Comparative Analysis of Paid and Free Articles on Medium.com

Florian Eder, Moritz Enderle

October 21, 2025

Contents

Abstract

TODO: Write an abstract

1 Data Acquisition

There is currently no publicly available dataset of articles on Medium.com we could use for our analysis. The platform also does not provide an official API for data access. Therefore, we had to build a custom data acquisition pipeline to scrape articles from Medium.com, focusing on both paid and free content.

1.1 Legal Considerations

As this process involves web scraping, which is strictly prohibited by Medium's Terms of Service, we got in touch with Medium's legal department to clarify the situation and obtain permission for our academic project.

Although we received permission to proceed, we ensured our scraping methodology does not overload or negatively impact Medium's services. This includes:

- Respectful crawling delays between requests
- Compliance with robots.txt directives where applicable
- Limiting the total number of scraped articles to a reasonable amount

All collected data is used exclusively for academic research purposes and will not be redistributed or commercialized.

1.2 Page Discovery

Traditional content discovery methods are mostly based around spiders crawling links from one page to another. However, this approach would introduce a significant bias, as most articles on Medium.com are linked only to other articles within the same field (e.g. technology articles link to other technology articles). This would lead to a dataset that is not representative of the overall article distribution on Medium.com.

This led us to discover Medium's sitemaps as a more suitable approach for page discovery. The main sitemap is located at `https://medium.com/sitemap/sitemap.xml` and contains references to 20440 individual sitemaps, each containing up to 7000 URLs. In total, this results in 32 million URLs [as of March 2025]. We completed the whole process over 6 hours with a Gaussian distributed delay with mean 0.05 seconds. From these URLs, we were able to filter out a portion of non-article URLs (e.g. user profiles, tag pages, etc.). This finally resulted in 14702 individual sitemaps and a total of just over 30 Million URLs.

1.3 Scraping the Articles

The articles on Medium.com are dynamically loaded using JavaScript, which means a simple HTTP request to fetch the HTML content is not sufficient. To overcome this issue, employed

playwright, a high level APIU for browser control, to fully render the pages before extracting the relevant data.

Due to the high number of articles to be scraped, we parallelized the process using multiple worker processes. Each worker operates independently, fetching URLs from the database, rendering the pages, and extracting the relevant data.

We extracted most data directly from the rendered HTML using CSS selectors. However, some data points are embedded in JSON-LD structured data within the page, which we parsed to extract additional metadata.

For member-only (paid) articles, we implemented a method to access the full content by purchasing the Medium membership and using the associated cookies during scraping.

1.4 Database Storage

We stored data in DuckDB, a columnar database optimized for analytical queries. We designed the schema to include tables for sitemaps, URLs, articles, authors, and comments, with relationships maintained through foreign keys. This enables efficient querying for our comparative analysis, as illustrated in Figure 1.

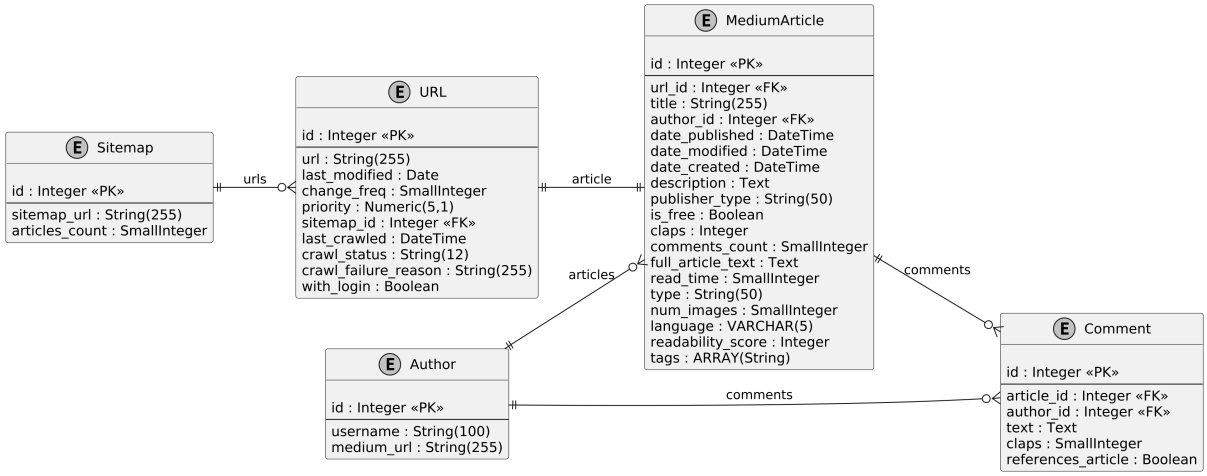

Figure 1: Entity-relationship diagram of the scraping data model.

