Reasoning-Based Refinement of Unsupervised Text Clusters with LLMs

Tunazzina Islam
Department of Computer Science
Purdue University
West Lafayette, IN 47907
islam32@purdue.edu

Abstract

Unsupervised methods are widely used to induce latent semantic structure from large text collections, yet their outputs often contain incoherent, redundant, or poorly grounded clusters that are difficult to validate without labeled data. We propose a **reasoning-based refinement framework** that leverages large language models (LLMs) not as embedding generators, but as semantic judges that validate and restructure the outputs of arbitrary unsupervised clustering algorithms. Our framework introduces three reasoning stages: (i) **coherence verification**, where LLMs assess whether cluster summaries are supported by their member texts; (ii) **redundancy adjudication**, where candidate clusters are merged or rejected based on semantic overlap; and (iii) **label grounding**, where clusters are assigned interpretable labels through a two-stage process that generates and consolidates semantically similar labels in a fully unsupervised manner. This design decouples representation learning from structural validation and mitigates the common failure modes of embedding-only approaches. We evaluate the framework in real-world social media corpora from two platforms with distinct interaction models, demonstrating consistent improvements in cluster coherence and human-aligned labeling quality over classical topic models and recent representation-based baselines. Human evaluation shows strong agreement with LLM-generated labels, despite the absence of gold-standard annotations. We further conduct robustness analysis under matched temporal and volume conditions to assess cross-platform stability. Beyond empirical gains, our results suggest that LLM-based reasoning can serve as a general mechanism for validating and refining unsupervised semantic structure, enabling more reliable and interpretable analysis of large text collections without supervision.

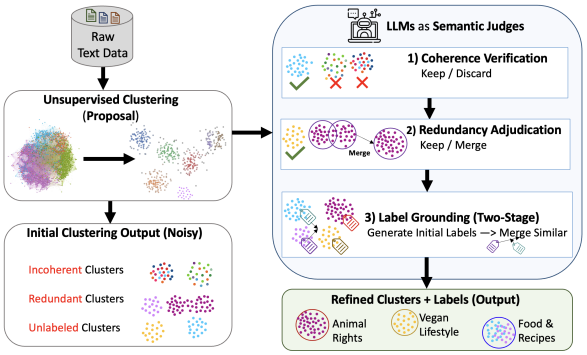


Figure 1: Overview of our framework. Unsupervised clustering generates initial cluster proposals that are often noisy. We treat these clusters as hypotheses and use LLMs as semantic judges to (1) verify coherence, (2) adjudicate redundancy, and (3) generate interpretable labels via a two-stage grounding process, producing refined, coherent, and distinct clusters.

1 Introduction

Large text collections are commonly analyzed using unsupervised methods to induce latent semantic structure, enabling downstream tasks such as summarization, monitoring, and social analysis without requiring labeled data (Angelov, 2020; Hoyer, 2004; Blei et al., 2003; Lee and Seung, 1999; Deerwester et al., 1990). Such methods are particularly important in domains where annotation is costly or infeasible, including social media, where discourse is noisy, rapidly evolving, and highly heterogeneous (Zappavigna, 2012). Despite their widespread use, unsupervised clustering pipelines often produce outputs that are noisy, redundant, or weakly interpretable, making it difficult to assess whether the induced structure meaningfully reflects the underlying semantics of the data.

A central challenge in unsupervised semantic structure induction is that the number of latent themes is typically **unknown a priori**. To address this, prior work has frequently adopted nonparametric formulations that allow the structure to grow with the data (Srijith et al., 2017), for example, through hierarchical Bayesian models such

Findings of the Association for Computational Linguistics: ACL 2026, pages 9899–9917

as the Hierarchical Dirichlet Process (HDP) (Teh et al., 2004). While such approaches mitigate the need to pre-specify the number of clusters, they do not resolve a more fundamental issue: whether the induced clusters are semantically coherent, non-redundant, and interpretable to humans, especially in short-text and high-noise settings such as social media.

More recent approaches improve unsupervised text clustering by learning stronger representations, typically using contextual sentence embeddings combined with geometric clustering criteria (Grootendorst, 2022; Reimers and Gurevych, 2019). These methods assess cluster quality through distance-based properties of the embedding space, such as separation or density, treating them as indicators of semantic coherence. However, prior work has shown that embedding geometry does not always align with human notions of meaning (Ethayarajh, 2019; Mimno et al., 2011). In practice, clusters can appear well separated numerically while remaining semantically incoherent, and multiple clusters may encode overlapping themes with only superficial lexical differences. As a result, representation-centric pipelines lack explicit mechanisms for verifying whether induced clusters are meaningful or interpretable from a semantic perspective.

In this paper, we take a different approach. Rather than proposing a new clustering algorithm or representation, we study how large language models (LLMs) (Brown et al., 2020) can be used as semantic reasoners to validate and restructure the outputs of arbitrary unsupervised clustering methods. Our key insight is that LLMs possess strong natural-language reasoning capabilities (Yao et al., 2022; Wei et al., 2022; Kojima et al., 2022) that can be leveraged to assess whether a proposed cluster is internally coherent, whether two clusters are meaningfully distinct, and whether an induced theme is well-grounded in the underlying texts. This enables a shift from purely statistical or geometric criteria toward explicit semantic validation.

We introduce a **reasoning-based cluster refinement** framework (Fig. 1) that treats clustering as a **proposal step** and uses LLM reasoning to adjudicate structure. The framework consists of **three** stages:
(1) coherence verification, where LLMs assess whether a cluster summary is supported by its member texts;
(2) redundancy adjudication, where candidate clusters are merged or rejected based on semantic overlap rather than embedding similarity alone; and
(3) label grounding, where clusters are assigned interpretable labels through a **two-stage** process that generates candidate labels and consolidates semantically similar ones.

Importantly, LLMs are used not as embedding generators, but as **semantic judges** that accept, reject, or revise structural hypotheses produced by unsupervised methods. Unlike topic modeling approaches that induce semantic structure from scratch, our framework treats clustering output as a hypothesis and uses LLM reasoning only to validate, prune, and ground that structure. This design decouples representation learning from structural validation, mitigating common failure modes of embedding-only pipelines. Our framework is agnostic to the choice of clustering algorithm and can be applied as a post-hoc refinement layer to existing unsupervised systems.

We evaluate the proposed framework on large-scale social media corpora drawn from two platforms with distinct interaction models: X (formerly Twitter) and Bluesky. While the empirical study focuses on *vegan* discourse—a socially impactful and contested topic—the domain serves primarily as a **testbed** for evaluating the framework under realistic noise, redundancy, and platform variation. To assess reliability in the absence of gold annotations, we conduct a human evaluation with expert annotators, achieving high inter-annotator agreement and demonstrating strong alignment between LLM-generated labels and human interpretations. We further perform robustness analysis under matched temporal and volume conditions to examine the stability of the induced structure across platforms. Our contributions are threefold:

1. We propose a reasoning-based framework for validating and refining the unsupervised semantic structure using LLMs as **semantic judges**.
2. We provide a systematic evaluation, including human validation, comparing reasoning-based refinement with embedding-only approaches.
3. We release cross-platform datasets and evaluation resources to support future research on interpretable and reliable unsupervised text analysis[1].

[1]Our datasets and code are available here: https://github.com/tunazislam/reasoning-based-refinement-llms-vegan

2 Related Work

A large body of previous work studies unsupervised methods for inducing latent semantic structure from text (Boyd-Graber et al., 2014; Blei et al., 2003). Classical probabilistic approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its non-parametric extensions, including HDP (Teh et al., 2004), address the problem of unknown cluster cardinality by allowing the number of latent components to grow with the data. While these methods provide principled statistical formulations, they often struggle with semantic coherence and interpretability, particularly in short and noisy texts such as social media posts (Mimno et al., 2011; Hong and Davison, 2010; Chang et al., 2009).

More recent approaches leverage representation learning to improve unsupervised structure induction. Contextual sentence embeddings (Reimers and Gurevych, 2019) combined with density-based or hierarchical clustering have become common, exemplified by methods such as BERTopic (Grootendorst, 2022). These approaches improve cluster separation in the embedding space, but fundamentally rely on geometric criteria as proxies for semantic validity. As a result, clusters may appear statistically well-formed while remaining semantically incoherent, redundant, or weakly grounded in natural language (Ethayarajh, 2019). Our work departs from representation-centric approaches by focusing on *validating* induced structure rather than improving representations.

