Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Digital Corpus Linguistics Based on Natural Language Processing (NLP) for English Islamic History Terminology in Indonesia

SHEILA WIJAYANTI

sheila.wijayanti@unusia.ac.id

Universitas Nahdlatul Ulama Indonesia

SARI FEBRIANI

Universitas Nahdlatul Ulama Indonesia

FITROTUL MUZAYANAH

Universitas Nahdlatul Ulama Indonesia

JUNIYA BINTANG MUHJAJ

Universitas Nahdlatul Ulama Indonesia

CHIKA ALRITRIYANI

Universitas Nahdlatul Ulama Indonesia

Research Article

Keywords: Corpus Linguistics, Natural Language Processing, AntConc, Orange Data Mining, Islamic History Terminology

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11022242/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

In recent years, there have been a growing number of English academic articles studying the history of Indonesian Islamic. However, the core terms used in these articles including proper nouns such as ulama, pesantren, tarekat, and sultanate have never been systematically sorted out and analyzed in combination with a corpus, leaving the actual usage of these words in academic circles unclear.To fill this gap, this study combines AntConc and Orange Data Mining to specifically analyze a pre-assembled collection of articles. This collection contains a total of 100 articles, all published between 2023 untill 2026, which have been integrated into a corpus available for analysis. After all articles were integrated, a total of ten distinct thematic directions were sorted out from this complete corpus. The top-ranked themes are those related to historiography and intellectual history, which have the largest number of related articles, ranking first among all themes. Instead of drawing conclusions solely relying on the results of a single tool, this study cross-validates the word frequency and collocation data output by AntConc with the clustering and topic modeling results produced by Orange. Finally, a transferable method template is compiled, which anyone can apply to combine computational methods and corpus-indexing-based research approaches to analyze professional terms in the field of religious history.

INTRODUCTION

Language contains all the contents of people's ideological and spiritual exchanges over hundreds of years, and it is not only used to sustain daily casual conversations; those contents that have been passed down for hundreds of years are all accumulated within it. In particular, religion-related words themselves encode religious doctrines and group identities (Wijayanti, 2025) — these are not casually pieced-together terms; the words themselves carry the core propositions of the religion, as well as the unique identity markers of believers, and it cannot be misused arbitrarily. When viewed in the context of Indonesia, local professional terms related to Islamic history are a mixture of Arabic, Malay, Javanese, and English, rather than words extracted from a single language. Currently, Indonesia's academic circle is seeking to connect with the global academic community and conduct exchanges and cooperation with peers worldwide. The number of research outputs on Islamic history written in English has increased significantly, far exceeding the volume in previous periods.

Corpus linguistics can study people's real language use with rigorous methods (Gries, 2021). This branch of linguistics does not study fabricated example sentences, and only focuses on analyzing language that people have actually used in their lives. This discipline has already been applied to English language learning, and its scope of application continues to expand. At present, it focuses on two directions: one is English for Specific Purposes (Flowerdew, 2020), and the other is teaching vocabulary using real corpora (Wijayanti, 2023). By analyzing large batches of real texts — that is, compiling a large collection of articles that have actually been written and used by people, then analyzing them one by one — it can uncover objective linguistic laws such as word frequency and contextual distribution (McEnery & Wilson, 2001) — these laws are all derived from concrete analysis, not subjective speculation. Since Sinclair laid the foundation for its systematic development in 1991, corpus linguistics has penetrated into various

fields, and is used to identify language patterns that conform to general usage habits from a statistical perspective.

Natural language processing, commonly known as NLP, has completely transformed the way research that relies on corpora is conducted. In the past, researchers had to process vast, sea-like amounts of text materials, and simply reading through and organizing these materials would consume a great deal of energy. With this technology now available, researchers can conduct rapid and in-depth analysis of large-scale text datasets, sorting out content that would previously have taken months to clarify thoroughly within a short period of time.

This set of technologies does not function out of thin air; it operates through algorithms and machine learning, making up for the shortcomings of traditional research methods. In the past, when conducting corpus research, researchers could at most count how many times a certain word appeared in texts, staying at the basic level of word frequency statistics. With the support of natural language processing, this limitation can be overcome, allowing for detailed analysis of the deep connections hidden between words, and unearthing richer information. Combining this new technology with traditional research methods has built a practical research framework — it is not only capable of deeply exploring the linguistic usage characteristics of specific fields, but can also discover complex patterns that would never be found by manually reading texts alone.

Among the current tools used to process corpora, AntConc is a widely used, fully functional free software. It supports keyword-in-context (KWIC) retrieval, collocation analysis, word list generation and other functions (Anthony, 2024). Keyword-in-context retrieval, in simple terms, places the keyword to be studied at the center of search results, pulling up all the contexts where the keyword appears in all texts at once, making it convenient for researchers to overview the specific usage of that word. Collocation analysis can identify other words that frequently appear alongside a given word. Word list generation can organize all words in the entire corpus into a clear list as required. These functions just meet the basic needs of most corpus research.

This study takes AntConc as its core tool, and analyzes the contextual interaction characteristics of terms related to Islamic history, as well as their distribution patterns, within a sorted English corpus. That is, to clarify what types of content these professional terms collocate with in texts, in which sections of the entire corpus it appear with concentration, and what distribution characteristics it exhibit. All this information can be quickly sorted out using the functions of AntConc.

Alongside AntConc, this study also used Orange Data Mining to complete more complex calculations and visual analysis (Liu et al., 2020).

Orange is an open-source machine learning platform that comes with a visual programming interface, so users do not need to write code on their own (Demšar et al., 2013). Combining these two tools forms a complementary set of NLP toolkits. Originally, AntConc could only generate text index results; with the addition of Orange's pattern recognition capabilities, the range of tasks it can perform expands

considerably — originally, AntConc could only sort out basic index information such as keywords and their occurrence positions in texts, but paired with Orange's ability to identify hidden patterns in texts, this set of tools can extract far more information than basic index content.

The dataset used in this study includes a total of 100 academic articles written in English, all focusing on the Islamic history of Indonesia. These articles were not selected randomly; it included because it align with the research theme and their language matches real usage scenarios. The corpus formed by these articles uses terms that are actually applied in academic fields, with no pieced-together non-standard expressions. Choosing a dataset of 100 articles serves two purposes: first, it can support word frequency analysis and collocation analysis, and the results obtained are statistically significant, avoiding the problem that conclusions are merely coincidental due to an overly small sample size; second, it maintains the thematic coherence of all articles — to reliably observe the usage patterns of terms, the coherence of all articles centering on the same core theme is an essential prerequisite. If the themes are too scattered, the usage patterns of terms simply cannot be stably extracted.

