Journal of Scientometric Research, 2024; 13(1):58-70.
https://www.jscires.org

Research Article

Comparing Research Topics through Metatags Analysis: A Multi-module Machine Algorithm Approaches Using Real World Data on Digital Humanities

Bhaskar Mukherjee, Debasis Majhi, Priya Tiwari, Saloni Chaudhary

Department of Library and Information Science, Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.

ABSTRACT

The present study extract, map and compare the lexical and semantic similarity of terms from author-provided keywords with machine extracted terms and topics from titles and abstracts of an inter-disciplinary field like 'digital humanities'. Author-provided terms (keywords) were first extracted and mapped through visualization software like Gephi and then these extracted terms were compared with terms extracted from title and abstract of the research articles through NLP based statistical modules. Also, the interdisciplinary of significant topics were measured through the Brillouin index. A set of 7483 articles downloaded from Scopus database on the domain of digital humanities and its associated fields were used for the purpose. We observed the researches on digital humanities are spread over a considerable number of concepts like 'Industry 4.0', 'topic modelling, 'open science'. Further, the machine algorithm-based extraction compared and identified a larger lexical similarity between these author-provided keywords and title-extracted keywords, rather than abstract-extracted keywords. Jaccard similarity of all author-keywords with machine extracted title keywords came 0.83 and SBERT BiEncoder_score was 0.7374. The top research areas extracted from titles, through unsupervised approach of term extraction resulted in topics like digital humanities approach, digital humanities visualization, indicating a strong connection to the discipline of digital humanities. The average interdisciplinarity index of top significant topics came between 1.217 and 1.284, with the highest index value for 'computational digital humanities'. As this study is based on real-world data, it is highly useful to understand how far machine algorithm-based text extraction can be helpful for information retrieval process.

Keywords: Text Analysis, Machine extraction algorithm, Keyword Extraction, Unsupervised algorithms, NLP, Digital Humanities.

Correspondence:
Bhaskar Mukherjee
Department of Library and Information Science, Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.
Email: mukherjee.bhaskar@gmail.com
ORCID: 0000-0003-2077-6976

Received: 07-08-2023;
Revised: 28-12-2023;
Accepted: 21-03-2024.

INTRODUCTION

In the ever-evolving quest for extracting the essence of knowledge through machine algorithm approaches, one has to embark on the journey of exploring the nuances and implications of metadata tags in scholarly research. Metadata tags such as title, abstract, and author-provided keywords of scholarly publications serve critical roles in mapping the content of the article. While the title serves as the first point or 'face' of contact for potential readers, the abstract provides a condensed overview of the research objectives, methodology, key findings, and conclusions, helping readers quickly assess the relevance and significance of the research. Therefore, title and abstract of a research paper should emphasize all significant words arranged in appropriate sequence as per the thought-content and subject of the paper.[1,2] A keyword (also known as index term or descriptor) on the other hand, is a term that captures the essence of the topic of a document or a search query.[3] Author-provided keywords are basically the key phrases or words provided by the author to reflect their personal view on the subject presented in a document.[4] However, a majority of research papers published nowadays lack significant keywords, compelling the researcher to spend his valuable time reading the entire document.[5] Evidences from the past also indicate that author produced keywords fail to showcase the interdisciplinarity of their publications, due to their biases or personal incompetencies.[6] Moreover, the manual process of assigning keywords is time-consuming and labor-intensive, making it less scalable.

Text similarity is a process to analyse the distribution of ideas, contents in a text corpus and establish interrelation among ideas, concepts with an emphasis on the findability of concepts. It analyses how 'close' two pieces of text are, either in surface closeness or meaning. The first is referred to as lexical similarity and later is referred to as semantic similarity. For lexical similarity, topics/texts sometimes display the distribution of knowledge in visual way and sometimes, in terms of their frequencies. The motto

OPEN ACCESS

EPUBL editing. publishing. technology

DOI: 10.5530/jscires.13.1.5

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

of such analysis is to discover the topic, the association and its occurrence. Although, this type of analysis also plays a significant role in resource indexing, cultural fusion and modelling,[7] it does not take into account the actual meaning behind words or entire phrases. Semantic similarity, on the other hand, measures the distance between the terms having likeness of their meaning or semantic content. In fact, it is a sub-domain of Natural Language Processing (NLP) that displays the results in terms of similarity score. Few well known techniques available to count the similarity score are Jaccard coefficient, Cosine similarity and recently SBERT cross-encoder and SBERT bi-encoder. The simplest way to understand the Jaccard similarity[8] may be explained as:

$$Jaccard = \frac{|\ tokens\_in\_string\_A \cap tokens\_in\_string\_B\ |}{|\ tokens\_in\_string\_A \cup tokens\_in\_string\_B\ |}$$

It is assumed that a threshold of 0.6 and above similarity score is effective in determining similarity between phrases/text corpuses. When the text becomes longer, this value could be smaller. Therefore, the similarity score can be smaller for abstract items than the title items. However, pre-processing and filtering of text enable to gain better similarity score even.

In the realm of text mapping, various types of analytical approaches have emerged to tackle the complexity of understanding and extracting meaningful insights from vast amounts of textual data. Visualizing software such as Gephi and VOSviewer, and machine-algorithm based approaches like TF-IDF (Term Frequency-Inverse Document Frequency), CounterVectorizer are now widely in use. Gephi, a powerful network analysis and visualization platform, focuses on visualizing and understanding the relationships between author keywords within a given corpus or dataset. By employing network analysis algorithms, Gephi facilitates identifying clusters, central nodes, closeness centrality, betweenness centrality, and patterns of keyword co-occurrence, enabling a comprehensive exploration of the underlying connection between the most prominent author-provided keywords. On the other hand, CounterVectorizer is a popular feature extraction approach-based algorithm in natural language processing that allows conversion of a collection of text documents into a numerical representation. It works by creating a vocabulary of unique words present in the corpus, removing punctuation, diacritics, numbers, and predefined stop-words and then counting the frequency of each word occurrence in each document. This approach was utilized to extract important keywords from the title and abstract fields in order to perform a linguistic similarity of the keywords obtained through Gephi.

CounterVectorizer simply counts the occurrence of a piece of word in whole text corpus, not necessarily most important keywords of a text. Unsupervised machine-extracted analysis, on the other hand, leverages automated algorithms which have been employed to extract relevant keywords from title and abstract fields and provide relevancy score of a term with respect to the whole text corpus. YAKE (Yet Another Keyword Extractor) is one of the significant unsupervised machine algorithms that employ a combination of statistical and linguistic methods to identify key terms in a document. It utilizes natural language processing techniques to assess the importance of each word within the context of the document, considering factors such as term frequency and document frequency.[9] Due to its functionality of depicting semantic relation between keywords, this algorithm has gained popularity in machine-extraction process.