Unsupervised structure induction has been widely applied to social media data to study public opinion, discourse dynamics, and emerging narratives (Momeni et al., 2018; Zhao et al., 2011). However, social media text presents persistent challenges for unsupervised methods, including short length, high lexical variation, sarcasm, and rapid topical drift. These properties exacerbate the gap between statistical coherence and human interpretability, limiting the reliability of downstream analyses based on automatically induced themes. Rather than proposing domain-specific heuristics or new topic models, our approach treats clustering outputs as *hypotheses* that require semantic validation.

LLMs have recently been used to support data annotation, weak supervision, and qualitative analysis across a range of NLP tasks (Wang et al., 2023; Ding et al., 2023; Wang et al., 2021). Prior work demonstrates that LLMs can generate labels, summaries, and explanations that align closely with human judgments, enabling scalable annotation in low-resource settings (Islam and Goldwasser, 2025a; Gilardi et al., 2023; Huang et al., 2023). These capabilities have motivated the use of LLMs for theme labeling and content summarization in social science and computational social science research. Most existing approaches, however, employ LLMs as generative annotators or topic induction tools, applying them directly to texts or clusters to produce structure (Brady and Islam, 2025; Pham et al., 2023). Islam and Goldwasser (2025c) used LLM-generated explanations to extract recurring themes and aspects. Unlike recent approaches that use LLMs to directly generate topics or cluster structures, our framework instead uses LLMs as *semantic judges* to validate, refine, and ground the structure produced by unsupervised methods.

A growing line of work explores LLMs-in-the-loop for improving unsupervised or weakly supervised systems (Islam and Goldwasser, 2025d,b; Dai et al., 2023). Islam and Goldwasser (2025b) guided their framework using a seed set of initial themes. Building on this, Islam and Goldwasser (2025d) assumed a predefined set of themes and focused on uncovering underlying arguments. Lam et al. (2024) developed a concept induction algorithm using LLM with human-guided abstraction called LLooM, which has a *seed* operator for accepting user-provided seed term/set. In contrast, our method operates without any initial seed set.

3 Problem Formulation

Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ denote a corpus of unstructured text documents, such as social media posts. The goal of unsupervised semantic structure induction is to organize $\mathcal{D}$ into a set of latent semantic groupings that support downstream analysis and interpretation, without access to labeled data.

In practice, existing unsupervised pipelines typically proceed by applying a clustering algorithm to learned text representations, producing a set of clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$, where K is not known a priori. We refer to each C_k as a *cluster hypothesis*: a candidate grouping proposed by an unsupervised method that may or may not correspond to a coherent semantic theme.

Our focus is not on generating cluster hypotheses, but on addressing a complementary and underexplored problem: *how to validate and refine*

cluster hypotheses in the absence of labeled data. Specifically, given an initial set of clusters $\mathcal{C}$, we aim to produce a refined semantic structure $\mathcal{C}^*$ that satisfies three design goals: (i) **semantic coherence**, where documents within a cluster support a common theme; (ii) **non-redundancy**, where distinct clusters correspond to meaningfully different themes; and (iii) **interpretability**, where each cluster can be grounded in a concise, human-readable description, potentially through consolidation of semantically overlapping label candidates.

3.1 Design Principles

Our framework is guided by three design principles.

Clustering as proposal. We treat unsupervised clustering as a proposal mechanism that generates candidate structure, rather than as a final decision. The choice of clustering algorithm is therefore interchangeable and not central to the framework.

LLMs as semantic judges. Large language models are used not to generate embeddings or clusters, but to reason over natural-language summaries and textual evidence. Given a cluster hypothesis, an LLM evaluates whether the hypothesis is semantically supported by its member documents.

Explicit reasoning checkpoints. We decompose validation into explicit reasoning stages that target specific failure modes of unsupervised clustering, such as incoherence.

3.2 Framework Overview

Fig. 1 illustrates the overall pipeline. The framework takes as input an initial set of cluster hypotheses $\mathcal{C}$ produced by any unsupervised clustering method, along with the underlying documents.

Stage 1: Coherence Verification. For each cluster hypothesis C_k, we first construct a concise natural-language summary that captures its latent theme. An LLM then evaluates whether this summary is supported by the documents in C_k, reasoning over representative examples. Clusters deemed semantically incoherent are discarded from the final cluster set.

Stage 2: Redundancy Adjudication. Even when clusters are individually coherent, multiple clusters may encode overlapping or redundant themes. To address this, we compare summaries of surviving clusters pairwise, and redundant clusters are merged, while genuinely distinct clusters are retained.

Stage 3: Label Grounding. For each refined cluster in $\mathcal{C}^*$, the LLM assigns interpretable labels through a **two-stage** grounding process, where initial candidate labels are generated, and semantically similar labels are consolidated.

The output of the framework is a refined set of clusters $\mathcal{C}^*$ with grounded labels, which can be evaluated using both automatic metrics and human judgment, as described in subsequent sections.

4 Experiments

In this section, we detail the dataset, experimental setup to implement our framework, results, and error analysis.

4.1 Dataset

Prior studies of online discourse on lifestyle choices have largely focused on single, centralized platforms such as X (Islam, 2019). In contrast, we include vegan discourse from both **X** and **Bluesky**, an emerging decentralized social network, to evaluate our framework across platforms with different interaction dynamics.

To extract tweets and posts regarding vegan lifestyle choices, we filter texts containing specific keywords, such as *vegan*, *veganism*, *plantbased*, *meatfree*. The full list of keywords is shown in Table 4 in App. A. For X, we collect tweets using the Tweepy[2] library via the Twitter streaming API subsequentially from October 2019 to February 2020. We have extracted 330464 tweets from 204670 different users. Later, we notice that there are 63751 suspended users. However, in this work, we did not use the whole dataset. To avoid dominance by prolific or automated accounts, we sample at the user level, limiting the number of posts per user. Finally, we have 20000 tweets from 275 different users. This design prioritizes thematic diversity over volume and follows common practice in social media discourse analysis.

For Bluesky, we use a lightweight pipeline to collect and store the posts from the Bluesky firehose[3] in real time. It consists of a *data collector* that connects to the firehose and collects the new posts. We **do not** download *comments* of corresponding posts. We have collected 13032 English posts from June 2025, of which 1752 are unique.

Disclaimer: The datasets analyzed in this study consist of publicly available posts from X and

[2] https://www.tweepy.org/

[3] https://docs.bsky.app/docs/advanced-guides/firehose

Bluesky. These data may contain offensive content or toxic language. The content is used solely for academic analysis and not for the dissemination of harmful material.

4.2 Framework Implementation

4.2.1 Clustering Texts

For the clustering process, we utilized a multi-step approach combining dimensionality reduction, clustering, and hyperparameter search optimization. We begin by clustering the texts using HDBSCAN (McInnes et al., 2017), a nonparametric clustering algorithm. First, the text data is preprocessed and transformed into numerical form using TF-IDF vectorization. While dense embedding models such as GTR-T5 (Ni et al., 2022), E5 (Wang et al., 2022) could serve as alternatives, we adopt TF–IDF to preserve interpretability and lexical transparency in early clustering stages, ensuring that the LLM-based refinement focuses on semantic consolidation rather than embedding quality. We then normalize the sparse matrix representation using MaxAbsScaler to ensure balanced scaling across features. To reduce the dimensionality, Truncated Singular Value Decomposition (SVD) is applied, capturing the most important information from the data in a lower-dimensional space. The UMAP (McInnes et al., 2018) is employed to further reduce dimensions and create embeddings that preserve the local structure of the data. The HDBSCAN is optimized by searching across different parameter configurations (e.g., min_cluster_size, min_samples, cluster_selection_method, metric) using the DBCV (Moulavi et al., 2014) score as an evaluation metric.

4.2.2 Refining the Clusters

To enhance the semantic coherence of the clusters obtained from HDBSCAN, we perform a multi-step refinement process involving LLMs and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019).

Generating Cluster Summaries. For each cluster, we utilize LLM with a zero-shot prompting manner to generate a concise summary from top-k (k = 5) texts assigned to each cluster. To construct representative summaries, we select the **top-5 documents closest to the cluster centroid in embedding space**. This choice reflects a trade-off between representativeness and prompt stability. In pilot experiments, we observe that cluster summaries stabilize beyond $k = 5$, with minimal qualitative changes when additional documents are included. Using a small representative set also helps avoid prompt dilution when clusters contain many heterogeneous texts, ensuring that the most central examples guide the summary generation. Prior work on LLM-assisted discourse analysis (Islam and Goldwasser, 2025d,b) adopts a similar representative-sampling strategy. We also verified robustness across $k \in \{3, 5, 7\}$, observing qualitatively consistent summaries and coherence judgments. Finally, limiting the prompt size reduces latency and API cost while maintaining stable cluster descriptions.