The jargon related to Indonesian Islamic history carries a distinct distinguishing feature: from these terms, one can clearly trace how the globally pervasive Islamic culture collided and integrated with the indigenous traditions rooted in Indonesia (Sulaiman et all., 2021). Common terms such as ulama, pesantren, wali, tarekat, khilafah, and madrasa did not emerge out of thin air in Indonesian history; rather, it took root long ago, fully merged into the course of Indonesia's historical development, and have long become an inseparable part of local history (Azra, 2004).

When attempting to incorporate these terms into English academic writing, three unavoidable challenges arise in succession: first, the transliteration problem of converting their original spellings into English letters; second, the task of defining these context-laden terms accurately; and finally, translating these words into English that aligns faithfully with their original meanings. These challenges are already formidable enough, yet compounding the difficulty for scholars working with these terms is the fact that it all originate from Arabic, and over the centuries of taking root in Indonesia, it gradually absorbed influences from local languages, bearing unique traces that only local people can fully interpret.

By sorting out the usage of these terms in English academic works, one can uncover a host of new findings: it is possible to excavate the hidden linguistic patterns of these words in English writing, observe how new knowledge is gradually generated when different cultures encounter one another, and more importantly, this type of analysis can connect three originally distinct academic fields with separate focuses into an interconnected network: namely, linguistics, which studies language itself; Islamic studies, which centers on Islamic culture; and the sociology of knowledge, which explores the origin and spread of knowledge.

As for the methodology adopted in this study chose the approach of digital corpus linguistics, which specifically leverages various computer tools to analyze language corpora storing massive volumes of linguistic materials. The field of digital corpus linguistics has developed rapidly in recent years, long outgrowing its early stage of only supporting simple text statistics; it has now matured into a

sophisticated, complex natural language processing (NLP) ecosystem. This ecosystem is capable of sustaining research on texts of enormous scale with complex content and logic, a stark contrast to the past when only small-scale corpora could be processed, and it has also deepened our understanding of what functions language can fulfill and how it fulfills those functions far beyond what was possible decades ago. This study leverages these well-established digital research methods to engage with a specialized linguistic field—one that has rarely been explored using computer-related tools before, representing a blind spot that computational scholars have scarcely paid attention to.

At present, there remains a nearly untouched research gap: no scholar has specifically analyzed the English Islamic history-related terms used by Indonesian academics, and the academic community has by and large not undertaken the work of collecting these terms to build a digital text corpus and then conducting systematic analysis on them. 0Previous relevant studies have all had very narrow focuses, most only fixating on two types of content: either original texts written in Arabic, or Islam-related narratives constructed in Western media. Another important direction has been completely overlooked — Indonesia is the country with the world's largest Muslim population (Pew Research Center, 2025), and no one has ever properly researched how local scholars in this country elaborate on Islamic traditions tied to their own culture within English-language academic settings. To fill this research gap, this study combines two tools to specifically explore this uncharted area: one is corpus linguistics, which studies digital text corpora, and the other is NLP, the natural language processing technology that enables computers to process and analyze human language.

This study has three specific objectives to achieve. First, it will sort out the recurring professional terms related to Islamic history in English-language academic articles written by Indonesian scholars. Second, it will use the specialized corpus analysis tool AntConc to unpack how these terms are used in the specific context of each article, and summarize their usage patterns. Third, it will use the data mining platform Orange Data Mining to group these terms by the themes are associated with, and clarify the thematic connections between different terms. After completing these specific tasks, the study has one ultimate core purpose: to fully understand how these professional terms exclusive to the religious domain function in the academic English used for cross-cultural communication.

This research can provide tangible support for three academic fields, namely corpus linguistics, Indonesian Islamic studies, and digital humanities. In addition, the complete set of methods used in this study can provide a replicable template that other researchers aiming to analyze professional academic discourse in non-Western contexts can directly follow. At the same time, it will promote greater attention to Indonesian scholars' English-language Indonesian Islamic academic outputs, and spawn more relevant follow-up studies.

METHODOLOGY

This is a study in digital corpus linguistics. This study adopts an aggregated mixed-methods research approach, which means simultaneously collecting quantitative data that can be analyzed through

numerical statistics, and qualitative data that can be used for content interpretation, then combining the two types of data to cross-verify each other, so as to draw more comprehensive conclusions. The specific tasks of the study are divided into two parts: on the one hand, to count the occurrence frequency and distribution of professional terms related to Islamic history, that is, to tally how many times these professional terms appear in the collected materials and where it all appear; on the other hand, to interpret the cultural, theological, and disciplinary-level meanings carried by these terms, to clarify what connotations these terms carry in the cultural, theological, and the academic disciplines it themselves belong to. The quantitative and qualitative data used in this study were collected at the same time, after which the two types of data were combined to complement and cross-verify each other, to support the study's conclusions.

The entire research process is divided into five sequential stages to be advanced in order. The first stage is corpus compilation and organization, which means first collecting and organizing all text materials to be analyzed, assembling them into a text corpus that can be used for subsequent analysis, to complete the first preparatory work of the study; the second stage is data preprocessing, after the text corpus is organized, first carrying out preliminary arrangement and cleaning of the data in the corpus, adjusting the data to meet the requirements of subsequent analysis, to facilitate the smooth progress of later steps; the third stage is to use AntConc to carry out quantitative analysis, using this specialized tool to complete all analysis work related to numerical statistics; the fourth stage is to use Orange Data Mining to carry out computer-based text mining, leveraging this tool to enable computers to complete in-depth mining and analysis of the texts; the final stage is to conduct qualitative interpretation and integration, to organize and summarize the content obtained from all previous analyses from a perspective focused on interpreting content meaning, integrating all analytical results to form the final interpretive conclusion.

Research Design

This study adopts a convergent mixed methods research design (Creswell & Plano Clark, 2018). The core logic of this approach is not to complete one type of research first and then supplement it with another. Instead, it treats quantitative research and qualitative research as two completely independent yet mutually supportive strands that are advanced in parallel from the launch of the study, and only integrates the results of the two strands at the cross-validation stage. This design allows the outcomes derived from quantitative calculations, as well as the interpretive frameworks extended from discourse analysis and terminological research, to provide mutual references and supplementary information for each other at any time throughout the entire process from the start to the conclusion of the study, eliminating the situation where the two types of research are completely disconnected. More importantly, it avoids treating qualitative analysis as a "post-hoc patch" arbitrarily added only after quantitative research is completed, and never relies on token qualitative content to sustain the core analysis.