Digital Humanities-A short note

The Italian Jesuit priest, Father Roberto Busa, is widely regarded as an early pioneer in the field now known as Digital Humanities (DH). In 1949, Busa embarked on an ambitious project to create an index of all the words in the works of St. Thomas Aquinas, comprising an astounding 11 million medieval Latin words. To accomplish this monumental task, Busa sought the support of IBM's CEO, Thomas Watson, and together they developed a punch-card lemmatized concordance.[10] Over the following decades, the field underwent various name changes, such as Humanist Informatics, Literary and Linguistic Computing, and Humanities Computing.[11] It was not until 2004, after 55 years of groundbreaking work, that Busa penned the foreword to "A Companion to Digital Humanities," officially introducing the term "digital humanities."[12]

Digital humanities has been interwoven into a number of connotations. Roth has identified various forms of understanding digital humanities such as 'digitized humanities' dealing with management of digitized archives, 'numerical humanities' putting emphasis on mathematical models and 'humanities of the digital' concentrating on digital media communication available for online communities.[13] Burghardt has explored yet another form known as public humanities, dealing with scholarly communication in the digital publishing, electronic learning and humanities.[14] Luhmann and Burghardt have pointed out that digital humanities is a highly interdisciplinary field with close ties to many neighbouring disciplines such as 'computational linguistics' and 'humanistic informatics'. Further, Puschmann and Bastos have attempted to compare academic networking platforms to serve communities in Digital Humanities.[15] They found that various emerging topics such as digital literature, digital archaeology and digital history are tracing the new computational areas of humanities research. Wang has examined 803 papers and 41 top keywords of DH research with strong relations to keywords such as digital history, digital libraries, text mining and humanities computing.[16]

However, despite the widespread adoption of the term, there is ongoing debate and disagreement regarding its precise definition and boundaries. Various scholars have offered different interpretations, ranging from using computational tools for humanities research to the critical investigation of humanities methods in the digital medium.[17] One of the earliest attempts by McCarty to define the field, then known as humanities computing was "a large methodological commons of techniques derived largely from and applicable across other disciplines.

These techniques depend for their application chiefly on the data in question rather than subject matter".[18] Svennson explained that "digital humanities may be described as a combination of models and practices where humanities scholars are increasingly turning to information technology both as a scholastic tool and cultural object in need of analysis".[19] John Unsworth expressed that "digital humanities is using computational tools to do the work of humanities" (p 67).[17] Julia Flanders asserted that "digital humanities is a critical investigation and the practice of the methods of humanities research in the digital medium" (p 69).[17] Ernesto Priego further had a bit divergent approach, "digital humanities is the scholarly study and the use of computers and computer culture to illuminate the human record".[17] Matthew K. Gold declared that "digital humanities is both a field with discernable set of academic linkages, practices, methodologies and a vague umbrella term used to describe the application of information technology to traditional humanistic inquiry".[17]

While the field continues to evolve, common hallmarks of digital humanities include the application of technology to research questions, interdisciplinary collaboration, iterative and experimental project approaches, critical examination of technology's role and impact, and the exploration of new questions and methods for analyzing data. In recent years, the research focus within digital humanities has expanded beyond textual analysis to encompass a wider range of objects and methodologies. This range consists of several prominent terms such as historical studies, data mining/text mining, archives, repositories, sustainability and preservation.[20] The growing involvement of diverse disciplines in digital humanities initiatives has highlighted the need for subject specialists to actively participate in research and provide insights on the further course of this rapidly emerging field.

Text Analysis-Researches

Several studies regard keyword as the core element of expressing topics[21,22] as these words are considered by author to be the most relevant to their research.[23] Word-frequency analysis has also been considered as the primary indicator for a topic's validity and highly occurred keywords are deemed to be the hottest topics.[23,24] Lu *et al.*[25] used author-keywords to represent topics and propose an effective ex-ante approach, namely Author-defined Keyword Frequency Prediction (AKFP) to detect research trends. This prediction relies on the Long-Short-Term Memory (LSTM) neural network. Keyword based analysis of research topics may be either frequency based or network based.[26] The network based analysis enables to identify co-occurred words, organization, countries and help in identifying the hotspots in research.[27,28] The temporal trend of Korean Research in medicine was analysed by using Gephi.[29] The network and clusters were shown in clusters through visualization method. Similarly, Jung and Lee[30] performed a keyword network analysis using Gephi on research papers on text mining that have been published among 45 disciplines. They used various centrality scores to measure the relative importance of terms in the text. Shen *et al.*[31] similarly visualized and analysed the hot topics in natural disaster research by using Gephi to generate a country collaboration and discipline collaboration network. In spite of having such advantages, this type of analysis is sometimes restricted to simple descriptions of the network[23] owing to the fact that this type of analysis is based on retrospective data evolution of keyword frequency in the future.

Few remarkable advances in artificial intelligence domain have led to significant production of literature related to supervised and unsupervised machine learning approaches.[32] Other extraction methods include N-Gram statistical information of words in the text corpus,[33] linguistic features,[34] or informative features like highlighted words.[35] In addition, extracting and clustering related words based on history of query frequency is also adopted.[36] In order to extract keyword that builds upon text statistical features to identify and rank the most important keywords in a text Campos[37] developed YAKE.

For understanding to what extent author keyword in an academic paper can be considered as a key element to understand the contents of the paper, Kwon,[38] using Brillouin Index, found that the interdisciplinarity degree of 80% of the keywords were as low as 0 to 0.499. The interdisciplinarity of author keywords in physical education was the least followed by social sciences and humanities. Only a few keywords had a fairly high interdisciplinarity and authors of these articles belong to different disciplines. Similarly, Chang and Huang[39] conducted a study on LIS field to explain the interdisciplinarity of the field. They found that the degree of interdisciplinarity has increased since 1978 to 2007 in LIS field, particularly in co-authorship and the sources of direct citation in LIS field are distributed across 30 disciplines.

