mining wage tops 100k
2. higher nickel price may boost goldfields production
3. pay tv sector booming report says
4. mining industry ends 2016 on positive note
5. expect oil prices to rise caltex ceo
6. banks pounce on interest rate hike
7. abc learning moves to calm investors

Cluster 4:

1. iran saudi agree to curb violence
2. ethiopian troops to stay in somalia for weeks
3. un reinforcements due in south sudan in 48 hours
4. loyalist forces reportedly end mauritania uprising
5. philippine troops clash with hostage kidnappers
6. palestinian rocket attacks breach gaza truce
7. israel intercepts gaza bound aid ship

Cluster 5:

1. fears aired for regional development initiative
2. self extinguishing cigarettes law welcomed
3. cautious welcome for corrective service changes
4. enrolment changes allow last minute votes
5. transport left out of city living strategies
6. confusion reigns over sustainability declaration
7. push for air quality report action

Cluster 6:

1. elderly job seekers discouraged because of age
2. alcohol restrictions no long term solution report
3. lifeline anticipates strain on mental health
4. bush remedies vs western medicine
5. survey highlights growing youth debt
6. rules tightened for child models
7. thousands join online tracking of flu symptoms

Figure 11: Top 7 headlines (based on proximity to cluster centroid) from 6 of the 20 clusters

As shown in Figure 11, the clusters produced by our model demonstrate strong internal semantic cohesion beyond simple keyword matching, with headlines grouped by high-level conceptual themes. While some clusters like Cluster 4 (International Conflict/Diplomacy) and Cluster 1 (Natural Disasters) show precise thematic focus, others maintain clear topical relationships despite broader scope.

Cluster Themes and Key Concepts:

Cluster 1: Natural Disasters & Weather
- Fires, floods, emergency response
- Weather events, damage reports
- Infrastructure impacts

Cluster 2: Local Government & Planning
- Council decisions, zoning
- Municipal boundaries
- Community development

Cluster 3: Business & Financial Markets
- Industry performance
- Price trends, wages
- Market indicators

Cluster 4: International Conflicts / Diplomacy
- Military operations
- Peace negotiations
- Regional tensions

Cluster 5: Policy & Regulation
- Legislative changes
- Public initiatives
- Regulatory reforms

Cluster 6: Social Issues & Health
- Public health
- Demographics
- Community welfare

4.2. Quantitative Evaluation

Metric	Type	Formula
SS	Cluster	$\frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$
CHI	Cluster	$\frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)}\times\frac{N-k}{k-1}$
TD	Semantic	$\frac{\text{unique Words}}{T\times N}$
TC	Semantic	$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\binom{N}{2}}\sum_{1\leq i\leq j\leq N}\log\frac{P(w_i,w_j)}{P(w_i)\times P(w_j)}$

Embed	DR Algo	Cluster Algo	SS	TC*	TD*	CHI	Sum
TF-IDF	–	K-means++ (k=10)	0.004	0.001	0.330	12	0/4
	LSA	K-means++ (k=10)	0.097	0.342	0.550	3143	0/4
	NMF	K-means++ (k=10)	−0.04	**0.426**	0.950	2081	1/4
Term Freq	LDA	K-means++ (k=10)	**0.250**	0.245	**0.963**	**7573**	3/4
Llama3.1-8B	AE (dim=30)	K-means++ (k=20)	0.126	–	–	75.2K	0/2
	AE (dim=10)	K-means++ (k=20)	0.152	–	–	97.3K	0/2
	AEL** (dim=30)	K-means++ (k=20)	0.107	–	–	151K	0/2
	AEL** (dim=10)	K-means++ (k=20)	**0.501**	–	–	**6.8M**	2/2

* Llama3.1-8B embeddings capture semantics beyond term frequencies, TD and TC were not applicable

** AEL refers to Asymmetric **A**uto **E**ncoder with **L**inear Decoder in Section 2.2.6

Our quantiative evaluation shows that among the statistical learning methods, LDA excels among 3/4 metrics. Whereas, when it comes to Deep Leanring methods, Autoencoder which combined our two proposed improvements has the highest score, far exceeding the 2nd place.

4.3. Reflection: Our Innovations

In addition to our iterative improvements, we implemented **several noteworthy innovations**:

1. Decoder Simplification: We astutely identified the issues that caused the Baseline Autoencoder (AE) to underperform. To address this, we replaced the multi-layer decoder with a single linear layer. This modification demonstrates our comprehensive understanding of neural network linearity principles and enhances the model's efficiency.

2. Custom Prompt Integration: We incorporated a custom prompt prefixing each headline before processing. This enhancement results in greater semantic purity within each cluster, effectively reducing the presence of "noisy headlines." The qualitative improvements achieved through this approach are discussed in detail in the Qualitative Evaluation section.