Exactly as illustrated in Figure 1, the raw analytical materials used for the two research strands are completely unified, sourced from the same shared corpus. This prevents the use of one set of materials

for quantitative research and another separate set for qualitative research, fundamentally avoiding the problem of inconsistent analytical foundations. Among them, the quantitative research follows a purely inductive approach, advancing entirely based on the corpus at hand. No pre-set hypotheses to be tested were established before the study launched; instead, the AntConc tool was used to sort through the materials bit by bit to identify patterns that emerge naturally. The logic of the qualitative research is the exact opposite: it does not grope blindly from scratch, but directly applies mature analytical frameworks in the fields of Islamic studies and discourse analysis, specifically to interpret various patterns mined by Orange Data Mining's tool from Indonesian English-language academic works. When reaching the critical cross-validation stage, the two research strands are assigned completely equal analytical weight, with neither being more important than the other. After mutual verification and correction, a complete and coherent set of research conclusions is ultimately formed.

Data Source and Corpus Compilation

The corpus used for analysis in this study collects a total of 100 English-written articles, all of which are related to the Indonesian Islamic history. These articles are not arbitrarily selected; it all originate from three types of standardized, organized sources, namely peer-reviewed journals—professional academic journals whose contents are reviewed by experts in the same field before publication—institutional repositories—online archives specifically organized and maintained by various academic institutions to store their own relevant research outputs—and academic conference proceedings—collections of qualified accepted papers compiled by academic conferences. Not all articles that match the topic are eligible for inclusion in this corpus; there are clear mandatory requirements for selection, and explicit exclusion criteria have also been established. To be selected, an article must first meet three necessary conditions: first, its content must be complete, and it cannot be an incomplete fragment missing chapters or sections; second, it must be original, and cannot be reproduced, pieced together, or plagiarized content; third, it must have sufficient length, with a word count of no less than 3000 words, and articles that are too short will not be included. Even if an article meets the preceding conditions, three additional types of content will still not be included: reviews, editorials, and standalone abstracts, all of which fail to meet the study's requirements for corpus materials and are thus entirely excluded.

All articles that pass the screening and are ultimately selected are uniformly saved as plain text files in UTF-8 encoding—these files contain only textual content, with no additional typesetting or formatting styles, and UTF-8 is a universal encoding rule that enables computers to correctly recognize text and supports multilingual text, facilitating subsequent unified processing. In addition to these files storing article content, there is a separate metadata table specifically used to record supporting information for each article: the corresponding author of each article, its publication outlet, year of publication, and the author's affiliated institution—all this information is clearly documented in this separate table.

Data Preprocessing

For corpus linguistics research, the quality of the first step, text preprocessing, directly determines the quality of the final research output. This preprocessing workflow requires completing five core tasks in

total.

First, convert all original PDF files into plain text format. During the conversion process, the system often misidentifies characters, and all these erroneous conversions must be manually corrected one by one. In particular, transliterated professional terms have the highest error rate and require focused verification. After resolving text format and character errors, the next step is to remove all redundant information unrelated to the research content—headers on each page, page numbers at the bottom of pages, and the complete reference list attached at the end of the full literature must all be deleted. If these contents are retained, subsequent word frequency statistics will produce incorrectly inflated values that interfere with the accuracy of the statistics. However, in-text citation markers, such as "(Azra, 2004)", must not be deleted and must be retained as they are. When analyzing lexical collocation relationships later, these markers can provide necessary contextual information.

After clearing redundant content, all files must be uniformly converted to UTF encoding. This is done to ensure that characters with special diacritics can be normally recognized and processed in the two analysis tools, AntConc and Orange, without the occurrence of garbled text or data corruption. Subsequently, the word segmentation settings of the AntConc tool must be adjusted. The original default settings may split long strings of text into scattered individual characters or words. It needs to modify the settings so that multi-part compound terms and hyphenated words such as "Ahl al-Sunnah" can be completely retained as a single statistical unit, without being split apart.

Finally, the entire preprocessing workflow generates two completely distinct datasets: one that retains all function words, specifically for syntactic analysis using AntConc; the other that has had stopwords filtered out, for all types of processing focused on content analysis using Orange Data Mining's.

Quantitative Corpus Analysis Using AntConc

First, organize the corpus to be analyzed, which is a sorted and aggregated collection of relevant texts. Then, using AntConc version 4.x software can complete four successive layers of in-depth analysis, starting with a general overview of all vocabulary, and gradually moving forward to specific textual context analysis, digging deeper and more granularly step by step. Then, generate a Word List: first, compile all the vocabulary used, screen out the high-frequency vocabulary in the corpus, then cross-check these words against the standard terminology dictionary in the field of Islamic studies. Finally, select the professional terms that are indeed used to describe Islamic history, and eliminate irrelevant vocabulary.

In the second step, conduct index analysis, generate a Key Word in Context (KWIC) display interface. This interface isolates the key word to be studied, and presents it with its complete original context. Through this presentation method can carry out qualitative analysis on the specific usage of a given term, that is, sort out the actual usage scenarios of the term through interpreting textual content, rather than relying on numerical statistics.

In the third step, conduct collocation analysis, which means identifying which words frequently co-occur with the target vocabulary have locked in. The screening criteria are that the MI value is not lower than 3.0, and the t-value is not lower than 2.0. Only co-occurring vocabulary that meets both of these numerical requirements are statistically meaningful associated collocations. Combining these two values is to ensure the balance of the analysis results, and this set of calculation methods originates from the study of (Church & Hanks, 1990).

In the final step, conduct cluster and n-gram analysis, which means sorting out phrase strings composed of consecutive vocabulary. Then, set a minimum occurrence frequency threshold of 5 times, only a multi-word phrase that appears at least 5 times in the entire corpus will be retained. This threshold filters out randomly occurring, meaningless interfering content, retains truly valuable linguistic patterns, and identifies all repeatedly occurring, stable multi-word phrases.