Verifying Cluster Coherence. For coherence verification, the LLM evaluates the semantic alignment between the generated cluster summary and the representative texts used to construct it. The evaluation operates on these representative samples rather than the full set of cluster documents. A cluster is flagged as incoherent when the summary fails to capture a consistent theme across multiple representative texts. In this context, *low semantic alignment* refers to cases where the LLM determines that the summary is not sufficiently supported by the representative documents, indicating that the cluster may contain heterogeneous or weakly related content. Fig. 2 shows an example of an incoherent cluster from our dataset.

Text1: Folic Acid for Pregnant Women. Folic acid is a B vitamin that is found in vegan diets, some kinds of fruits and cereals and animal products.
Text2: Not to be an annoying vegan but you're all fucking disgusting for consuming animal products
Text3: me retreating to my vegan discord after arguing with animal abusers https://t.co/e26X6XUtsf
Text4: @KingDavey1000 @herbivore_club @BackRoa61286135 It's not false, though. You fight what should be your comrades in the fight against animal abuse more than you fight that abuse. Because we're "that type of vegan" doesn't mean we deserve your ire when we're fighting the same cause.
Text5: "Vegan" Clout Crew coming in to defend animal abuse.

Figure 2: Example of the incoherent cluster from X.

Identifying and Merging Redundant Clusters. We identify clusters that are semantically similar and merge them to reduce redundancy. SBERT is used to generate an embedding for each cluster summary. Cosine similarity between cluster embeddings is calculated to identify clusters with high semantic overlap. Clusters exceeding a similarity threshold are merged. We select the appropriate threshold using grid search. Fig. 3 in shows examples of merged cluster summaries from X (Fig. 3a) and Bluesky (Fig. 3b) datasets.

4.2.3 Generating Cluster Label

Each cluster initially produces a single candidate label derived from its summary. After this step, we

Summary of cluster 1: The text encourages starting your day with motivation and inspiration. They emphasize various themes such as money, business, leadership, wisdom, writing, education, beauty, online success, technology, and lifestyle. The hashtags indicate a focus on personal and professional growth, power, style, and lifestyle choices like veganism. The messages are repeated to highlight the importance of these concepts, especially on a Sunday.

Summary of cluster 18: The texts encourage people to get their daily dose of motivation and inspiration before starting the day. It includes a wide array of hashtags related to money, business, leadership, wisdom, writing, education, beauty, online success, technology, veganism, love, life, jobs, style, and several Friday-themed feelings, all targeted towards an audience in the USA.

(a) Merged clusters (X).

Summary of cluster 3: The posts highlight several trends and product releases in the beauty industry with a focus on sustainability. Recent innovations include eco-friendly and cruelty-free nail polish, sustainable skincare from Amazon using recycled materials, and an artisanal Swiss skincare collection made from natural vegan ingredients. Amazon's organic beauty range is noted for its scientific approach and commitment to eco-friendly practices. A repeating theme is the availability of affordable, high-quality makeup. The posts consistently emphasize sustainable beauty and cruelty-free products.

Summary of cluster 10: These posts discuss the emerging trends in the beauty industry focusing on innovation and sustainability. Key points include: Matte Lipstick Trend: There's a shift towards eco-friendly and innovative formulas using plant-based ingredients and cruelty-free certifications. This is transforming the beauty landscape. Amazon's Role: Amazon is noted for its role in revolutionizing clean beauty by offering innovative, cruelty-free products that meet environmental standards and provide skincare benefits. Innovations: New sustainable, plant-based eyeliner formulas highlight the move towards cleaner beauty, emphasizing the importance of cruelty-free ingredients. Organic Beauty Movement: Amazon is shaking up the beauty industry with ethically-sourced skincare and plant-based makeup, encouraging consumers to support

(b) Merged clusters (Bluesky).

Figure 3: Example of merged cluster summaries.

compute pairwise SBERT similarity between generated labels. Labels with similarity above a threshold of 0.85 are grouped to identify semantically redundant themes. For each group of similar labels, the LLM generates a consolidated label representing the merged semantic category. This process allows multiple preliminary labels to be unified into a single final label (Table 6 in App. B.4), forming a **two-stage labeling** procedure consisting of initial label generation followed by semantic consolidation.

4.2.4 Assigning Label to Individual Text

After the final set of consolidated labels is produced, individual documents are reassigned using LLM to ensure alignment between the refined semantic taxonomy and document-level content. This way, we assign the generated label to each text. The LLMs assign the most appropriate label to each text based on its content. Prompts are designed to guide the LLMs in selecting the best-fitting label from the set of generated labels.

4.3 Experimental Details

For the HDBSCAN clustering model, we use a data-driven approach to estimate the best number of topics by maximizing the DBCV score. We retain the default settings for $cluster\_selection\_method$, and $metric$ parameters, while we grid search the $min\_samples$ and $min\_cluster\_size$ to get more sensible topics (Detail in App. B.1).

We grid search the merge similarity threshold $\tau \in \{0.75, 0.80, 0.85, 0.90\}$ using standard metrics: Silhouette Score (S, higher is better) (Rousseeuw, 1987), Davies–Bouldin Index (DB_i, lower is better) (Davies and Bouldin, 2009), and cluster count (C). We select $\tau = 0.85$ as the best trade-off—strong scores with controlled cluster count (see App. B.2 and Table 5 for grid search results).

For the LLM part of implementation, we use GPT-4o[4] (OpenAI, 2024) with the default parameters. Finally, we have 14 and 22 generated labels for X and Bluesky, respectively, after following the steps mentioned in our framework. The generated labels are shown in Table 6 in App. B.4. Label distributions are detailed in App. B.5. Prompt details are provided in App. B.3 in Fig. 6 and Fig. 7. Cost is provided in App. B.6.

4.4 Baselines

To evaluate whether applying the LLM-based refinement process increases the coherence of clusters generated by HDBSCAN, we measure cluster quality, semantic coherence, and conduct statistical significance testing. We have the following two baselines:

HDBSCAN without Refinement: Evaluating the original clusters from HDBSCAN.

SBERT-Based Refinement: Using SBERT for cluster refinement without LLM assistance.

4.4.1 Cluster Quality:

Cluster quality evaluation is essential to verify that a clustering method produces meaningful groupings. To this end, we report both Silhouette Score and Davies–Bouldin Index, two widely used and complementary metrics. As shown in Table 1, LLM-based refinement consistently improves semantic coherence across both platforms, as measured by silhouette scores. In contrast, SBERT-based refinement achieves stronger separation on X according to DB_i, while differences are not statistically significant on Bluesky. Overall, LLM-based refinement improves semantic coherence and human-aligned labeling quality while achieving competitive separation compared to SBERT-based refinement.

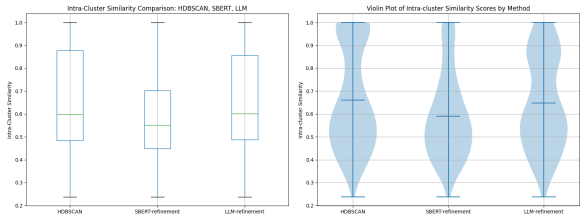
4.4.2 Semantic Coherence:

We measure intra-cluster coherence as the mean cosine similarity of all text pairs in a cluster ($\tau \geq 0.85$

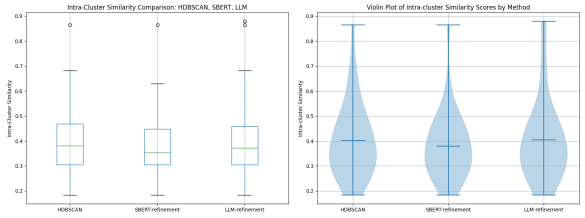
[4] https://openai.com/index/hello-gpt-4o/

Dataset	Metric	HDBSCAN	SBERT-rf	LLM-rf
X	C	359	250	232
	S (↑)	0.122	0.156	**0.674**
	DB_i (↓)	2.322	**0.569**	0.635
Bluesky	C	37	34	36
	S (↑)	-0.017	0.052	**0.979**
	DB_i (↓)	2.739	0.282	**0.227**

Table 1: Cluster quality comparison. rf: refinement.