Motivation

Our approach in the present study differs from the existing studies. With the increasing complexity and size of information in every domain, extracting keywords manually from large diversified text corpus is a challenging task. We focus the present study on how such task can be performed by NLP based approaches applied to real-world dataset using multiple modules. Our attempts were to map the author-provided terms first and then compare the lexical and semantic text similarity of author-provided keywords with the machine extracted title and abstract keywords. Furthermore, we identified top areas of research from the title and abstract corpus through unsupervised natural language processing algorithm and examined the extent of interdisciplinarity of a field like digital humanities, in terms of countries, institution's and author's involvement. We have neither relied on any learning graph theory based model nor supervised set of data or training stage and did not restore any linguistic tool to control the vocabulary.

Research Questions

The primary intent of this article is to advance the use of machine algorithm techniques to map the topics in digital humanities

research, by introducing a novel approach that has gained popularity in natural language processing and data science. The specific questions that we would like to excavate are:

RQ1-To what extent the author-provided keywords are lexically similar with machine extracted keywords;

RQ2-Among title and abstract, which field of machine algorithm-based extraction shows more semantic similarity with author provided keywords; and

RQ3-How significant topics can be extracted through unsupervised model for an interdisciplinary subject like 'digital humanities' and what extent of interdisciplinarity these significant terms show in terms of journal, organization, country, or domain?

METHODOLOGY

Data and Source

To obtain the data, we searched the Scopus database using the phrase search query in Title, Abstract, and Keywords' field. For the example in phrase search query "Digital Humanities", we performed the following search query TITLE-ABS-KEY ("Digital Humanities"). Similar search methods have been used for connecting keywords related to digital humanities, such as "*Digital Humanities*" *"Humanistic informatics,"* *"Literary and Linguistic computing"*, *"Computational Humanities"*, *"Digitized Humanities"*, *Numerical Humanities"*, *"Public Humanities"*, *"Computer and Humanities"*, *"Humanities of the Digital"*, *""Digital History"*, *""Digital Literature"*, *"Digital Archaeology"*, *"Computer and the Humanities"*, *"Computer and the Human"*, *"Linguistic Computing"* and "*Humanities Computing*" in the 'Title, Abstract, Keywords' field. These terms have been ascertained and selected after a careful examination of our backbone studies.[13-16] For this purpose, the search was conducted during the early days of January 2023 and a total of 7483 results were retrieved (research article and conference proceedings, books, book chapters). The downloaded articles were checked, refined, duplicates removed and formats other than research articles were removed. Thereafter, 6939 articles published in English language only during 1971 to 2022 were considered where 2219 articles lack the author's keywords and 488 articles lack an abstract. The majority of the papers were from the fields of computer science, social science, and the arts and humanities and more than 70% documents were open access.

Extraction of significant keywords

In order to extract keywords from the article, we approached three metatags of an article: author-keyword, title and abstract. For extracting significant terms from author-provided keywords, we used already existing dynamic algorithm-based visualization tool-Gephi that would not require learning graph theory. For extracting significant terms from titles and abstracts, we developed python-based algorithm using pre-existing module library. Author-provided keywords were imported in Gephi and the number of dynamic nodes and edges were counted. Using statistical feature, different centrality scores were measured in data laboratory. To measure the network of various concepts i.e., which nodes have more close relation and which node has more influence over the network with the central idea of digital humanities, we calculated the closeness centrality and betweenness centrality. In visualization, closeness centrality reveals how close a node is to the vertices. Whereas, betweenness centrality indicates how often a node occurs on all the shortest paths between two nodes. The value is calculated by dividing the number of nodes with the number of shortest paths linking these two nodes. The larger the value, higher the influence is within a network by virtue of their control over information passed between others. A node may have a larger presence in the network, but not necessarily their presence has major control over information passing between nodes.

As the author provided keywords may have limitations because of uncontrolled vocabulary; singular and plural form can be used in terms formality, to understand the similarity of author-provided terms with machine extracted terms, we applied two different approaches of machine extraction in the title and abstract of selected articles: term extraction and topic extraction. Through term extraction, we compared the lexical and semantic similarity between both series of terms, while through topic extraction we identified the significant topics and measured the interdisciplinary of these significant topics. The overall steps involved in the phase are: (1) Preparing the corpus of terms assigned by authors (human assigned) using pre-existing labelled program; (2) Pre-processing of candidate terms from title and abstract through personally developed algorithms; (3) feature extraction, n-gram generation and computing keyword score through personally developed python modules. (4) Post-processing involving data validation/similarity and classification into proper groups; and (4) Analysing the overall results from the post-processed corpus.

For pre-processing, we exploited *Natural Language Toolkit (NLTK)* Library and Panda Library of Python. First, tokenization was performed in which each word; symbol, special characters, etc were considered as a single token and converted into lower case. In the context of the scientific publication analysis, generic stop word lists fail to serve the purpose as they sputter many field-specific insignificant and redundant terms in the dataset.[40] To overcome this pitfall, an effort was made to create a dictionary of stop words taking into consideration the list created by Sarica and Luo[40] [Annexure-I]. After removing the stop words, lemmatization was done using WordNet Lemmatizer to correct the context of the inflected terms (like digital Scholarships >lemmatization > digital scholarship) and develop the 'bag of words' using the corpus. A Bag of Words (BoW) is a representation of text used in machine learning that keeps track of the occurrence of words in the text while ignoring syntax and word order.

After the pre-processing step, we performed machine algorithm techniques on the corpus. We used the *sklearn* library and from *sklearn.feature_extraction* module we have used *CounterVectorizer* for extracting the most frequent keywords using the

ngram_range=(2,3). *Scikit-learn* library in python provides this powerful tool to transform the given text into a matrix based on the frequency of words. This matrix or numerical vectors can then be used as an input for various machine learning algorithms, such as clustering or classification, to identify patterns and trends in the data. On the other hand, the *YAKE* technique was used to extract the keywords from the text corpus using the *yake.KeywordExtractor* module and *max_ngram_size* is fixed at 3 and the *deduplication_thresold* value is 0.9. YAKE is an unsupervised extraction technique that automatically extracts the keywords without being contingent on any external datasets or discipline. The overall procedures of algorithms that have been followed in this study are as follows:

Algorithm procedures 1: Extract keywords from target text through CounterVectorizer.	
pandas, nltk, numpy read_csv	Importing modules in algorithm. For corpus reading in pandas.
.lower() from nltk.corpus import stopwords stop_words.extend	Converting all terms into lower case. Pre-defined NLTK stop word imported. New stop words added to NLTK list.
import WordNetLemmatizer	Lemmatizing the words.
import CountVectorizer ngram_range=2,3	From sklearn.feature_extraction. N-gram value setting at 2,3.
Algorithm procedures 2: Extract keywords from target text through YAKE.	
import yake .read_csv .lower() from nltk.corpus import stopwords stop_words.extend import WordNetLemmatizer	Importing module. For csv reading in pandas. Converting all terms into lower case. Pre-defined NLTK stop word imported. New stop words added to NLTK list. Lemmatizing the words.
yake.KeywordExtractor()	Yake function calling.
deduplication_threshold = 0.9	Deduplication function.
deduplication_algo = 'seqm'	Deduplication algorithm.
max_ngram_size = 3	Maximum number of ngram size setting.