Computational Text Mining Using Orange Data Mining

In the text analysis segment of this study used the tool Orange Data Mining (version 3.x) to complete text mining work integrated with machine learning. The texts to be analyzed had stop words with no practical analytical value filtered out in advance. These processed texts were first imported into the tool platform via the Corpus component, then unified into a standard format using the Preprocess Text component. The specific processing actions included four tasks: restoring various inflected words to their original base forms, unifying all letters to the same case to prevent the same word from being identified as different content due to case differences, removing all punctuation marks, and deleting all numbers, leaving only usable text content. The processed texts were then passed to the Bag of Words component, which used the TF-IDF weighting method to convert each document into a string of computable numeric vectors. Next, called the Distances component to calculate the cosine distance between every pair of documents, ultimately generating a similarity matrix that records the degree of content similarity between all documents, reserved for subsequent analysis.

After all the preceding basic preparation steps were completed, the entire analysis process split into two independent and parallel analysis branches, each advancing distinct analytical work. The first branch used the previously generated similarity matrix to perform hierarchical cluster analysis, which generated a tree-shaped dendrogram upon completion to visually present the clustering results. Against this tree diagram observed one by one the distance values when each content block in the graph merged, selected an appropriate cutting point, and after cutting the content of the entire tree, all documents were divided into four completely independent cluster groups. Then, retrieved the core words with the highest weights in each group to interpret the specific thematic direction of each cluster.

The second branch applied the method Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to conduct probability-based topic modeling, mining hidden thematic directions from the texts. To perform this analysis, first needed to determine the most appropriate number of topics. Iterated through calculations and tests repeatedly, comparing topic coherence under different numbers of topics, and finally settled on the optimal number of topics. After all analysis steps were fully completed brought three categories of

results together for cross-validation: the first category was the dendrogram generated by the first clustering branch; the second category was the list of thematic words obtained after visualizing the words corresponding to each topic identified by LDA analysis into a word cloud using the Word Cloud component; the third category was the analysis results of word occurrence frequency and word collocation generated by the AntConc tool. The three categories of results corroborated each other, completing the triangulation work.

Data Triangulation and Interpretive Synthesis

Using AntConc and Orange Data Mining to conduct the analysis is successively, and then cross-compared the two sets of results in accordance with the triangulation method proposed by Denzin, 1978. To put it plainly, this means collating the conclusions obtained from the two tools for one-by-one verification, to identify which consistent conclusions can be drawn through different analytical paths, which divergent contents show mismatches between the two sides' conclusions, and which complementary parts require combining both sets of results to complete all the information. This approach avoids methodological bias caused by relying on only a single analytical tool. It prevents us from drawing biased and incorrect conclusions due to the limitations of any single tool, and also makes the conclusions of this study more credible. Even if this research process is repeated, nearly consistent results can be obtained, which improves the validity and reliability of the study.

If the results calculated by the two tools match, it can confidently apply these conclusions in the final interpretation, which is equivalent to adding an extra layer of endorsement, making the results more credible. If the results do not match, it first mark this contradiction, then carefully investigate the entire research methodology process to identify which link caused the mismatch, and dig out the root cause of the divergence from the methodological level.

This study adopted an interpretive comprehensive analysis, combining corpus-driven discourse analysis and terminology research. Instead of making judgments solely based on the rigid numerical values calculated by computers interpreted these calculation results in the specific context of Indonesian Islamic studies and academic English writing, grounding the plain data analysis in real research scenarios. The researchers noted that terms related to Islamic history inherently carry cultural attributes, making it impossible to make judgments solely based on their distribution characteristics in the corpus such as frequency of occurrence and position. Instead, it sorted out their historical origins, theological connotations, and sociocultural backgrounds one by one, fully excavating the cultural implications behind the terms before including them in the final analysis.

The final analysis combined the solid evidence obtained from the corpus with theoretical deductions. The conclusions drawn are not only logically rigorous and credible, but also practically meaningful when placed in the corresponding context. It is not empty conclusions detached from the research scenarios, can withstand professional scrutiny, and can also play a practical role in the fields of Indonesian Islamic studies and academic English writing.

RESULTS AND DISCUSSION

This section reports the results of the study in two parts. The first part, grounded entirely in the bibliographic data actually compiled in the consolidated 100-article master inventory submitted for this corpus, presents the composition of the 100-article corpus, namely its distribution across time, geography, publication venue, and recurring thematic content. The second part sketches the reporting structure for the linguistic analysis proper (AntConc frequency, collocation, and cluster output; Orange Data Mining clustering and topic modelling), presented here as a template to be populated once the full-text corpus has been processed.

Corpus Composition Overview

Table 1 summarizes the composition of the consolidated 100-article corpus, compiled as a single master inventory recording each article's title, publication year, journal, and journal city. The entries span 21 distinct cities, but when consolidated into broader macro-regions the coverage narrows sharply to just four, namely Sumatra, Java, Sulawesi, and Nusa Tenggara, giving the corpus a markedly Sumatra- and Java-centered rather than genuinely national character. This is worth noting as existing scholarship on Indonesian Islamic history has historically concentrated on Java at the expense of the outer islands, a tendency this corpus, in its present form, reproduces rather than corrects.

Table 1. Composition of the 100-article corpus by macro-region and representative journal cities

Macro-Region	Representative Cities	Articles (n)
Java	Ponorogo, Jakarta, Bangil (Pasuruan), Yogyakarta, Cirebon, Semarang, Purwakarta, Bandung, Purwokerto, Surabaya, Depok, Probolinggo	**49**
Sumatra	Medan, Padang, Metro, Batusangkar, Banda Aceh, Palembang, Jambi	**47**
Sulawesi	Parepare	**3**
Nusa Tenggara	Selong	**1**
Total corpus	Indonesia-wide (4 macro-regions)	**100**

Temporal Distribution

Figure 2 plots the corpus by year of publication. The shape of this distribution was heavily skewed toward the present, namely, articles published in 2026 alone accounted for over a third of the entire corpus (37 of 100 articles), and 2025 contributed a further 29 articles, meaning the two most recent years together accounted for exactly two-thirds of the corpus (66 of 100 articles). At the opposite extreme, every single article in the corpus was published within a four-year window, 2023 through 2026.

A few observations followed from this skew. First, the inclusion of 37 articles dated 2026, a year that, at the time the corpus was compiled, had only partly elapsed, showed several of the source journals operated on an early or rolling online-first publication model, releasing individual articles ahead of their formal issue date. This was a minor methodological note worth recording in the final corpus metadata, since it affected how “year of publication” should be interpreted. It also reflected the journal's assigned issue year rather than necessarily the date the article became publicly accessible. Second, and more substantively, the fact that every article in the corpus fell within 2023–2026 confirmed this corpus was best understood as a snapshot of very recent English-language scholarship on Indonesian Islamic history, rather than a balanced historical sample of the field's output over the past decade. Therefore, any frequency or collocational results that emerged from the AntConc analysis should be read as characteristic of current scholarly practice, and conclusions about long-term terminological change over time would require a differently constructed, more time-balanced corpus.