(a) Box plot (X data). (b) Violin plot (X data).



(c) Box plot (Bluesky data). (d) Violin plot (Bluesky data).

Figure 4: Comparing intra-cluster similarity across HDBSCAN, SBERT-refinement, and LLM-refinement.

on sentence embeddings). Fig. 4 plots the score distributions for X and Bluesky. On X data (Figs. 4a, 4b), HDBSCAN and LLM-refinement yield consistently higher similarity than SBERT-refinement. Their medians are comparable ($\sim$ 0.60), but LLM-refinement slightly edges ahead. Both methods produce many clusters with very high cohesion (0.9–1.0), while SBERT-refinement rarely does. On Bluesky (Figs. 4c, 4d), HDBSCAN and LLM-refinement again align, showing higher medians ($\sim$ 0.38–0.40) and broader upper tails than SBERT-refinement ($\sim$ 0.35). All methods cluster most data in the moderate range (0.3–0.5), with occasional highly cohesive outliers; HDBSCAN and LLM-refinement reach $\sim$ 0.87–0.89, SBERT-refinement lower.

4.4.3 Statistical Significance Test:

To statistically test whether the coherence scores differ significantly between the initial and refined clustering, we compute the difference among the three methods: HDBSCAN w/o refinement, HDBSCAN with SBERT-refinement, and HDBSCAN with LLM-refinement. Table 7 in App. C summarizes the statistical significance analysis of the clustering results for both the X and Bluesky vegan datasets. We first apply a non-parametric Kruskal–Wallis (Kruskal and Wallis, 1952) test to determine whether there are overall differences in the clustering quality metrics. For the X dataset, the Kruskal–Wallis test indicates a significant difference among the methods ($H = 16.187$, $p < 0.001$), prompting post-hoc pairwise comparisons (independent samples) using the Mann–Whitney U (Mann and Whitney, 1947) test. These comparisons reveal that HDBSCAN with SBERT-refinement significantly outperformed both HDBSCAN ($p < 0.001$) and LLM-refinement ($p < 0.01$), while the difference between HDBSCAN and LLM-refinement is not statistically significant ($p = 0.4808$). For Bluesky, the Kruskal–Wallis test does not reveal significant differences across methods, and none of the Mann–Whitney U pairwise comparisons reached significance.

Model	X Acc. (%)	Bluesky Acc.(%)
LDA	30.4	36.2
BERTopic	38.7	42.4
SBERT	56.2	53.6
TopicGPT	72.8	68.4
Llama 3.2	66.6	60.0
Mistral Large 2	71.6	71.8
GPT-4o	**78.4**	**89.8**

Table 2: Assignment comparison w.r.t. human judgment.

4.5 Evaluation

We perform a human evaluation by asking annotators to assess whether the LLM-assigned theme adequately matches the content of **randomly** (random seed=42) selected 500 tweets from X and 500 posts from Bluesky datasets. Two annotators (experts in NLP and CSS) provide the judgment, and the inter-annotator agreement is 0.82 (almost perfect agreement) using Cohen's Kappa coefficient (Cohen, 1960). We use multiple baselines, including SBERT-based assignment, topic modeling: LDA (with 10 topics for X, 5 topics for Bluesky), and BERTopic for comparison. For assignment, we compare three LLMs: GPT-4o, Mistral Large 2 (mistral-large-2407[5]) (Jiang et al., 2023), Llama 3.2 (llama-3.2-90b-text-preview[6]) (Touvron et al., 2023). Table 2 shows the human evaluation results on three LLMs for assignment as well as SBERT assignment and LDA, BERTopic baselines. LLM-based labeling achieved high alignment with human judgments ($\sim$ 90% on Bluesky), outperforming traditional and hierarchical topic modeling

[5] https://mistral.ai/news/mistral-large-2407/
[6] https://www.llama.com/

as well as SBERT baselines. This demonstrates that LLMs can serve as effective unsupervised annotators, enabling scalable theme assignment without costly manual labeling.

Evaluation Fairness and Cluster Pruning. Our refinement process may remove clusters that are deemed incoherent during the coherence verification stage. Importantly, clusters are not discarded to optimize evaluation metrics; rather, a cluster is flagged as incoherent only when its generated summary is not semantically supported by representative member texts. This procedure is applied consistently to all clusters produced by each method before evaluation.

To ensure fair comparison, all evaluation metrics are computed on the final cluster structures produced by each method. We explicitly report the number of clusters (C) in Table 1 to make structural changes transparent. Notably, baseline refinement methods such as SBERT-based refinement also modify cluster cardinality, indicating that changes in cluster count are not unique to our approach.

Furthermore, improvements are not limited to geometric metrics. As shown in Table 2, LLM-based refinement yields substantial gains in human-aligned assignment accuracy, suggesting that improvements reflect better semantic consolidation rather than metric inflation. We also observe that, on the X dataset, the statistical behavior of LLM-refined clusters is comparable to HDBSCAN ($p = 0.48$), and on Bluesky, no method significantly dominates (Table 7 in App. C), indicating that performance gains are not driven by trivial pruning effects.

4.6 Error Analysis

We conduct an error analysis to examine where the best model's (GPT-4o) assigned themes diverge from human judgment. In X, the theme *veganism impacts, challenges, and discussions* overlaps with *advocacy*, *lifestyle*, and *ethics*, causing short personal posts to be misclassified. Food mentions often trigger *dining experiences* labels even for ads or generic content, while positive remarks are sometimes conflated with promotion or activism (e.g., *animal rights*). In Bluesky, abstract themes such as *social and ethical commentary* are frequently misclassified due to implicit moral cues, sarcasm, or vague language. We also observe keyword over-reliance, with mentions of *skincare* or donations misclassified as ethical consumption or advocacy. Error analysis details are provided in App. D.

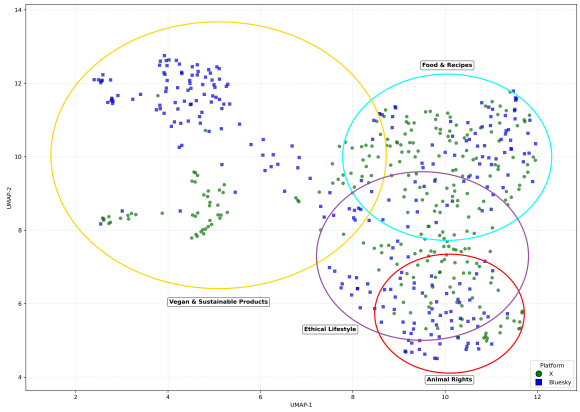


Figure 5: UMAP projection of vegan discourse from X (green) and Bluesky (blue) with broader themes.

5 Qualitative Analysis

To determine whether the differences in linguistic tone across the two platforms can be attributed to differing thematic focuses, we map embedding vectors of 1000 texts from both platforms (same 500 from previously randomly selected from each platform) onto a 2-dimensional projection using UMAP. **To enhance clarity, we show broader themes that are closely situated within the same density region** (Fig. 5). We observe distinct but overlapping clusters: Bluesky posts dominate the 'Vegan & Sustainable Products' region, while both platforms contribute to 'Food & Recipes', 'Ethical Lifestyle', and 'Animal Rights'. These latter themes often blend practical advice with advocacy, reflecting how lifestyle choices are framed as moral commitments. The examples of the posts for each broader theme are shown in Table 3.

Table 8 in App. B.5 highlights example themes that are unique to each platform. On X, discourse centers on informational and aspirational themes (e.g., daily motivation, advocacy), reflecting its broadcast-oriented use. In contrast, Bluesky emphasizes conversational and satirical content, including humor and sociopolitical critique. Theme distributions (Fig. 8 in App. B.5) further show that X foregrounds advocacy, while Bluesky leans towards lifestyle and consumer-oriented themes, highlighting platform-specific discourse patterns.