Measuring Interdisciplinarity

In the last stage, to analyse the extent of interdisciplinarity existing in the significant terms of titles, we explored the degree of interdisciplinarity using Brillouin index[41] by analyzing the journals, country, organization and subject of the significant terms. Following formula has been adopted to study the interdisciplinarity.

$$BID = \frac{\log N! - \sum(\log ni!)}{N}$$

where N refers to number of observations, and n_i refers to the number of observations in category i. In measuring the country, N refers to the total number of countries involved in all the publications and n_i refers to the quantity of countries involved in the publication of a selected topic. As the total number of countries, organizations and journals for N were beyond 171, to measure the Brillouin index we applied long_log formula. However, for calculating the log value for n_i, we chose a sampling process. A reference value (organization name, journal name etc.) appearing more than one time in a single article was counted only once and the name of organization, journals appearing more than nine times in the whole dataset was considered as sample under corresponding n_i for a topic. The domains of the terms were decided by studying the department of the author-affiliation field. For example, author affiliations with Department of Economics were kept to the discipline of economics. Authorial affiliations with Centre for Computational Engineering were classified under the discipline of computer science and engineering. In case it remained undecided, we first looked at the subject domain of the publishing journal and then, the abstract of the published article. After following all these steps, if the decision remained unsatisfactory, we excluded the record from the sample. Themes of the keywords were finally classified according to Scopus Subject Headings. The classification of significant terms was distributed among 33 classes from Scopus list (Annexure-II).

RESULTS

Mapping of terms of Author-assigned keywords

Table 1 shows the occurrence of significant terms in author-provided keywords and their relative role in the network. The term 'cultural heritage' has the highest occurrence and also the highest closeness centrality, but not the highest betweenness centrality.

It means the keyword 'cultural heritage' although is close to many nodes but does not play an important role in the flow of information from one node to other. Closeness centrality of digital libraries is higher than 'linked data', 'big data' reveals that although later have higher occurrence in the corpus, but they are not situated close to many nodes. On the other hand, keywords like 'GIS visualization', 'crowdsourcing' appeared less in number but these terms play a significant role in the flow of information. The term 'Digital Humanities' establishes triangles with quite a number of subjects among which artificial intelligence with machine learning established highest number of triangles followed by Metadata with linked data. And the Collaborative coefficient (CF) is observed to be higher with the subjects like Digital libraries Δ Digitization, Cultural Heritage Δ ontology, and Machine learning Δ Artificial intelligence. In visualization, triangles are established between three adjacent nodes and a higher value of clustering coefficient indicates a higher degree to which the three nodes in a graph tend to cluster together. It indicates the strength between important nodes over other nodes present in the network.

Table 1: Connections and strengthens of author-provided keywords on 'Digital Humanities'.

Author Keywords	Occu	CC	BC	Tri	Potent Node of Triangle	CF
Cultural heritage	105	0.356076	0.006813	968	Ontology	0.052797
Digital libraries	83	0.351587	0.007177	1016	Digitization	0.034554
Machine learning	78	0.335279	0.006488	984	Text encoding initiatives	0.037692
Artificial intelligence	74	0.342683	0.012986	1196	Machine learning	0.03368
COVID-19	64	0.335836	0.016972	742	Pandemic	0.024824
Digital history	63	0.329811	0.003018	432	Public History	0.031929
GIS Visualization	54	0.340947	0.006662	649	Spatial History	0.027692
Linked data	51	0.332651	0.001954	799	Semantic Web	0.076533
Big Data	49	0.334563	0.007672	572	Artificial Intelligence	0.029932
Pedagogy	46	0.318457	0.003762	365	Linguistics	0.07085
Semantic web	45	0.327945	0.00268	436	Linked Open Data	0.050428
Metadata	44	0.343779	0.004367	1130	Linked Data	0.056784
Crowdsourcing	43	0.33993	0.004186	616	Citizen Science	0.036193
Open access	43	0.327116	0.004128	899	Open Science	0.066445
Text mining	43	0.324433	0.001744	553	Topic Modelling	0.049485
Text encoding initiative	41	0.292649	0.002605	228	Industry 4.0	0.091529
Digital literature	39	0.309085	0.001113	398	Digital archaeology	0.043105
Social media	38	0.332984	0.005861	561	Computational Sol. Sc.	0.044104
Ontology	35	0.32563	0.002617	622	Cultural Heritage	0.052797
Digital Scholarship	34	0.284899	0.000644	556	Digital History	0.096227

CC=Closeness Centrality, BC=Betweenness Centrality, Tri= Triangles, Occu=Occurrence, CF=Clustering Coefficient.

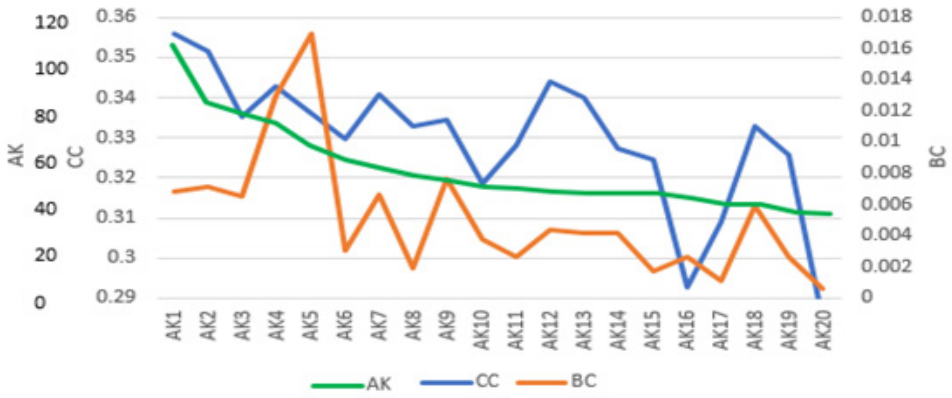

Figure 1: Relationship between AK, BC & CC.