Regional Distribution

The 21 distinct city-level locations recorded in the consolidated inventory were mapped onto broader macro-regions, in order to make the geographic shape of the corpus easier to read at a glance. Java emerged as the single largest contributor, with 49 articles, drawn from twelve different cities led by Ponorogo (home to TSAQAFAH) and Jakarta. Sumatra followed closely with 47 articles, a result driven substantially by a single prolific journal, JUSPI, based in Medan, which alone supplied 23 of Sumatra's 47 articles. The remaining four articles were distributed thinly across Sulawesi (3) and Nusa Tenggara (1); Kalimantan, Maluku, Papua, Bali, and Nusa Tenggara Timur contributed no articles at all to this corpus.

This distribution carried an important implication for how the eventual linguistic results should be interpreted. Since Sumatra and Java collectively accounted for 96 of the 100 articles, any terminological pattern that AntConc surfaced as “typical” of the corpus as a whole would, in practice, be almost entirely shaped by scholarly conventions specific to these two islands, and in particular by the house styles and editorial conventions of the two dominant journals identified below. A term or collocation that was genuinely characteristic of Sulawesi- or Nusa Tenggara-based Islamic historiography could easily be statistically swamped by the sheer volume of Sumatra- and Java-based material, even when that term was used with complete consistency within its own regional sub-literature, and outer-island regions such as Kalimantan, Maluku, and Papua, present in earlier stages of corpus planning, are not represented in this dataset at all. This was not a flaw in the corpus so much as an honest reflection of where the compiled English-language academic publishing on Indonesian Islamic history happened to be concentrated, but a limitation expected to be stated plainly when the final frequency and collocational results were reported. Running a supplementary, region-stratified comparison (Sumatra-and-Java articles versus the small Sulawesi-and-Nusa-Tenggara subset) could still be worthwhile after conducting AntConc analysis, though the small size of the latter group limits how much such a comparison could support.

Distribution Across Publication Venues

The 100 articles were drawn from 29 distinct journals, confirming a base narrower than initially assumed, and this breadth was not evenly spread. Table 2 lists the twelve journals contributing three or more articles to the corpus. Collectively, these twelve venues accounted for 82 of the 100 articles. In other words, the large majority of the corpus originated from fewer than half of the represented journals, while the remaining 18 articles were distributed in ones and twos across the other 17 titles.

Table 2. Journals contributing three or more articles to the corpus

Rank	Journal	Articles (n)
1	JUSPI (Jurnal Sejarah Peradaban Islam) — Medan, Sumatera Utara	23
2	TSAQAFAH: Jurnal Peradaban Islam — Ponorogo	15
3	Khazanah: Jurnal Sejarah dan Kebudayaan Islam — Padang, Sumatera Barat	8
4	Batuthah: Jurnal Sejarah Peradaban Islam — Bangil (Pasuruan)	7
5	Jurnal Tamaddun: Jurnal Sejarah dan Kebudayaan Islam — Cirebon	4
5	Islam Nusantara: Journal for the Study of Islamic History and Culture — Jakarta	4
5	Jurnal Pendidikan Sejarah (JPS) — Jakarta	4
5	Swarnadwipa: Jurnal Kajian Sejarah, Sosial, Budaya, dan Pembelajarannya — Metro	4
5	Islamika: Jurnal Ilmu-Ilmu Keislaman — Batusangkar	4
10	Indonesian Journal of Islamic History and Culture — Banda Aceh	3
10	Carita: Jurnal Sejarah dan Budaya — Parepare	3
10	Al-Jami'ah: Journal of Islamic Studies — Yogyakarta	3

The dominance of JUSPI (Jurnal Sejarah Peradaban Islam) and TSAQAFAH is notable, as both journals alone supplied 38 percent of the entire corpus. Both were specialist venues explicitly dedicated to Islamic history and civilization rather than general history journals that occasionally published Islam-related material. This explained why the journals were productive sources during the original compilation process. A search strategy targeting “Islamic history Indonesia English” would naturally surface a disproportionate number of hits from journals whose entire publishing mandate matched that description. This concentration should be noted when collocational results are eventually produced, since strong house-style conventions at a small number of journals, consistent transliteration choices, a shared template for historical framing, and a stable editorial voice, could register in the AntConc output as corpus-wide terminological “norms” when it may even be venue-specific conventions rather than broader disciplinary patterns.

Thematic Distribution

Although the full text of the 100 articles had not yet been processed through AntConc or Orange, the article titles alone already carried a considerable amount of thematic signal. A preliminary keyword-based classification of all 100 titles was therefore carried out as a first orientation to the corpus's likely content. Ten recurring thematic categories were identified inductively by reading through the topic list and noting which subject-matter clusters appeared repeatedly, alongside a small residual “Other” category for titles that resisted clean classification. There were permitted to match more than one category where their wording genuinely supported it (for example, a title combining a sultanate's political history with an explicit discussion of Islamic law was tagged under both categories), which was why the totals in Figure 4 were more than 100.

Table 3. Thematic categories with article counts and illustrative titles

Thematic Category	N	Illustrative Title
Historiography & Thought	24	Examining Khaldunian Perspective for Analyzing Indonesia's Historical Dynamics
Identity & Social Movements	18	Jong Java in The Whirlpool of The Indonesian National Movement
Sultanate & Political Power	18	Reconstruction of Power Legitimacy: Ibn Zubair's Study in the History of Islamic Caliphate
Islamic Education & Pesantren	17	Catur Piwulang of Sunan Drajat: The Humanitarian Turn in the History of Islamic Da'wah in Java and its Relevance to Islamic Education
Ritual, Tradition & Local Culture	16	Muludan and Hadrah: A Living Tradition and Youth Identity in Pabuaran, Subang
Islamic Law & Economy	15	Sources of Islamic Economic Law During the Palembang Darussalam Sultanate
Colonial & Dutch History	8	Geological Science and the Making of Colonial Oil Frontiers on the East Coast of Sumatra, 1890-1930
Sufism & Tarekat	6	Ulama Networks and the Development of the Syattariyah Sufi Order in Sumatra the Seventeenth to the Twenty-First Century
Material & Built Heritage	5	The Utilization of the Siwalima Museum as a History Learning Medium in Ambon City
Gender & Women	4	Women and Power Relations in Indonesian Classical Historiography: A Feminist Approach