1.5 Monitoring

We implemented telemetry to track scraping health via structured worker logs aggregated into daily roll-ups for latency, error, and retry trends. We also configured the pipeline to push batch metrics to Weights & Biases such as success ratios, premium hit rates, and fed into dashboards with alert thresholds.

2 Data Analysis

2.1 Dataset Overview

Our initial scraping process, initiated from the 32 million URLs derived from the sitemaps, culminated in a raw dataset of **65, 248** fully scraped articles.

Ensuring a Fair Comparison

Medium.com introduced the ability for paid memberships in early 2017. As we want to ensure comparability between free and paid articles, taking into account any articles published before this date would introduce a significant bias. Furthermore, to allow for sufficient adoption of the commercial plan by authors and stabilize publication patterns, we only consider articles published after the first of January 2020. This filtering is illustrated in Figure 2, which shows the distribution of articles per month from 2020 to 2025, highlighting the increasing proportion of paid content over time.

This filtering step resulted in a final analysis dataset of **33, 510** articles published between January 2020 and May 2025, contributed by **24, 639** unique authors, and containing a total of **83, 064** responses.

Of this final, cleaned dataset, **33.6**% (**11, 262** articles) were classified as member-only (paid), and **66.4**% (**22, 248** articles) were free. This composition provides a strong basis for comparative statistical analysis. The primary data points successfully extracted for this analysis include the article's text, estimated reading time, clap count, response count, author follower count, and premium status.

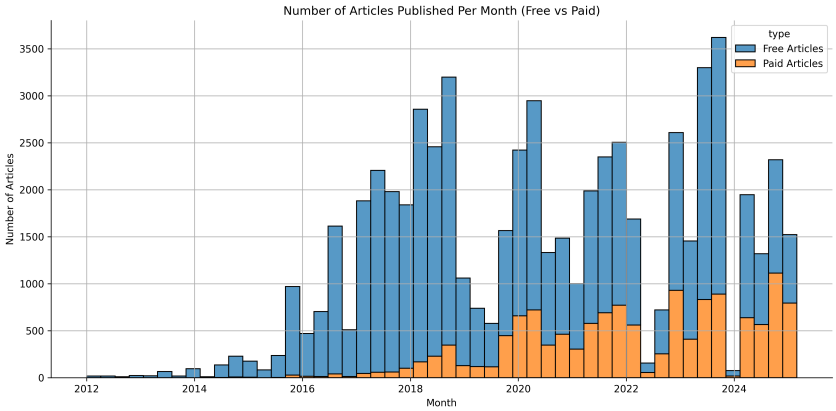

Figure 2: Distribution of scraped articles per month (2020-2025), segmented by premium status. The chart illustrates the growing relative proportion of member-only content in the dataset over time.

2.2 Descriptive Statistics

Table 1 presents the core descriptive statistics for the engagement and length metrics, segmented by the article's premium status. These statistics highlight a clear difference in average engagement, particularly in clap count, where paid articles appear to significantly outperform free articles.

Table 1: Descriptive Statistics for Key Metrics (Post-Jan 2020)

Metric	Status	N	Mean	Median	Std. Dev.
Clap Count	Paid	11,262	407.9	207.0	519.0
	Free	22,248	122.4	40.0	272.1
Response Count	Paid	11,262	4.9	2.0	7.0
	Free	22,248	1.2	0.0	3.1
Reading Time (min)	Paid	11,262	6.0	5.0	2.9
	Free	22,248	5.9	5.0	3.2

2.2.1 Hypothesis Testing (Two-Sample t-tests)

Due to the observed differences in mean engagement metrics and reading time, we performed independent two-sample t-tests to formally assess the statistical significance of these differences. Given the large sample sizes, the Central Limit Theorem allows us to proceed with t-tests despite the non-normality and heteroscedasticity of the underlying populations, focusing on differences in the sample means.

The null hypothesis (H_0) for each test is that there is no difference between the means of paid and free articles for a given metric ($\mu_{paid} = \mu_{free}$).

The results of the t-tests, presented in Table 2, reveal statistically significant differences between paid and free articles for all examined metrics. For clap count and response count, the t-statistics exceed 50, with p-values effectively at zero, indicating extremely strong evidence against the null hypothesis of no difference in means. This suggests that paid articles have substantially higher levels of engagement compared to free articles. Conversely, for reading time, the t-statistic of 2.26 yields a p-value of 0.0235, which, while below the conventional 0.05 threshold, indicates a marginally significant difference, implying that paid articles are, on average, slightly longer in reading time than free ones. These findings underscore the potential impact of premium status on content engagement and length within the Medium.com ecosystem.