6 Temporal Drift and Robustness

Our X corpus spans late 2019–early 2020, while Bluesky data is from June'25, raising concerns

Broader Theme	Platform	Example Posts
Animal Rights	X	...Stop this cruelty! #AnimalsLivesMatter #animalcruelty #vegan #animalrights PETITION: Justice for Monkey Hanged from Tree .....
	Bluesky	Wool industry is infested with violence & documented cruelty in nearly every shearing shed that investigators entered. One of the shearers is seen punching sheep in the nose and face. Another is seen jabbing a sheep.....
Food & Recipes	X	...and I are taco lovers & all things Mexico so we can't wait to try Veggie Lad's #vegan recipe...
	Bluesky	If you want some next-level delicious tofu, try out this vegan orange chicken! #vegan #vegansky #veganfood #tofu.....
Vegan & Sustainable Products	X	Why you should be using shampoo bars to have healthier hair and protect the environment #ecofriendly #CrueltyFree #vegan.....
	Bluesky	Delve into the world of Amazon's organic beauty with eco-friendly, cruelty-free products that redefine skincare #OrganicBeauty #SustainableBeauty.....
Ethical Lifestyle	X	Celebrate earth day everyday by going vegan.....
	Bluesky	Veganism doesn't actually mean zero animal byproduct, it's about reducing harm to animals so you can in fact be vegan and consume animal byproduct, it just has to be more ethical...

Table 3: Examples of posts by broader theme.

about confounding historical and platform effects. To address this, we identify the densest 28-day window of X activity (2020.01.14 – 2020.02.10) and down-sample Bluesky to match post count. Cluster quality improves for both platforms under this balanced setup, but the relative pattern sharpens: X shows higher cohesion (silhouette 0.60 vs. 0.08) and tighter separation (Davies–Bouldin 0.61 vs. 1.14). A χ^2 test (Cochran, 1952) on the *theme* $\times$ *platform* table confirms the association is **non-random** ($\chi^2 = 80.0$, $df = 23$, $p \approx 3 \times 10^{-8}$). Hence, our findings are **descriptive** signals that persist after equalizing volume and narrowing time window—**not artifacts of dataset imbalance**. See the App. E for full methodology and visualizations (Fig. 9). We emphasize that these are **descriptive contrasts—not causal estimates of platform design** because external factors (e.g., COVID-19, market shifts) may influence discourse.

7 Broader Societal Impact

This work has meaningful implications for understanding digital public discourse. Our approach offers a powerful tool for *sociotechnical* analysis. This enables researchers, policymakers, and stakeholders to gain timely insights into understanding trends, evolving public sentiments, and consumer behavior at scale.

Our case study illustrates the importance of analyzing narratives across both centralized (X) and decentralized (Bluesky) platforms. We show that different platforms amplify distinct facets of the same movement—ranging from `ethical advocacy` to `humor`, `product marketing`, and `community identity`. These distinctions reflect broader shifts in how digital infrastructure shapes discourse, activism, and ideology.

Our findings offer actionable insights for stakeholder groups such as vegan advocacy organizations, e.g., PETA[7], TVS[8], and consumer protection watchdogs, e.g., FTC[9]. For instance, an advocacy group aiming to mobilize support for `animal rights` might prioritize Bluesky's *community-driven environment*, while consumer protection agencies concerned with greenwashing could scrutinize high-volume *product promotions* on X. By highlighting differences in thematic structure and coherence, our framework can inform where and how to engage diverse online audiences.

8 Conclusion and Discussion

We introduce a reasoning-based framework that uses LLMs as semantic judges to validate and refine unsupervised text clusters. Rather than inducing structure from scratch, the approach improves semantic coherence and human-aligned labeling while maintaining competitive separation relative to strong embedding-based baselines. Experiments with cross-platform social media data and robustness analysis demonstrate the reliability of the induced structure.

Our method is particularly beneficial in settings where initial clustering produces semantically noisy or redundant structures, such as short-text or high-variation corpora. When clustering quality is already near-optimal, refinement yields smaller gains, suggesting that the framework is most impactful as a semantic validation layer rather than a replacement for strong clustering methods.

[7] https://www.peta.org/
[8] https://www.vegansociety.com/
[9] https://www.ftc.gov/

9 Limitations

Our experiments focus on English-language social media data and a specific discourse domain (veganism), which may limit direct generalization to other languages or domains. While the framework itself is domain-agnostic, applying it to new settings may require adapting prompts or evaluation criteria. Finally, although human validation demonstrates strong alignment with LLM-based labeling, we **do not** claim that LLM judgments fully replace expert annotation in high-stakes settings.

In our experiment, as the X data predates Bluesky by ∼5 years, our analysis offers descriptive—not causal—contrasts. The 28-day subsample narrows but does not remove this gap—because residual historical confounds (e.g., COVID-19, market shifts) cannot be ruled out without contemporaneous data from both platforms. Future work might collect 2025 X sample to eliminate this residual confound.

Additionally, as LLMs are trained on extensive human-generated text, they may embed human biases (Islam, 2026; Blodgett et al., 2020; Brown et al., 2020), which are not addressed in this study. We only used pre-trained LLMs and did not consider fine-tuning due to the resource constraints.

We adopt TF–IDF to preserve interpretability and lexical transparency in early clustering stages, though dense embedding models could serve as alternatives. Future work could replace TF–IDF with dense encoders such as E5 or GTR-T5 to evaluate whether dense initializations further enhance cluster coherence in cross-platform discourse analysis.

We emphasize that our work does not propose a new topic model nor generate topics directly with LLMs; instead, it focuses on post-hoc semantic validation of arbitrary unsupervised clustering methods. Our current evaluation focuses on intra-cluster embedding similarity, Silhouette Score, Davies–Bouldin Index, and human validation. Incorporating document-level coherence metrics (Rahimi et al., 2024; Korenčić et al., 2018; Ramrakhiyani et al., 2017) adapted for clustering outputs can be explored as future work.

10 Ethical Considerations

To the best of our knowledge, we did not violate any ethical code while conducting the research work described in this paper. We report the technical details for the reproducibility of the results. The author's personal views are not represented in any results we report, as it is solely outcomes derived from machine learning or AI models.

The social media data used in this study may contain offensive, biased, toxic, or harmful language. Such content reflects user-generated discourse and does not represent the views of the authors or institutions. All analyses were conducted for research purposes only.

11 Acknowledgments

We would like to thank Lightning AI Studio for providing the computing resources. Also, we are thankful to the anonymous reviewers for their thoughtful suggestions.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*.

Su Lin Blodgett and 1 others. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*.

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Alexander Brady and Tunazzina Islam. 2025. Latent topic synthesis: Leveraging llms for electoral ad analysis. *arXiv preprint arXiv:2510.15125*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

William G Cochran. 1952. The $\chi 2$ test of goodness of fit. *The Annals of mathematical statistics*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *EPM*.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*.

David L Davies and Donald W Bouldin. 2009. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Patrik O Hoyer. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Tunazzina Islam. 2019. Yoga-veganism: Correlation mining of twitter health data. *arXiv preprint arXiv:1906.07668*.

Tunazzina Islam. 2026. Who gets which message? auditing demographic bias in llm-generated targeted text. *arXiv preprint arXiv:2601.17172*.

Tunazzina Islam and Dan Goldwasser. 2025a. Can llms assist annotators in identifying morality frames?-case study on vaccination debate on social media. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 169–178.

Tunazzina Islam and Dan Goldwasser. 2025b. Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating llms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 859–884.

Tunazzina Islam and Dan Goldwasser. 2025c. Post-hoc study of climate microtargeting on social media ads with LLMs: Thematic insights and fairness evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15838–15859, Suzhou, China. Association for Computational Linguistics.

Tunazzina Islam and Dan Goldwasser. 2025d. Uncovering latent arguments in social media messaging by employing llms-in-the-loop strategy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7397–7429.

Albert Q Jiang, , and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Damir Korenčić, Strahil Ristov, and Jan Šnajder. 2018. Document-based topic coherence measures for news media text. *Expert systems with Applications*, 114:357–373.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Leland McInnes, John Healy, Steve Astels, and 1 others. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *JOSS*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Elaheh Momeni, Shanika Karunasekera, Palash Goyal, and Kristina Lerman. 2018. Modeling evolution of topics in large-scale temporal text corpora. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. 2014. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and 1 others. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

OpenAI. 2024. Hello gpt-4o. *https://openai.com/index/hello-gpt-4o/, 2024*.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian's, Malta. Association for Computational Linguistics.

Nitin Ramrakhiyani, Sachin Pawar, Swapnil Hingmire, and Girish Palshikar. 2017. Measuring topic coherence through optimal word buckets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 437–442, Valencia, Spain. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

PK Srijith, Mark Hepple, Kalina Bontcheva, and Daniel Preotiuc-Pietro. 2017. Sub-story detection in twitter with hierarchical dirichlet processes. *Information Processing & Management*, 53(4):989–1003.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17.

Hugo Touvron and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Michele Zappavigna. 2012. Discourse of twitter and social media. *Discourse of Twitter and Social Media*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.