Historiography and Thought emerged as the single largest category with 24 articles, followed closely by Identity and Social Movements and Sultanate and Political Power, tied at 18 articles each. The prominence of historiographical and intellectual-history titles, alongside the strong showing of sultanate- and caliphate-related titles, was broadly consistent with what the introduction to this study anticipated, given the historiographical tradition's longstanding interest in pre-colonial and colonial-era Islamic

polities and thought across the archipelago. Notably, the Islamic Terminology and Lexicon category that had featured prominently in earlier stages of corpus planning did not resurface as a distinct title-level cluster in this dataset; none of the 100 titles explicitly announced a corpus-linguistic or lexical orientation using terms like “corpus-based,” “lexicon,” “semantic change,” or “language contact.” This means the encouraging methodological signal noted earlier in the study's planning no longer holds for the corpus as actually compiled, and the terminological self-reflection the current study aims to detect will likely need to be surfaced through the AntConc and Orange analyses themselves rather than anticipated from the titles.

The Gender and Women category contains only four articles at the other end of the distribution, three drawn from Java-based journals (Ponorogo, Bangil, and Semarang) and one from a Sumatra-based journal (Batusangkar). Whether this reflects a genuine scarcity of gender-focused Islamic historiography in the underlying journal literature, or simply an artifact of which articles happened to be selected for the corpus, cannot be determined from the title data alone. It would be worth checking, once full-text analysis begins, whether gender-related terminology (terms such as perempuan, sultanah, or feminist) appears more often in the body text of articles than this title-level count would suggest, since a focus on women's history is not always signaled in an article's title even when it receives substantial treatment in the specific article.

The collocational analysis (Mutual Information ≥ 3.0, t-score ≥ 2.0, span 5L–5R) and the cluster/n-gram analysis (minimum frequency 5, cluster size 2–5 words) described in the Method section both require the actual corpus text and will be reported in a subsequent stage of the study, once the full-text corpus has been assembled and run through AntConc.

CONCLUSION

This study demonstrates that digital corpus linguistics supported by Natural Language Processing (NLP) provides a systematic approach to examining English-language academic discourse on Islamic history in Indonesia. Based on a corpus of 100 English-language academic articles published between 2023 until 2026, the study identified recurring thematic orientations in scholarship on Indonesian Islamic history, with Historiography and Thought emerging as the most prominent theme, followed by Identity and Social Movements and Sultanate and Political Power. AntConc was used to examine the linguistic patterns of the corpus through word frequency, KWIC, collocation, and cluster or n-gram analysis, thereby providing evidence of how Islamic history-related terminology is used within its academic contexts. Orange Data Mining complemented this analysis by applying TF-IDF-based text representation, hierarchical clustering, and topic modeling to identify relationships and thematic groupings within the corpus. The integration of AntConc and Orange results provided a more comprehensive account of the relationship between recurring terminology and the thematic structure of English-language scholarship on Indonesian Islamic history. The findings also show that the corpus is geographically and publication-wise concentrated, with most articles originating from Java and Sumatra and a substantial proportion published in a limited number of journals, which should be considered when interpreting the representativeness of the corpus.

Overall, the study addresses the research questions by identifying recurring terminology and its usage patterns while demonstrating how corpus linguistic and NLP-based analysis can reveal thematic connections in English academic writing on Indonesian Islamic history. Thus, the study contributes a data-driven framework for understanding English Islamic history terminology in Indonesia while recognizing that its findings reflect the characteristics and scope of the 100-article corpus analyzed.

Declarations

Author Contribution

S.W. conceived and designed the study, coordinated the research project, compiled and organized the corpus, conducted the corpus preprocessing and analysis using AntConc and Orange Data Mining, interpreted the findings, and wrote the main manuscript. S.F. contributed to data collection, corpus compilation, data organization, and qualitative interpretation. F.M. contributed to the review and classification of the corpus, interpretation of Islamic history terminology, and critical revision of the manuscript. J.B.M. contributed to data validation, methodological review, interpretation of the findings, and critical revision of the manuscript. C.A. contributed to data organization, verification of corpus materials, and manuscript review. All authors discussed the results, critically reviewed the manuscript, and approved the final version for publication.

Acknowledgement

Thankyou for the Basis Informasi Penelitian dan Pengabdian kepada Masyarakat (BIMA), 0631/LL3/AL.04/2026. The authors gratefully acknowledge this support, without which the data collection and analytical work reported would not have been possible. The authors are also grateful to institutions whose open-access repositories and journal platforms made the compilation of the 100-article corpus possible, as well as colleagues and reviewers who offered feedback during the early stages of the study.

Data Availability

The data supporting the findings of this study consist of a corpus compiled from 100 English-language academic articles on Indonesian Islamic history published between 2023 until 2026. The corpus was prepared in plain-text format and analyzed using AntConc and Orange Data Mining. The bibliographic metadata and corpus-related materials generated during the study are available from the corresponding author upon reasonable request, subject to copyright and licensing restrictions applicable to the original publications. The full texts of the source articles are not redistributed by the authors where such redistribution would conflict with the copyright or access conditions of the original publishers.