Table 2: Two-Sample T-Test Results for Key Metrics

Metric	t-statistic	p-value
Clap Count	54.690392	< 0.001
Response Count	53.863954	< 0.001
Reading Time (min)	2.264893	0.023528

TODO: maybe extend this part?

3 Topic Modeling

3.1 Introduction

Topic modeling using embeddings represents a modern approach to uncovering thematic structures in text data, leveraging dense vector representations to capture semantic similarities. Unlike traditional methods like Latent Dirichlet Allocation (LDA), embedding-based techniques

utilize pre-trained language models to generate contextual embeddings, allowing for more nuanced topic discovery that accounts for word polysemy and context.

In our analysis of Medium.com articles, we use embedding-based topic modeling to compare thematic distributions between paid and free content. This method also enables us to identify clusters of semantically similar articles and assess whether certain topics offer a higher propensity for paid content.

3.2 Data Preprocessing & Generation of Embeddings

Preprocessing for embedding-based models is not as intensive as for traditional NLP models, as embeddings inherently capture semantic relationships. However, the scraped content was saved in a markdown format, which included various non-textual elements such as images, code snippets, and formatting syntax. To ensure the quality of the text data used for topic modeling, we implemented a preprocessing pipeline which removed these non-textual elements, retaining only the core textual content of each article. This step was crucial to prevent noise from affecting the embedding generation and subsequent topic modeling.

For generating embeddings, we utilized the `prdev/mini-gte` model developed by QTACK [5] in early 2025. This distilled version of the original General Text Embeddings (GTE) model [2] achieved remarkable performance in the MTEB v2 English Benchmark [4], outperforming larger models while ensuring efficient use of memory and computational resources. This was especially suitable for our large dataset of Medium articles, allowing us to generate high-quality embeddings without prohibitive resource consumption.

3.3 Methodology

We conduct topic modeling through density-based clustering on a low-dimensional manifold learned from sentence embeddings. The overall workflow consists of the following steps:

Inputs and balanced sampling Two embedding matrices are loaded: one for free articles and one for member-only articles. A metadata file provides per-article labels and titles. To mitigate class imbalance, we apply stratified sampling, drawing `N_SAMPLES_PER_CLASS = 11000` articles per class.

Manifold learning We reduce the sampled embeddings to three dimensions using UMAP [3] with `n_neighbors = 20`, `min_dist = 0` and `n_components = 3`. The choice of a small `min_dist` emphasizes dense, well-separated aggregates to facilitate subsequent density-based clustering. The 3D space can be visualized easily while retaining sufficient structure for clustering.

Density-based clustering We fit DBSCAN [1] on the 3D UMAP coordinates using `eps = 0.3` and `min_samples = 150`. DBSCAN is non-parametric in the number of clusters and explicitly models noise (label `-1`), which is appropriate for heterogeneous web text where some points do not belong to any dense theme. Cluster labels are assigned per point and carried forward for analysis.

Topic labeling from representative titles For interpretability, we derive short human-readable topic names for clusters. For each non-noise cluster, we select 50 representative article

titles by sampling uniformly across the cluster's UMAP extent. We then label each cluster, trying to capture its thematic essence.

2D visualization We further map the non-noise 3D UMAP coordinates to two dimensions with t-SNE (`perplexity = 30`, `learning_rate = 200`, `random_state = 42`). This creates a 2D scatter plot where clusters can be visually separated and annotated with convex hulls. With color, we indicate the within-cluster free-to-paid article ratio.

Hypothesis tests within clusters Let $Y_i \in \{0, 1\}$ be the indicator for article i being free (1) or paid (0). Let p_0 denote the overall free proportion in the full embedding population ($p_0 = 0.5$), computed as the number of embedded free articles divided by the total across both classes. For each non-noise cluster c, we test the null hypothesis $H_0 : \mathbb{E}[Y \mid c] = p_0$ against the two-sided alternative using a one-sample t-test on the cluster's Y values. We report cluster size, sample mean $\hat{p}_c$, t-statistic, and p-value in Table A.