A Data Collection

The full list of keywords can be seen in Table 4.

vegandiet, veganfood, veganlife, veganlover, veganlifestyle, vegancommunity, veganfoodshare, vegans, veganfoodlove, veganjourney, plant-based, cruelty-free, dairyfree, crueltyfree, meat-free, govegan, veganfoodporn, animal rights.

Table 4: List of the keywords for data collection.

B Experiments

B.1 HDBSCAN Hyperparameters

We grid search for the number of $min\_samples$ on $\{2, 3, 5, 10\}$. The $min\_cluster\_size$ number is selected based on a grid search whose values are sensitive to the number of input data points. Suppose $|D|$ denote the number of data points, then the grid parameters

Dataset	τ (C, S, DB_i)
X	0.75 (126, 0.378, 1.037), 0.80 (174, 0.519, 0.838), **0.85 (232, 0.674, 0.635)**, 0.90 (292, 0.825, 0.493)
Bluesky	0.75 (22, 0.589, 0.852), 0.80 (32, 0.859, 0.498), **0.85 (36, 0.979, 0.227)**, 0.90 (37, 1.0, $6.2e-08$)

Table 5: Threshold selection using grid search.

for HDBSCAN used in our method include $\{5,10,15,0.05\times|D|,0.1\times|D|,0.2\times|D|,0.25\times|D|\}$. We set the $n\_neighbors$ parameter in UMAP embedding to $min\_cluster\_size$. In our framework, for X, the best parameters are $min\_cluster\_size$: 15, $min\_samples$: 3 to obtain the DBCV score of 0.35. For Bluesky, the best parameters are $min\_cluster\_size$: 10, $min\_samples$: 2 to obtain the DBCV score of 0.53.

B.2 Cluster Threshold Selection

We grid search the merge similarity threshold $\tau \in \{0.75, 0.80, 0.85, 0.90\}$ and evaluate each setting using Silhouette Score (S), Davies–Bouldin Index (DB_i), and the resulting number of clusters (C). S: higher means points fit their own cluster better than others. DB_i: lower means clusters are compact and well separated. Results are shown in Table 5. While quality metrics improve with higher τ, 0.90 yields an explosion in small, redundant clusters that hurt interpretability and summarization. We therefore select 0.85 as the best trade-off—strong scores with controlled cluster count.

B.3 Prompt Design

Prompt templates used in our work are illustrated in Fig. 6. To summarize the top-5 texts, the prompt template is shown in Fig. 6(a). Fig. 6(b) provides a prompt template for checking cluster coherency, and Fig. 6(c) represents a prompt template for generating a cluster summary label. Prompt template for assigning a label (from the list of generated summary labels) to individual text is shown in Fig. 6 (d). Fig. 7 shows the prompt example of assigning text to the label.

B.4 Generated Labels

Table 6 shows the generated labels by our framework.

Generate a brief, clear and concise summary (in 100 words) from the following texts:
text1: **
text2: **
text3: **
text4: **
text5: **

(a) Prompt template for generating summary.

Do the following five texts share a common theme with this summary? Please answer with a 'yes' or 'no'.
text1: **
text2: **
text3: **
text4: **
text5: **
Summary: **

(b) Prompt template for cluster coherency check.

Based on the following summary of texts, what is the most appropriate label for this cluster summary?
Cluster Summary: **

(c) Prompt template for generating cluster summary label.

Assign one of the following themes to the text based on its content.
Themes: **{list of generated summary labels}**
Text: **

(d) Prompt template for assignment.

Figure 6: Prompt templates (shown as zero-shot).

B.5 Theme Distribution Analysis

Fig. 8 presents a comparative analysis of the distribution of assigned themes (generated by GPT-4o) across two social media platforms—X (formerly Twitter) and Bluesky—as part of our study on vegan-related discourse.

Platform X (Left Bar Chart). The discourse on X is concentrated around a few dominant themes: 'Veganism impacts, challenges and discussions', 'Veganism advocacy and animal rights', 'Daily motivation and inspiration'. These themes reflect a strong focus on ideological, motivational, and advocacy-related content, aligning with X's identity as a platform for public debate and activism. Although there is some thematic diversity, long-tail topics such as *Vegan birthday celebrations* or *None of the above* appear infrequently.

Bluesky (Right Bar Chart). Bluesky shows a more top-heavy distribution, with a focus on lifestyle and product-related themes: 'Sustainable and ethical beauty products', 'Humorous perspectives on vegan lifestyles and food', 'Vegan cooking recipes and cuisine'. This indicates that Bluesky users engage more with practical, day-to-day content rather than advocacy or ideological discussions. The themes also display greater granularity, including niche interests such as *Tour announcements*, *Pet products*, and *Plant-based lifestyle and acceptance*.

These differences suggest distinct discursive cultures:

Assign one of the following themes to the text based on its content.

List of Themes: 'Sustainable and Ethical Beauty Products', 'Veganism and Ethical Living', 'Vegan Cooking, Recipes and Cuisine', 'Humorous Perspectives on Vegan Lifestyles and Food', 'Vegan Snacks and Treats', 'Nutrition and Healthy Living', 'Pet Products and Vegan Options on Amazon', 'Animal Rights and Advocacy', 'Vegan Birthday Celebrations', 'Vegan Products and Community', 'Critiques and Humor in Social and Political Contexts', 'Vegan and Plant-Based Trends', 'Tour Announcements', 'New Music Releases', 'The Benefits and Innovations in Plant-Based Solutions', 'Natural Health Supplements', 'Vegan Lifestyle and Digital Projects', 'Social and Ethical Commentary', 'Plant-based lifestyle and acceptance', 'Vegan Lifestyle and Events', 'Plant-Based Living and Community Advocacy', 'Vegan Lifestyle and Media'

Text: This is a f***ing horror show 🩸 Animals butchered like objects. Workers trapped in trauma just to survive. All for cheap flesh on your plate. This isn't food — it's violence. Wake up. Opt out. Burn this system down. #MeatIsMurder #AnimalExploitation #GoVegan #NoMoreExcuses

Theme: Animal rights and advocacy

Figure 7: Prompt example of mapping text→theme (Bluesky dataset). The *black* colored segment is the *input* prompt and the *blue* colored segment is the generated *output* by LLMs.

Dataset	Labels
X	Daily Motivation and Inspiration, Gluten-Free and Vegan Publication Updates, Veganism Advocacy and Lifestyle Promotion, Vegan and Vegetarian Recipes and Cookbook, Veganism and Plant-Based Ethical Lifestyle, Veganism Advocacy and Animal Rights, Vegan Food and Lifestyle Celebrations, Vegan Desserts Recipe, Handmade Vegan Soaps Promotion, Veganism Impacts, Challenges, and Discussions, Vegan and Vegetarian Dining Experiences, Vegan and Gluten-Free Food Promotions, Promotion of Low-Sugar Vegan Tea Products, Promotion of Vegan Haircare Products.
Bluesky	Sustainable and Ethical Beauty Products, Veganism and Ethical Living, Vegan Cooking, Recipes and Cuisine, Humorous Perspectives on Vegan Lifestyles and Food, Vegan Snacks and Treats, Nutrition and Healthy Living, Pet Products and Vegan Options on Amazon, Animal Rights and Advocacy, Vegan Birthday Celebrations, Vegan Products and Community, Critiques and Humor in Social and Political Contexts, Vegan and Plant-Based Trends, Tour Announcements, New Music Releases, The Benefits and Innovations in Plant-Based Solutions, Natural Health Supplements, Vegan Lifestyle and Digital Projects, Social and Ethical Commentary, Plant-based lifestyle and acceptance, Vegan Lifestyle and Events, Plant-Based Living and Community Advocacy, Vegan Lifestyle and Media.

Table 6: Generated labels by LLMs from X and Bluesky data.