The Figure 1 displays the relationship between number of Author Keywords (AK), their Closeness Centrality (CC) and Betweenness Centrality (BC).

As indicated in the Figure 1, terms with high closeness centrality and betweenness centrality do not necessarily have larger number of occurrences. However, terms having high betweenness centrality also have high closeness centrality, except the term COVID-19. Leydesdorff[42] found that betweenness centrality was an indicator of interdisciplinarity of journal in local citation environment and after normalization. In our study we found that terms related to GIS visualization, big data, metadata, and open access, social media have high BC and CC value, suggesting these terms have more interdisciplinary connection than others.

To visualize how various domains of knowledge are attached with digital humanities, we applied Gephi visualization tool. The lines between the two nodes indicate the linkage between the two terms and deepness of colour in the lines shows how much strong these links are.

As indicated in Figure 2 (selecting only those nodes that have 30 instances in the data), there are so many concepts with which digital humanities establishes linkage such as digital libraries, cultural heritage, text mining open access, ontology big data, etc. When we analyzed the nodes from the author extracted keywords through Gephi, we observed a number of nodes that were quite broad and had no use without the connotation. Terms like 'literature, 'acceptable', 'amount' 'predominantly' 'reasonably' etc. have no meaning until the connotation of these terms are defined. When it comes to a subject having interdisciplinary relationship, occurrence of such terms becomes problematic. We address this gap by rigorously identified insignificant, uninformative stopwords and develop a list of stopwords that are irrelevant in the subject. Our text corpus consists of 1426884 tokens (word, bi-gram, tri-gram) from almost 15246 titles and abstracts. Annexure-I enlists 100 such stopwords that this study has identified. We have appended these stopwords with the existing stopwords of NLTK, USPTO and the stopwords identified by Sarica and Luo.[40]

Lexical similarity of terms identified through Feature extraction

Next, we have employed machine algorithm-based feature extraction approach to the title and abstract of the 6939 articles. Table 2 indicates the lexical similarity obtained through a comparison of author-provided keywords with the machine-extracted title and abstract keywords, utilizing CounterVectorizer.

Author-provided keywords, on the character and word level, are much more similar to the machine extracted keywords from titles of research publications, in comparison to machine extracted keywords from abstracts of research publications. Various abstract keywords such as humanities social, digital media, cultural studies, network analysis and humanities project are lexically dissimilar to author-provided keywords. A majority of title keywords are lexically similar to the author-provided keywords, with the exception of only few keywords such as humanities pedagogy and art history. One drawback that may be observed while comparing text similarity of author-provided keywords with the machine extracted keywords is that few terms like ontology and digitization, that received attention in Gephi, do not receive the same extent of attention through machine extraction process.

Semantic similarity of machine extracted terms

In order to provide a robust assessment of keyword similarity, BERT is employed to compare similarity between author-provided and machine-extracted keywords using contextual embedding. First, both sets of keywords are tokenized and encoded into a sequence of token IDs suitable for BERT. The model then generates contextual word embedding capturing the meaning of each token based on its context. Pooling is applied to obtain representative embedding for each keyword sequence.

Next, similarity measures like SBERT BiEncoder_score and SBERT CrossEncoder_score are computed between the pooled embeddings, indicating their similarity. As is clear from Table 3, the similarity measure of title extracted keywords has a greater value than abstract extracted keywords, asserting that title extracted keywords are much more similar to author-provided keywords than abstract extracted keywords.

Unsupervised text extraction of most important keywords

YAKE employs unique statistical criteria to gather contextual information and word dispersion in an article. By analyzing factors like word casing, frequency, placement, context connectivity, and sentence difference, it assigns a score to each term to determine the saliency of each term. By identifying and extracting key-phrases that frequently appear and are contextually significant, YAKE can

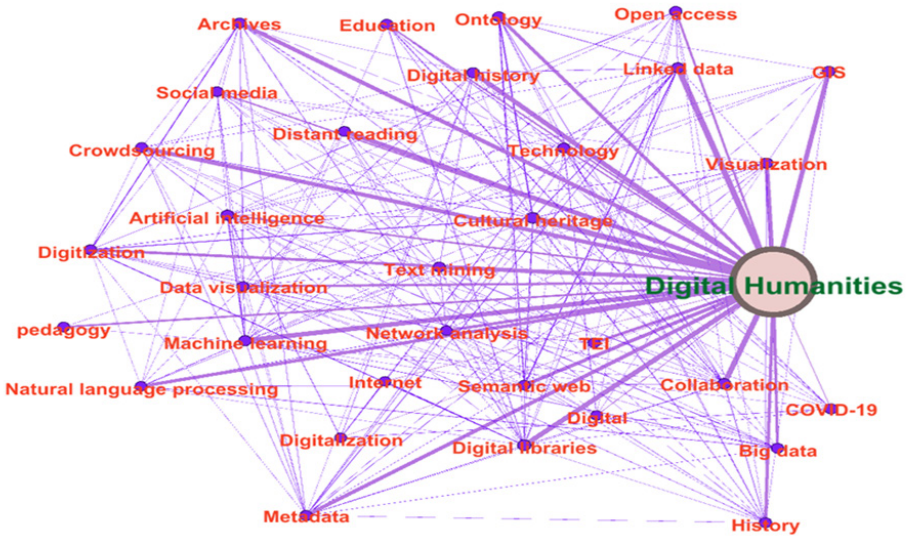

Figure 2: Connection between top nodes on Digital Humanities.

Table 2: Lexical similarity of author-keywords and machine extracted feature keywords.

Sl. No.	Author keywords	N	Title Words	N	Abstract Words	N
1	Cultural heritage	105	cultural heritage	153	cultural heritage	407
2	Linked data	83	COVID 19	100	big data	238
3	Big Data	80	big data	70	COVID 19	230
4	Machine learning	78	linked open data	68	machine learning	211
5	Artificial intelligence	74	machine learning	53	digital library	209
6	Digital libraries	74	digital scholarship	53	social media	206
7	Digital history	63	digital library	53	open access	203
8	COVID-19	51	open data	48	distant reading	201
9	GIS Visualization	51	humanities pedagogy	48	linked data	200
10	Pedagogy	51	artificial intelligence	48	humanities social	198
11	Semantic web	51	Culture study	44	arts humanities	193
12	Metadata	45	open access	41	digital history	187
13	Crowdsourcing	43	text mining	39	digital media	181
14	Open access	43	social media	36	cultural studies	176
15	Text mining	43	art history	35	open data	175
16	Text encoding initiative	41	digital history	34	artificial intelligence	169
17	Distant reading	38	distant reading	32	network analysis	166
18	Social media	38	semantic web	32	GIS	156
19	Ontology	35	Text encoding initiative	31	digital scholarship	154
20	Digital Scholarship	34	crowdsourcing	21	humanities project	144

N=Frequency of term in the whole corresponding corpus.

highlight important recurring phrases and concepts within the text.