3.4 Results

Figure 3 presents the 2D t-SNE visualization of the clustered articles, with color coding to indicate the proportion of free articles within each cluster. A clear gradient emerges from left to right, with clusters on the left predominantly free (green) and those on the right predominantly paid (red). The thematic labels assigned to each cluster reveal, that this gradient corresponds to a shift from tech-based topics (e.g., software engineering, cloud computing, AI) to more personal and lifestyle-oriented themes (e.g., relationships, personal growth, health).

We can further group the identified clusters into higher-level themes based on their labels. Table 3.4 summarizes these themes, along with their sizes, average share of free articles, and variances. In this table, the previously observed gradient is reinforced: technical themes such as "Emerging Tech & Engineering" exhibit higher average ratio (0.635), while personal themes like "Health & Well-being" show a lower ratio (0.167).

A special mention goes to the "Crypto & Web3" cluster, which stands out with the highest mean free proportion of 0.862. This suggests that articles in this domain are mostly free.

Statistical significance was assessed for each cluster using one-sample t-tests comparing the proportion of free articles within the cluster to the overall population proportion of 0.5. Almost all clusters exhibited highly significant deviations ($p < 0.001$), indicating that the distribution of free versus paid articles is not uniform across topics. This further supports the notion that the topic of an article is strongly associated with its premium status on Medium.com.

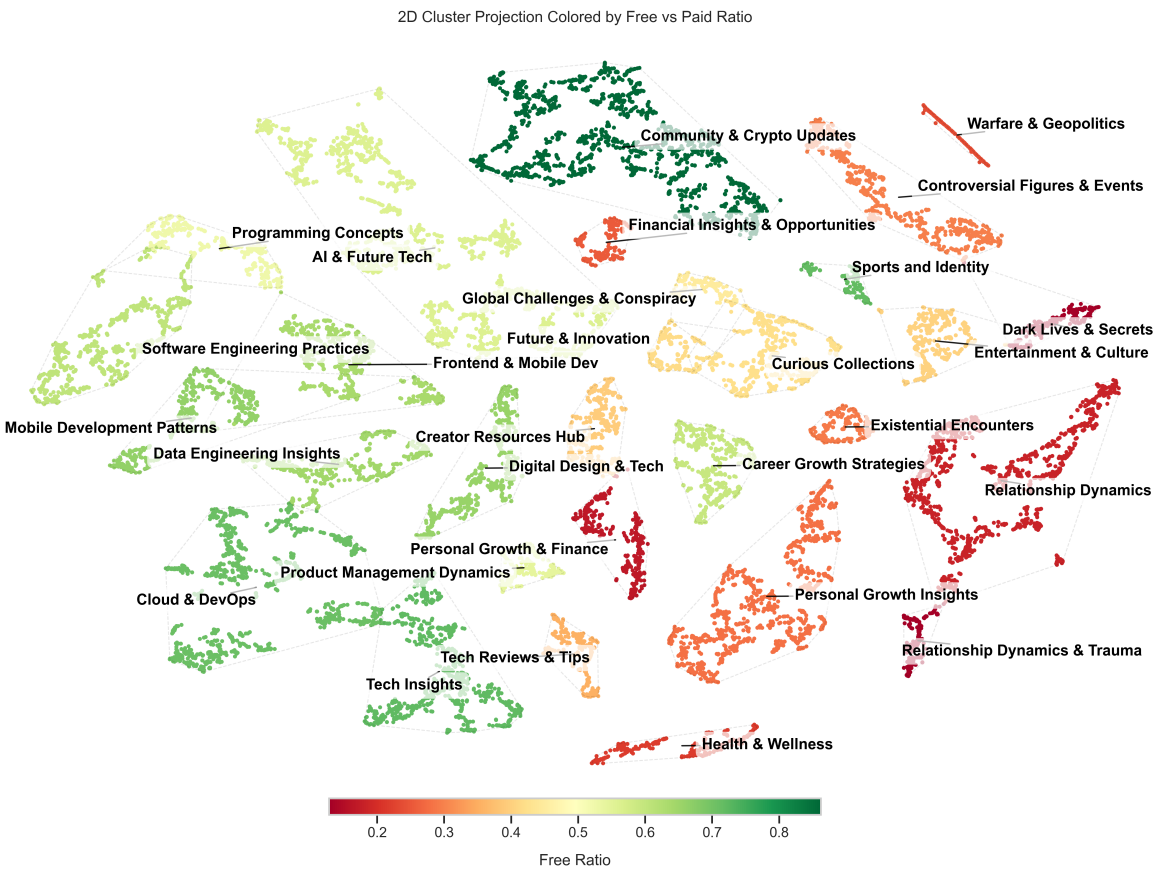

Figure 3: 2D Visualization of Topic Clusters illustrating the Ratio of Free to Paid Articles. The Color Gradient from Green to Red Indicates Increasing Proportion of Paid Articles within Each Cluster.