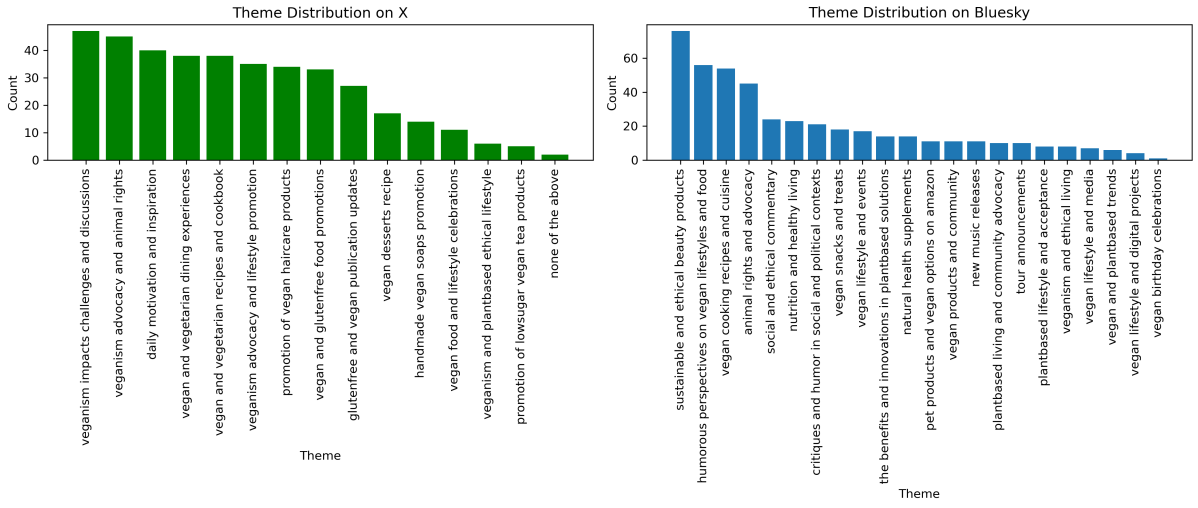


Figure 8: Distribution of assigned labels/themes using GPT-4o on X (green) and Bluesky (blue) datasets.

Dataset	Test / Comparison	Statistic	p-value	Significance
X	Kruskal–Wallis H-statistic	16.187	0.0003	$p < 0.001$
	Mann–Whitney U: HDBSCAN vs LLM	42706.0	0.4808	n.s.
	Mann–Whitney U: HDBSCAN vs SBERT	52922.5	0.000165	$p < 0.001$
	Mann–Whitney U: SBERT vs LLM	24018.0	0.001829	$p < 0.01$
Bluesky	Kruskal-Wallis H-statistic	0.2809	0.8690	n.s.
	Mann–Whitney U: HDBSCAN vs LLM	680.0	0.8816	n.s.
	Mann–Whitney U: HDBSCAN vs SBERT	674.5	0.6044	n.s.
	Mann–Whitney U: SBERT vs LLM	582.0	0.7289	n.s.

Table 7: Statistical test results for cluster metric comparisons. n.s. = statistically not significant (p-value > 0.05).

- **X** favors advocacy-driven and community-mobilizing narratives.

- **Bluesky** promotes a lifestyle-oriented discourse, with greater emphasis on ethical consumerism and humor.

Such findings support the hypothesis that platform audience composition significantly influences the thematic framing of social media discourse.

Table 8 shows the themes that are unique to each platform. On X, posts often center on 'Daily Motivation and Inspiration' and 'Gluten-Free Vegan Publication Updates', reflecting its role as a hub for *informational and aspirational content*. In contrast, Bluesky features 'Humorous Perspectives on Vegan Lifestyles' and 'Critiques in Social and Political Contexts', showing a more *conversational and satirical* tone. From theme distributions (Fig. 8), we notice that X emphasizes advocacy and motivational discourse (e.g., 'Veganism advocacy and animal rights'), while Bluesky leans toward lifestyle and consumer themes (e.g., 'Sustainable beauty products', 'Vegan recipes'). These distinctions underscore how each platform fosters different facets of vegan discourse.

B.6 Experimental Cost

Overall, in our experiment, for X data, GPT-4o costs $\approx$ \$22, and for Bluesky data, GPT-4o costs $\approx$ \$14. Latency (per $1k$ posts) $\approx 6 - 10$ minutes.

C Statistical Tests

Table 7 summarizes the statistical significance analysis of the clustering results for both the X and Bluesky vegan datasets. We first apply a non-parametric Kruskal–Wallis test to determine whether there are overall differences in the clustering quality metrics (Silhouette Score and Davies–Bouldin Score). Then we do pairwise comparisons (independent samples) using the Mann–Whitney U test.

D Error Analysis Details

For X, the top-3 most frequent themes associated with errors are as follows: 'veganism impacts, challenges, and discussions' (12 times), 'vegan and vegetarian dining experiences' (11 times), 'veganism advocacy and lifestyle promotion' (11 times). For Bluesky, the top-3 most frequent themes associated with errors are as follows: 'social and ethical commentary' (8 times), 'sustainable and ethical beauty products' (4 times), 'animal rights and advocacy' (4 times). In Table 9, we show a few cases where the themes assigned by GPT-4o do not align with human evaluations.

D.1 Error Analysis: X Data

Possible reasons for misclassifications of 'veganism impacts, challenges, and discussions': This theme might overlap with others like *advocacy*, *lifestyle*, and *ethics*. The model may be assigning this category when the content is not analytical or discussion-based, e.g., a short personal story.

Platform	Unique Theme	Example Posts
X	Daily Motivation and Inspiration	5 BEST EXERCISES FOR PEOPLE w/ BACK PAIN #fitness #nutrition #vegan #mindfulness #motivation #inspiration #SaturdayMotivation.....
X	Gluten-Free and Vegan Publication Updates	@TheHecticVegan: The Hectic Vegan Magazine Issue 5 #glutenfree #vegan
Bluesky	Humorous Perspectives on Vegan Lifestyles and Food	Vegan chefs consider it a huge insult when you cut yourself & drip your own blood on to their burgers to enhance the flavor.
Bluesky	Critiques and Humor in Social and Political Contexts	The funny thing about vegans is that they think they're morally superior for not eating meat because of deforestation when their entire diet is based on exploiting farmers get f****d (moral) vegans.

Table 8: Examples of platform-exclusive themes and representative posts.

Possible reasons for misclassifications of 'vegan and vegetarian dining experiences': Confusion between promotion and personal experience. Overgeneralization: GPT might label any mention of food or restaurants under this, even if the content is not about personal experiences (e.g., advertisements or random food mentions). False positive on food references: Posts like `I'd prob cry if anyone ever cooked vegan for me` or `hosting vegan cooking class` should not qualify as a *dining experience*.

Possible reasons for misclassifications of 'veganism advocacy and lifestyle promotion': GPT-4o may assign this theme to any positive vegan content, even if it's not clearly promotional or advocacy-related. Posts like `People who act like they can't enjoy vegetarian meals because they eat meat are so annoying`. On the other hand, the model might confuse with more cause-driven themes like animal rights or activism, such as posts like `@Sydney843 @QUBFoodProf @DiscoStew66 Organic vegetables almost always use animal byproducts in fertilisers and therefore are not vegan. Vegans should only ever eat non-organic unless from explicitly from a vegan farm.`

Additionally, for very few instances, LLMs provide a 'None of the above' response when a user posts about an unrelated topic but includes the hashtag #vegan.

D.2 Error Analysis: Bluesky Data

The theme 'social and ethical commentary' is prone to misclassifications due to its abstract and context-dependent nature. Posts under this theme often involve implicit moral reasoning, sarcasm, or indirect critique, which GPT-4o may struggle to interpret without deeper discourse understanding. For instance, vague or philosophical remarks like `If only humans were rational beings` can be easily mistaken for general lifestyle reflections or advocacy, especially when accompanied by hashtags such as #plantbased. Moreover, the absence of explicit topical anchors (e.g., product names, events, or actions) makes it difficult for the model to confidently associate the content with ethical or societal discourse.

We notice that posts mentioning skincare, gel, or beauty are classified as 'sustainable and ethical beauty products', showing the model's keyword over-reliance—ads and self-care tips with no ethical or sustainability angle.

Another four errors involved with 'animal rights and advocacy' where the model treats any emotional plea or donation link (e.g., `. . . please, my friend, donate if you can`) as full-blown activism, conflating charity appeals with broader advocacy.

Moreover, posts dominated by hashtags with minimal textual context also led to misclassifications; context-light, hashtag-heavy content encouraged overreliance on surface cues.

E Temporal Drift & Robustness Analysis

We quantify temporal drift by plotting monthly volume and correlating LLM-assigned themes with posting month (Table 10). Results confirm that several themes (e.g., 'vegan hair-care promotions', 'veganism advocacy and animal rights') peaked in late 2019, whereas Bluesky's June'25 content focuses on 'plant-based ethics'.

