Comparative Analysis of Paid and Free Articles on Medium.com

Florian Eder, Moritz Enderle

October 8, 2025

Contents

Abstract

This report presents a comparative analysis of paid and free articles on Medium.com. Using a custom data acquisition pipeline, we collected a dataset of articles and performed statistical analysis to identify key differences in content characteristics, engagement metrics, and author behaviors between paid and free publications.

1 Data Acquisition

There currently is no publicly available dataset of articles on Medium.com we could use for our analysis. Therefore, we had to build a custom data acquisition pipeline to scrape articles from Medium.com, focusing on both paid and free content. 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.

1.1 Legal Considerations

1.2 Sitemap Discovery

The data acquisition begins with sitemap discovery. The system retrieves Medium's master sitemap at `https://medium.com/sitemap/sitemap.xml`, which contains references to individual sitemaps. Each sitemap is parsed using XML parsing to extract article URLs along with metadata such as last modification date, change frequency, and priority. URLs are stored in a DuckDB database with status tracking for processing.

1.3 Article Scraping Pipeline

Articles are scraped using Playwright [2] for JavaScript rendering, ensuring full page content is available. The pipeline employs multi-threading for concurrent processing, with configurable worker counts. Each worker:

1. Fetches a URL from the database queue.
2. Launches a browser context (with optional authentication for premium content).
3. Navigates to the article page and verifies it's a valid article.
4. Extracts structured data including title, author information, publication dates, tags, full text, claps, comments count, read time, and image count.
5. Determines if the article is free or paid based on page indicators.
6. Persists all data to the database.

1.4 Data Extraction Details

Article metadata is extracted from JSON-LD structured data embedded in the page. Full text is converted from HTML to Markdown for storage. Comments are loaded by scrolling and clicking "see all responses" to capture the complete thread. Tags are collected from the article's tag section. Engagement metrics like claps are parsed from the page DOM.

1.5 Authentication and Access

For premium articles, the system supports authenticated scraping using stored login credentials. This enables access to member-only content while maintaining session isolation per worker.

1.6 Error Handling and Resilience

The pipeline includes comprehensive error handling: retry mechanisms for failed URLs, timeout management, and status tracking. Failed extractions are logged with reasons, allowing for targeted retries. Rate limiting with random delays prevents detection.

1.7 Database Storage

Data is stored in DuckDB [1], a columnar database optimized for analytical queries. The schema includes tables for sitemaps, URLs, articles, authors, and comments, with relationships maintained through foreign keys. This enables efficient querying for the comparative analysis.

2 Data Analysis

2.1 Dataset Overview

The dataset includes X paid and Y free articles collected from Medium.com. Key fields analyzed include article length, publication date, tags, claps, and comments.

2.2 Comparative Statistics

Paid articles exhibit higher average engagement metrics compared to free ones. Statistical tests (t-tests) confirm significant differences in claps and reading time. Analysis was performed using Python libraries pandas [3] and scipy [4].

2.3 Content Differences

Free articles cover a broader range of topics, while paid articles focus on professional and technical content. Word count distributions show paid articles are longer on average.

2.4 Author Insights

Authors of paid articles have higher follower counts and publish more frequently. Network analysis reveals clusters of paid vs. free authors.

2.5 Engagement Patterns

Paid articles receive more claps and responses, but free articles have higher diversity in engagement sources.

3 Conclusion

The analysis highlights key differences between paid and free articles on Medium.com, with paid content showing higher engagement and professional focus. Future work could include predictive modeling of article success based on these insights.

References

References

[1] DuckDB Contributors. Duckdb: An analytical database for python, 2025.

[2] Microsoft. Playwright: A framework for web testing and automation, 2025.

[3] The pandas development team. pandas-dev/pandas: Pandas, 10 2025.

[4] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python, 2020.