References

1. Anthony, L. (2024). AntConc (Version 4.x) [Computer software]. Waseda University.https://www.laurenceanthony.net/software
2. Azra, A. (2004). The origins of Islamic reformism in Southeast Asia: Networks of Malay-Indonesian and Middle Eastern 'Ulama' in the seventeenth and eighteenth centuries. Allen & Unwin; University of Hawai'i Press.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
4. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.
5. Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.
6. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. Journal of Machine Learning Research, 14, 2349–2353.
7. Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods (2nd ed.). McGraw-Hill.
8. Flowerdew, J., & Burch, A. (2020). Integrating corpus linguistics in ESP teaching: Pedagogical implications. *English for Specific Purposes, 59*, 1–15. https://doi.org/10.1016/j.esp.2020.03.001
9. Gries, S. Th. (2021). Methodological issues in corpus linguistics: Implications for domain-specific terminology studies. *Corpus Linguistics and Linguistic Theory, 17*(1), 1–19. https://doi.org/10.1515/cllt-2020-0007
10. McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction (2nd ed.). Edinburgh University Press.
11. Liu, H., Hamida, L., & Wijayanti, S. (2020). From Collocation to Cultural Perspectives on Polygamy: The Study of Semantic Prosody of Poligami in Ayat-Ayat Cinta 1 and Catatan Hati Seorang Istri 1. *PAROLE: Journal of Linguistics and Education, 10*(2), 164-175. https://doi.org/10.14710/parole.v10i2.164-175
12. Pew Research Center. (2025). Countries with the most Muslims & global Muslim population change, 2010–2020. https://www.pewresearch.org/religion/2025/06/09/muslim-population-change/
13. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.
14. Sulaiman, M., & Yusof, A. (2021). Digital humanities and Islamic studies: Emerging approaches and challenges. *Journal of Islamic Studies and Culture, 9*(1), 12–23. https://doi.org/10.15640/jisc.v9n1a2
15. Wijayanti, S. ., Zulianingrum, A. ., & Muhjaj, J. B. . (2025). A Study on Microlearning Through TikTok: Utilizing Social Media Videos for English Language Learning. *Jurnal Ilmiah Profesi Pendidikan*, *10*(4), 3691–3696. https://doi.org/10.29303/jipp.v10i4.3910
16. Wijayanti, S., & Karimah, S. A. (2023). Efforts to increase English vocabulary skills using puppet theater and flashcards at TK Islam Amaryllis, Cibubur. Community Empowerment, 8(9), 1329–1335.