Because our X corpus covers Oct'19–Feb'20 while the Bluesky crawl is restricted to June'25, direct comparison risks conflating platform and historical effects. To gauge the impact of this mismatch, we extract the densest 28-day window of X activity (2020.01.14 – 2020.02.10) by applying a rolling 28-day sum over daily post counts and selecting the peak. We then down-sample Bluesky to the same number of posts. Next, we recompute the clustering quality for each platform.

Table 11 shows that the relative pattern persists—and even sharpens. X remains far more cohesive than Bluesky: silhouette score increases from

Platform	Theme	Example Posts
X	veganism impacts, challenges, and discussions	.... y'all all the f**king vegan stables are gone. no rice, no beans, no bread, no chips, no oat milk, no kombucha, no peanut.
		I keep striking out in the meat sections of my local grocery stores. This pandemic just might force me into veganism. https://t.co/EwsFYindAn
X	vegan and vegetarian dining experiences	... I'd prob cry if anyone ever cooked vegan for me.
		We look forward to hosting our vegetarian cooking class tonight! See you at 18h00! #FreshEarth #CookingClasses http://t.co/WpSz5Uqfki
X	veganism advocacy and lifestyle promotion	People who act like they can't enjoy vegetarian meals because they eat meat are so annoying.
		@Sydney843 @QUBFoodProf @DiscoStew66 Organic vegetables almost always use animal byproducts in fertilisers and therefore are not vegan. Vegans should only ever eat non-organic unless from explicitly from a vegan farm.
Bluesky	social and ethical commentary	If only humans were rational beings ... #plantbased theconversation.com/why-theres-a...
		vegans for sure but for second place it def depends i find mostly other people mention the height unless it's someone under the age of like 23 and it for the bi/mulitracial depends on the mix and which parent is which lmao
Bluesky	sustainable and ethical beauty products	HybridGel Fill #EnailCouture #NailzByDragon #DragonzClaw #VeganNailTech #NailTech #VeganNails #208VeganNailTech #BoiseNailTech #BoiseNails #BoiseIdaho #UniqueNails #HybridGelNails #AcrylicNails #Manicure #Pedicure #GelManicure #GelPedicure #HybridGel #Polygel #HappyGel #AcrylGel
		Seven of Limes is back! This #StarTrekDiscovery - inspired soap was a big hit last year. Watch these citrusy swirls all come together here: youtu.be/hcJ93curg_0? #soapmaking #vegansoap #howitsmade #sevenoflimes
Bluesky	animal rights and advocacy	@afivegantenna.bsky.social Pls my friend donate if you can and Share the post with the donation link. bsky.app/profile/9ahm
		@liberalvegan.bsky.social bsky.app/profile/abed... Please, my friend, donate if you can and write a quote.

Table 9: Example posts from X and Bluesky where GPT-4o's assigned themes **do not** align with human judgment.

Theme	pearson_corr_with_month
Promotion of vegan haircare products	0.771
Veganism advocacy and animal rights	0.560
Veganism advocacy and lifestyle promotion	0.495
Vegan and glutenfree food promotions	0.459
Veganism impacts challenges and discussions	0.440
Vegan and vegetarian dining experiences	0.352
Handmade vegan soaps promotion	0.350
Vegan and vegetarian recipes and cookbook	0.277
Veganism and plantbased ethical lifestyle	0.250
Glutenfree and vegan publication updates	0.232
Vegan food and lifestyle celebrations	0.226
Vegan desserts recipe	0.185
Daily motivation and inspiration	0.136
Promotion of lowsugar vegan tea products	0.070

Table 10: Theme–time correlation on X.

$0.40 \rightarrow 0.60$ for X but remains low for Bluesky ($0.15 \rightarrow 0.08$). Separation improves once volume is equalized. Davies-Bouldin scores drop substantially for both platforms ($\approx 3 \rightarrow 1$), yet X still exhibits tighter clustering ($0.61 < 1.14$). Differences are not driven by sheer data volume or longer time span on X; instead, they reflect how discourse is organized.

We plot the distribution of themes (Fig. 9) in the balanced snapshot. We visualize the topical mix in the balanced snapshot by counting posts per GPT-4o assigned theme and coloring bars by platform (X vs. Bluesky); totals are equal across platforms by construction. Fig. 9 shows a strongly heavy-tailed (Zipf-like) distribution: a single promotional theme—'promotion of vegan hair-care products'—dominates the X slice, while Bluesky contributions are diffuse across several themes such as 'sustainable & ethical beauty products', 'animal rights & advocacy', and 'humorous perspectives on vegan lifestyles'. In other words, even after equalizing volume and narrowing the time window, the platforms exhibit different topical mixes: X concentrates attention in one large, commerce-oriented theme; Bluesky spreads attention across multiple smaller, lifestyle/advocacy themes. This pattern is consistent with our qualitative reading and aligns with the per-platform cohesion scores: X's large promotional cluster yields higher within-theme lexical cohesion, whereas Bluesky's broader mix produces many small clusters. We treat these as **descriptive contrasts**, **not causal claims about platform design**.

Then we perform a chi-square test (Cochran, 1952) on the theme×platform contingency table. A χ^2 test of independence on the *theme* $\times$ *platform* contingency table confirms a non-random association between discourse themes and platform in the balanced snapshot ($\chi^2 = 80.0$, $df = 23$, $p \approx 3.18 \times 10^{-8}$).

Corpus	Platform	#Posts	#Clusters	Silhouette (↑)	Davies–Bouldin (↓)
Full	X	939	14	0.40	3.12
	Bluesky	2479	21	0.15	3.06
Balanced 28-day snapshot	X	40	7	0.60	0.61
	Bluesky	40	17	0.08	1.14

Table 11: Cluster quality metrics by platform and corpus.

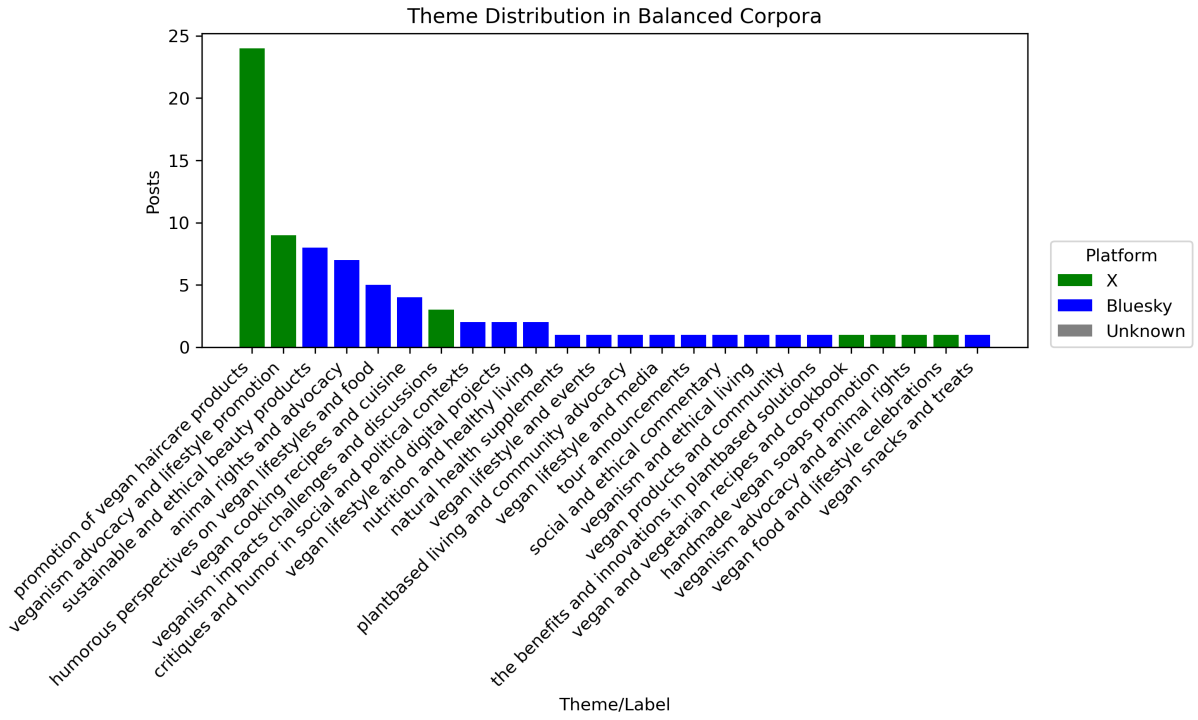


Figure 9: Theme distribution in balanced corpora.