Table 4 consists of most prevalent areas of research as identified from title and abstract corpus. Lower the score, greater the significance of the term. Accordingly, the top areas of research as identified from the title corpus are digital humanities approach, linked open data, digital humanities visualisation, and digital humanities platform. Similarly, top areas of research as identified from the abstract corpus include digital humanities projects, humanities social science, computational humanities curriculum, technologies and digital humanities research. It also indicates that most of the title extracted research areas have a highly relevant connotation to the discipline of digital humanities than the abstract extracted research areas.

Interdisciplinarity of top significant terms

Next, to measure the interdisciplinarity in terms of Brillouin's index of diversity, the methodology of choosing sample discussed earlier has been followed. The result, ranked according to decreasing BID, is shown in Table 5. The higher index value was observed for the topic computational digital humanities (1.2843) followed by digital cultural heritage (1.2740). Of these top twenty significant terms, the least interdisciplinarity has been observed for the term digital scholarship humanities (1.2175). Low value of 'n' does not mean that the said theme is unavailable in all journals [SD=12.24]/organization [SD=14.49]/domain [SD=5.45] or country [SD=1.75] etc., their appearance is relatively low on those journal/domains/organization that have been chosen as the sample in our analysis.

Table 3: Semantic similarity between author keywords and machine extracted keywords.

Similarity Methods	YAKE Extracted Keywords	
	Title	Abstract
SBERT BiEncoder_score	0.7374	0.6197
SBERT CrossEncoder_ score	0.6055	0.5210

While comparing the overall diversity ranks with the unique rank of each variable, (journal, organization, country, and domain) it was seen that when articles of a topic were proportionally distributed across different journals/organization/country/

Table 4: Top areas of research identified through unsupervised model.

Topic No.	YAKE_Title	Score	Abst No.	YAKE_Abstract	Score
T1	digital humanities projects	0.0000562	A1	digital humanities projects	0.00000209
T2	digital humanities approach	0.000103	A2	humanities social sciences	0.00000302
T3	digital humanities visualization	0.000135	A3	computational digital humanities	0.00000924
T4	digital scholarship humanities	0.000161	A4	digital humanities research	0.00000926
T5	digital humanities platform	0.000184	A5	digital humanities communities	0.0000120
T6	digital cultural heritage	0.000218	A6	technologies digital humanities	0.0000203
T7	digital humanities pedagogy	0.000275	A7	digital cultural heritage	0.0000213
T8	digital humanities literature	0.000294	A8	digital resources humanities	0.0000259
T9	digital humanities archaeology	0.000306	A9	digital humanities scholarship	0.0000268
T10	digital environmental humanities	0.000333	A10	digital humanities tools	0.0000308
T11	digital humanities collaboration	0.000361	A11	digital humanities approaches	0.0000324
T12	digital humanities research	0.000366	A12	digital humanities platform	0.0000329
T13	digital arts humanities	0.000606	A13	digital humanities collections	0.0000363
T14	computational digital humanities	0.000849	A14	practice digital humanities	0.0000400
T15	digital humanities social sciences	0.00193	A15	linked open data	0.0000444
T16	digital humanities framework	0.00219	A16	digital humanities literature	0.000312
T17	humanities semantic web	0.00300	A17	digital humanities pedagogy	0.000689
T18	linked open data	0.00104	A18	digital humanities visualization	0.000924
T19	digital humanities metadata	0.00283	A19	strategies digital humanities	0.00145
T20	open science humanities	0.00677	A20	digital humanities curriculum	0.0018

Table 5: Interdisciplinarity of significant keywords.

TN	TNA	BID	Overall Rank	Journal (Sample=93)		Organization (Sample=154)		Country (Sample=119)		Domain (Sample=33)	
				n	IRank	n	IRank	n	IRank	n	IRank
T14	74	1.2843	1	47	3	56	3	34	4	22	1
T6	151	1.2740	2	58	1	81	5	40	6	28	3
T5	86	1.2734	3	68	5	70	4	38	1	26	5
T12	110	1.2705	4	59	2	80	1	44	2	25	11
T8	136	1.2652	5	75	6	85	7	44	19	27	2
T7	110	1.2625	6	85	9	53	6	40	3	27	16
T19	105	1.2588	7	81	7	74	8	38	12	27	17
T9	90	1.2582	8	78	17	70	2	40	15	24	6
T1	74	1.2543	9	45	14	90	12	55	20	25	12
T13	113	1.2435	10	55	19	89	11	36	10	26	14
T15	81	1.2396	11	58	8	88	9	35	13	25	8
T2	77	1.2332	12	47	4	93	18	42	16	24	7
T20	87	1.2332	13	68	15	91	17	35	9	25	9
T3	88	1.2316	14	45	10	90	13	35	7	26	13
T16	98	1.2287	15	57	16	90	14	34	5	24	4
T10	145	1.2280	16	79	18	110	20	41	14	28	20
T17	95	1.2223	17	63	12	88	10	32	18	23	15
T11	90	1.2204	18	58	20	91	16	35	8	22	18
T18	97	1.2200	19	60	11	90	15	31	11	24	19
T4	97	1.2175	20	66	13	110	19	42	17	25	10

TN=Topic number, TNA= Total number of Article, n=number of occurrences of item under individual field, IRank= Rank individual category, BID= Brillouin's index of diversity, Rank of individual variable is based on Brillouin's index of diversity.

domain, the Brillouin's index of diversity was not necessarily increasing. Domain of specialization of contributing authors and journals of published articles shows more related or lower rank than the organization of contributing authors or their country.