https://doi.org/10.31603/ce.9785

Figures

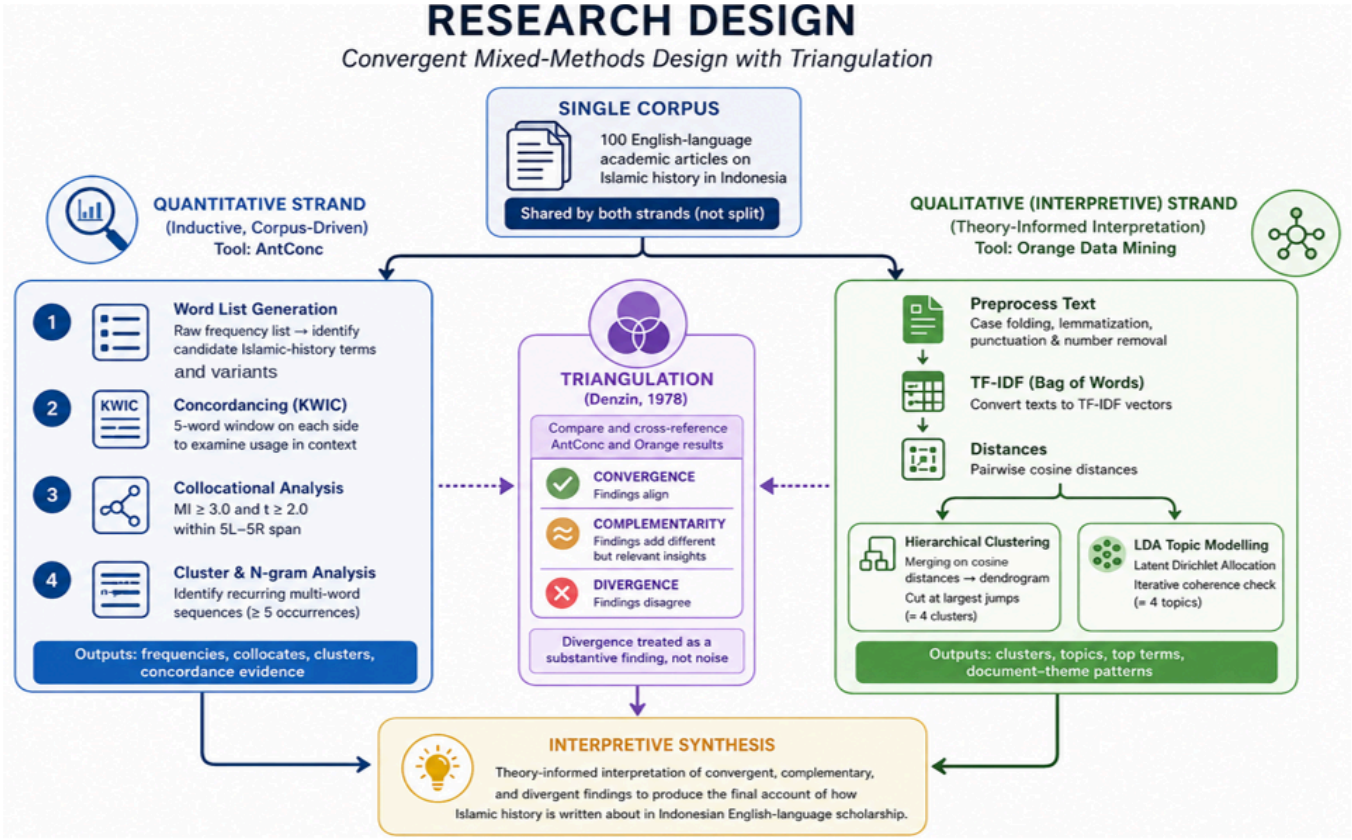


Figure 1

Overview of research design

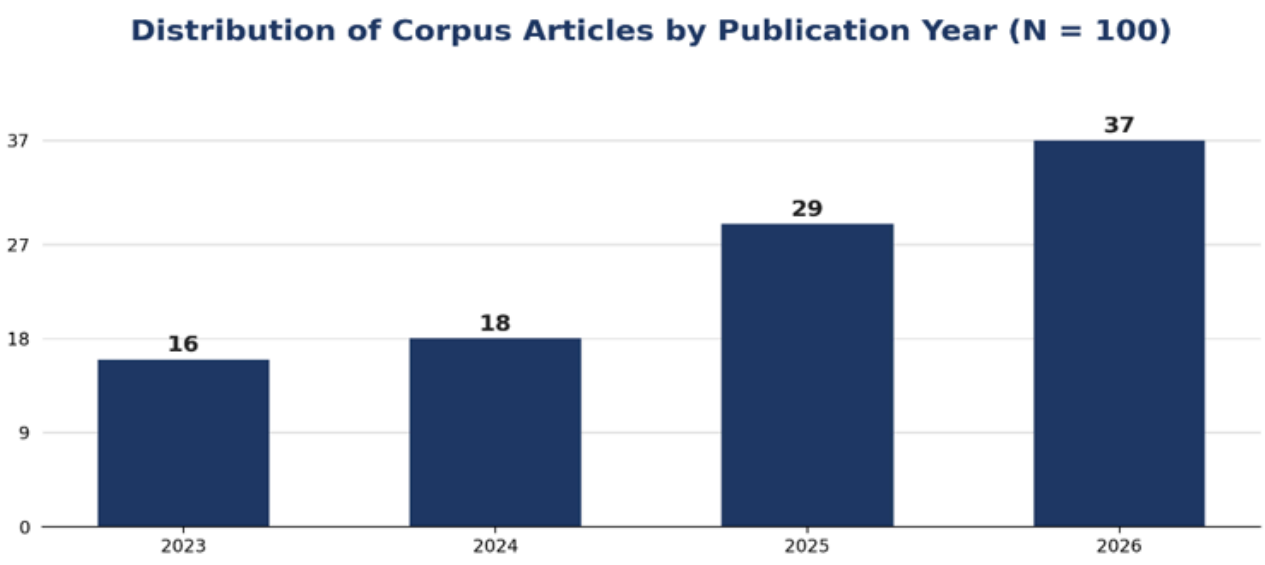


Figure 2

Distribution of corpus articles by publication year (N = 100)

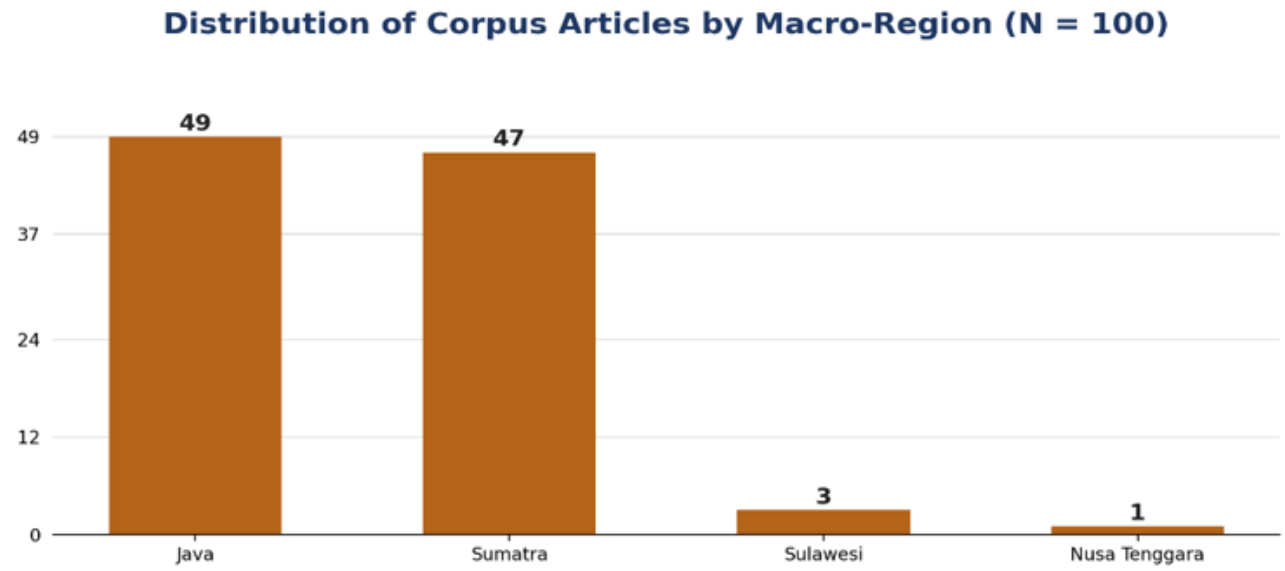


Figure 3

Distribution of corpus articles by macro-region (N = 100)

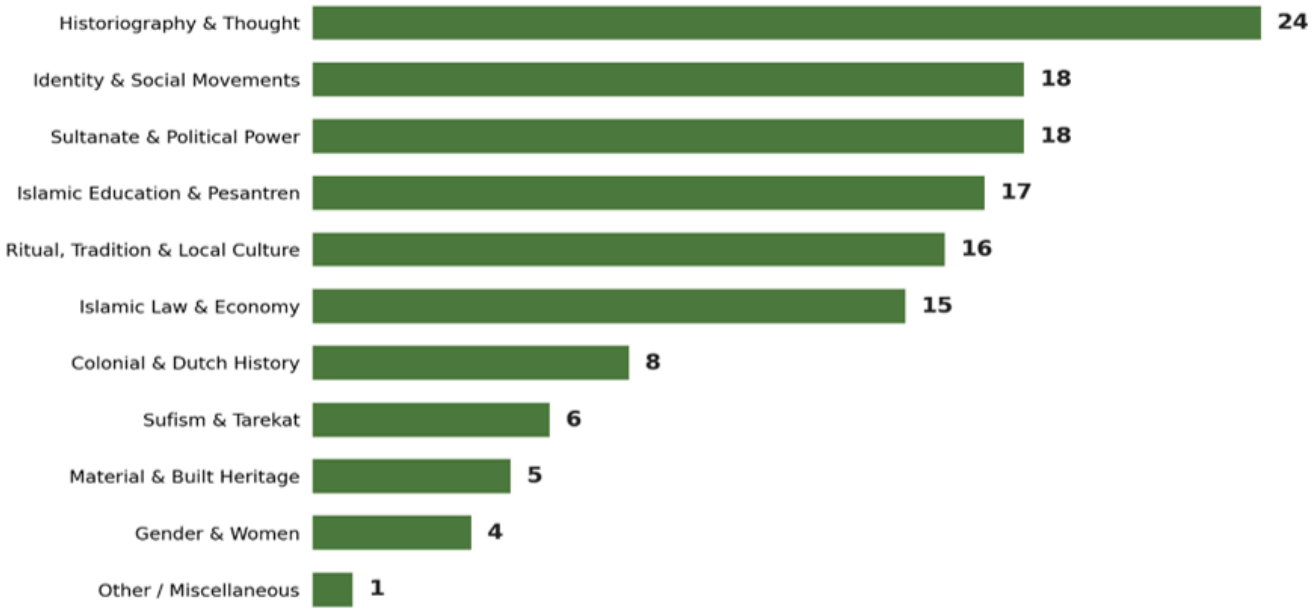


Figure 4

Thematic distribution of corpus articles classification (N = 100; multi-label)