Theme	Size	Mean Free	Variance
Crypto & Web3	**1,489**	**0.862**	**0.119**
Emerging Tech & Engineering	**5,746**	**0.635**	**0.225**
Product, Business & Growth	**1,973**	**0.521**	**0.239**
Society, Culture & Global	**1,487**	**0.331**	**0.217**
Personal & Emotional Life	**2,515**	**0.240**	**0.183**
Health & Well-being	**492**	**0.167**	**0.144**
Creativity & Curiosity	**1,358**	**0.520**	**0.216**

Table 3: Summary of Higher-Level Topic Clusters with Mean Free Proportions

3.5 Discussion

Our embedding-based topic modeling reveals a pronounced thematic divide on Medium.com, with technical domains like "Crypto & Web3" and "Emerging Tech & Engineering" predominantly free (e.g., 86.2% free in crypto), potentially to foster community engagement and attract a broad readership. Conversely, personal and lifestyle themes such as "Health & Well-being" and "Relationship Dynamics" lean heavily toward paid content (e.g., 16.7% free in health), suggesting authors monetize intimate or specialized narratives. The t-SNE visualization's left-to-right gradient underscores this shift from collaborative tech discourse to introspective storytelling,

supported by highly significant t-test results ($p < 0.001$ for most clusters), indicating topic choice strongly predicts premium status.

Beyond engagement disparities, this pattern may reflect Medium's ecosystem dynamics: free tech articles could serve as lead magnets for building author followings, while paid personal content caters to readers seeking depth or exclusivity. However, limitations include potential biases in embedding models toward mainstream topics and the subjective nature of cluster labeling.

4 AI Generated Content

TODO: Flo's Teil

5 Conclusion

TODO: Write a conclusion

A Appendix

Topic	Size	Mean Ratio	t-statistic	p-value
Community & Crypto Updates	**1489**	**0.138**	**40.563**	**0.000**
Tech Insights	**765**	**0.278**	**13.663**	**0.000**
Sports and Identity	**155**	**0.284**	**5.949**	**0.000**
Cloud & DevOps	**1004**	**0.291**	**14.586**	**0.000**
Mobile Development Patterns	**477**	**0.338**	**7.496**	**0.000**
Digital Design & Tech	**528**	**0.343**	**7.603**	**0.000**
Data Engineering Insights	**489**	**0.354**	**6.755**	**0.000**
Frontend & Mobile Dev	**564**	**0.363**	**6.735**	**0.000**
Software Engineering Practices	**694**	**0.393**	**5.746**	**0.000**
Career Growth Strategies	**457**	**0.407**	**4.042**	**0.000**
AI & Future Tech	**1720**	**0.441**	**4.952**	**0.000**
Product Management Dynamics	216	0.458	1.226	0.221
Programming Concepts	354	0.475	0.957	0.339
Global Challenges & Conspiracy	165	0.552	-1.326	0.187
Curious Collections	**539**	**0.579**	**-3.704**	**0.000**
Future & Innovation	**180**	**0.583**	**-2.261**	**0.025**
Entertainment & Culture	**344**	**0.599**	**-3.735**	**0.000**
Creator Resources Hub	**383**	**0.606**	**-4.229**	**0.000**
Tech Reviews & Tips	**264**	**0.648**	**-5.015**	**0.000**
Controversial Figures & Events	**603**	**0.703**	**-10.910**	**0.000**
Existential Encounters	**210**	**0.710**	**-6.672**	**0.000**
Personal Growth Insights	**1021**	**0.721**	**-15.725**	**0.000**
Financial Insights & Opportunities	**197**	**0.746**	**-7.920**	**0.000**
Warfare & Geopolitics	**153**	**0.765**	**-7.694**	**0.000**
Health & Wellness	**292**	**0.784**	**-11.788**	**0.000**
Relationship Dynamics	**1097**	**0.820**	**-27.503**	**0.000**
Personal Growth & Finance	**370**	**0.832**	**-17.098**	**0.000**
Dark Lives & Secrets	**200**	**0.870**	**-15.520**	**0.000**
Relationship Dynamics & Trauma	**187**	**0.872**	**-15.155**	**0.000**

Table 4: Proportion of Free Articles per Topic with Cluster Size and T-Test Results

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[2] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023.

[3] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[4] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[5] QTACK. mini-gte: A compact text embedding model. Hugging Face, 2024. Accessed: 2025-10-17.