DISCUSSION

The primary purpose of this paper is to map and compare the text similarity (lexical and semantic) of topics extracted from author-provided keywords with the title and abstract-extracted keywords. We analyzed a subject field that is reshaping the field of humanities and digital pedagogy. Before starting the lexical and semantic similarity processes, we first tried to visualize topics from the author-provided keywords by utilizing Gephi visualization software. We observed the researches on 'digital humanities' is spread over a considerable number of concepts like 'Industry 4.0', 'topic modelling, 'open science' along with related concepts like 'digital history', 'digital library' or 'cultural heritage'. The number of triangles and clustering coefficient of the former terms are higher as compared to others. The feasible explanation for the appearance of these terms could be attributed to the multidisciplinary of digital humanities, involving a variety of knowledge bases, disciplines, and specialties. A considerable number of terms on COVID-19' although seemed interesting but it provided a space for researches on evidence that COVID-19 is not a unique phenomenon in human history and by excavating the digital history one can learn lessons on how to handle pandemic beyond clinically. Higher closeness value of keywords like 'cultural heritage', 'digital libraries', 'artificial intelligence' or "GIS visualization' means direct and close connection of these terms with digital humanities. These terms also have 'independence' with respect to other keywords within the field of digital humanities. The correlation coefficient, betweenness and closeness centrality is 0.2745. On the other hand, high betweenness value of these nodes indicates that the local position of these keywords (node) with respect to the position of the nodes that it sits between is high; therefore they are important in the network.

We performed machine algorithm-based extraction through two state-of-the-art techniques using different modules of python library on the title and abstract keywords separately and combined and compared to what extent machine extracted terms are similar with author-keywords. We found large similarity in terms of character and word, to the title extracted keywords as compared to abstract extracted keywords. The Jaccard similarity of all author-keywords with machine extracted title keywords is

0.83 and 0.52 for abstract. This shows that title of the publications is highly representative of the subject of the paper. As title represents the main idea, it holds such words that are lexically similar with keywords. However, a few keywords like ontology and digitization that receive greater attention in Gephi are dropped in the machine extraction process. One of the possible reasons may be due to the Gephi's visual approach which might lead to subjective interpretations, while machine extraction tends to be more objective. Variations could arise due to data preprocessing, parameter settings, and algorithmic biases. Furthermore, while combining title keywords with abstract keywords and compared to author provided keywords, we found a large set of new words which might represent the subject of the paper but they faced a downfall in Jaccard similarity index. The basic reason behind their deterioration is that while adding abstract extracted keywords, it provides some more specific words which may not be of choice to the authors.

Although CounterVectorizer behaved well with our dataset for extraction of significant keywords from title, this module overlooked the semantic similarity. For this, we computed similarity measures like SBERT BiEncoder_score and SBERT CrossEncoder_score for both the title and abstract corpus. As can be observed from Table 3, there is a higher semantic similarity between the author-provided keywords and the title extracted keywords, as compared to the abstract extracted keywords. This can be attributed to the nature of the information conveyed in these respective elements. Titles and author-provided keywords are typically concise and carefully crafted to summarize the main theme of the document, making them more aligned in context and meaning. On the other hand, abstracts are more comprehensive and contain a broader range of information, leading to potential variations in semantic representation. BERT's ability to capture contextual understanding and word associations allows it to better identify the close relationship between title and author-provided keywords, resulting in a higher similarity score compared to abstract keywords.

As our intension was not only to know the content similarity of terms between two sets of metatags, but also to identify the top significant areas of publications through unsupervised model of term extraction and then the extent of inter-disciplinarity in terms of country, institutions, subject domain which exists in these top research fields. For this purpose, YAKE algorithm was utilized, which successfully identified and extracted keyphrases, shedding light on crucial recurring phrases and concepts within the domain of digital humanities. The results from the title and abstract corpus analysis revealed intriguing insights. The top research areas extracted from titles, such as digital humanities approach, linked open data, and digital humanities visualization, indicated a strong connection to the discipline of digital humanities. Conversely, the research areas identified from abstracts, like digital humanities curriculum and digital humanities projects, suggested a more diverse range of topics. This observation implies that titles provide a more focused and specific representation of research themes compared to abstracts, which tend to encompass broader aspects of the subject matter.

The average interdisciplinarity index in our study came between 1.217 and 1.284. The lower interdisciplinarity value for a topic does not mean that they appear in a smaller number of journals or domains, rather the topic may have skewed distribution in journals or domains. In some journals or domains, these topics are more prevalent as a result of which their total number comes in higher ranged, but they are not well distributed in different journals or domains. Examining the interdisciplinarity index of each topic, 'computational digital humanities' showed greater index value. This is because the topic frequently appears in diversified journals being published from different countries or under different domains or by authors affiliated with organizations belonging to different countries. When the rank of individual variable was compared with overall rank, it was seen that journal and domain of research are more sensitive variables to measure the interdisciplinarity compared to organization and country. And digital scholarship humanities followed by digital humanities metadata had the greatest difference in their ranks because these topics are field-specific and are rarely used in other fields.

Limitations

This research relies on the information found in the publications, but there are constraints associated with concentrating on a specific topic search within the digital humanities. If the terms used for the topic search are not present in the title, abstract, or keywords, there is a likelihood that literature pertaining to the chosen topic may not be retrieved from the database.[43] The results obtained from this study may not be readily generalizable to the entire field of Digital Humanities as it is predicated on a specific dataset, and the diversity and scope of research topics in Digital Humanities could exhibit variations. Lastly, this study primarily relies on data from the Scopus database. While Scopus is a comprehensive academic resource, the choice of a single database may introduce some bias into analysis, potentially overlooking relevant research not indexed within Scopus. Therefore, readers should interpret above findings with these limitations in mind, and further research incorporating multiple databases is encouraged to provide a more comprehensive perspective on research topics in Digital Humanities.

CONCLUSION

This research is significant in that it examined to what extent author-provided keywords are lexically and semantically similar with machine extracted terms from title and abstract and how these machine-extracted terms can be used to measure the interdisciplinarity. This research was felt necessary because author-provided keywords suffer from various limitations like uncontrolled vocabulary; multiple terms can mean one theme; and both singular and plural form can be used in formal terms, thus having difficulties in identifying the significant domains of a field. Through this study we tried to underline the importance of

machine extracted terms and justify how these terms can be used for further research by measuring interdisciplinarity characteristics of the significant terms. We observed machine extracted terms show almost equal lexical and high degree semantic similarity with author-provided keywords. And the machine extracted significant terms shows deeper level of subject contents than only lexical meaning, thus suggesting that machine extraction can be used as an alternative mechanism to understand the significant topics of a field and to measure the interdisciplinarity. While the network visualization of author-provided keywords shows how one term is connected with another, through machine extracted terms we were able to establish what extent of one variable of a topic plays role in measuring interdisciplinarity of a topic. Finally, we observed that interdisciplinarity of a topic does not necessarily depend on the total number of occurrences of a topic in a journal or domain, rather it depends on the proportional distribution of topic in different journals, domains etc. The importance of machine extracted text extraction helps researchers to gain a deeper understanding for dynamics of research trends and provide guidance for important topic detection. Therefore, these algorithms can be used not only to detect emerging or important topics but also to reveal obsolete topics and outdated technologies in advance, which can prove to be useful for policymaking or decision-making process, thereby combatting unnecessary economic losses. In addition, the proposed method of unsupervised text extraction can also be utilized in other prediction tasks having uneven data distribution. In future, this study may be expanded by analyzing other multidisciplinary subjects to uncover other aspects more closely related to the author's keywords.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Cals JW, Kotz D. Effective writing and publishing scientific papers, part II: title and abstract. Journal of Clinical Epidemiology. 2013; 66(6): 585. doi: 10.1016/j.jclinepi.2013.01.005, PMID 23434329.
2. Dewan P, Gupta P. Writing the title, abstract and introduction: looks matter! Indian Pediatr. 2016; 53(3): 235-41. doi: 10.1007/s13312-016-0827-y, PMID 27029687.
3. Howcroft G. A beginner's guide to metadata and keywords. Ed Bull. 2007; 3(3): 75-7. doi: 10.1080/17521740701788437.
4. Kevork EK, Vrechopoulos AP. CRM literature: conceptual and functional insights by keyword analysis. Mark Intell Plan. 2009; 27(1): 48-85. doi: 10.1108/02634500910928362.
5. Divakar G, Tyagi N, Saxena P, Verma V. Real world applications and research directions for machine learning: challenges and defes. IOSR J Eng. 2022; 12(1): 28-32.
6. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2): 165-93. doi: 10.1007/s40745-015-0040-1.
7. Park J, Topic mapping: A view of the road ahead. Lect Notes Comput Sci. 2005; 3873: 1-13. doi:10.1007/11676904_1
8. Jaccard P. The distribution of the flora in the Alpine Zone.1. New Phytol. 1912; 11(2): 37-50.
9. Campos R, Mangaravite V, Pasquali A, Jorge AM, Nunes C, Jatowt A. YAKE! Collection-independent automatic keyword extractor. Lect Notes Compt Sci. 2018; 10772. Available from: http://link.springer.com/10.1007/978-3-319-76941-7_80
10. Hockey S. A companion to digital humanities. In: Blackwell Publishing; 2011 . Available from: https://companions.digitalhumanities.org/DH/?chapter=content/9781405103213_chapter_1.html.
11. Nyhan J, Flinn A. Computation and the humanities: towards an oral history of digital humanities. Springer Nature; 2016. Available from: https://link.springer.com/10.1007/978-3-319-20170-2.
12. Sula C A, Hill, H V. The early history of digital humanities: : an analysis of computers and the humanities (1966–2004) and literary and linguistic computing (1986–2004). Digit. Scholarsh. Humanit. 2019; 34(1): i190-i206. doi: 10.1093/llc/fqz072
13. Roth C. Digital, digitized, and numerical humanities. Digit Scholarsh Humanit. 2019; 34(3): 616-32. doi: 10.1093/llc/fqy057.
14. Burghardt M. Theorie und Digital humanities-eine Bestandsaufnahme [internet]; 2020. Digital Humanities Theorie [cited Nov 8 2023]. Available from: https://dhtheorien.hypotheses.org/680.
15. Puschmann C, Bastos M. How digital are the digital humanities? An analysis of two scholarly blogging platforms. PLOS ONE. 2015; 10(2): e0115035. doi: 10.1371/journal.pone.0115035, PMID 25675441.
16. Wang Q. Distribution features and intellectual structures of digital humanities: A bibliometric analysis. J Doc. 2018; 74(1): 223-46. doi: 10.1108/JD-05-2017-0076.
17. Gold MK, editor. Debates in the digital humanities. Minneapolis: University Of Minnesota Press; 2012. 516 p.
18. McCarty W. Humanities Computing. Palgrave Macmillan; 2005. Available from: https://link.springer.com/book/9781403935045.
19. Svensson P. Humanities computing as digital humanities. Digit Humanit Q. 2009; 003(3).
20. Weingart S. the scottbot irregular. 2014; 1;2015. Submissions to digital humanities [cited Jul 17 2023]. Available from: http://www.scottbot.net/HIAL/?p=41041.
21. Peset F, Garzón-Farinós F, González L, García-Massó X, Ferrer-Sapena A, Toca-Herrera J, *et al.* Survival analysis of author keywords: an application to the library and information sciences area. J Assoc Inf Sci Technol. 2020; 71(4): 462-73. doi: 10.1002/asi.24248.
22. Trevisani M, Tuzzi A. Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories. Knowl Based Syst. 2018; 146: 129-41. doi: 10.1016/j.knosys.2018.01.035.
23. Huang TY, Zhao B. Measuring popularity of ecological topics in a temporal dynamical knowledge network. PLOS ONE. 2019; 14(1): e0208370. doi: 10.1371/journal.pone.0208370, PMID 30699118.
24. Zhao W, Mao J, Lu K. Ranking themes on co-word networks: exploring the relationships among different metrics. Inf Process Manag. 2018; 54(2): 203-18. doi: 10.1016/j.ipm.2017.11.005.
25. Lu W, Li X, Liu Z, Cheng Q. How do author-selected keywords function semantically in scientific manuscripts? Knowl Organ. 2019;46(6):403-18. doi: 10.5771/0943-7444-2019-6
26. Choi J, Yi S, Lee KC. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. Inf Manag. 2011; 48(8): 371-81. doi: 10.1016/j.im.2011.09.004.
27. Chang YW, Huang MH, Lin CW. Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. Scientometrics. 2015; 105(3): 2071-87. doi: 10.1007/s11192-015-1762-8.
28. Duvvuru A, Radhakrishnan S, More D, Kamarthi S, Sultornsanee S. Analyzing structural and temporal characteristics of keyword system in academic research articles. Procedia Comput Sci. 2013; 20: 439-45. doi: 10.1016/j.procs.2013.09.300.
29. Kim SK, Oh Y, Nam S. Research trends in Korean medicine based on temporal and network analysis. BMC Complement Altern Med. 2019; 19(1): 160. doi: 10.1186/s12906-019-2562-0, PMID 31277641.
30. Jung H, Lee BG. Research trends in text mining: semantic network and main path analysis of selected journals. Expert Syst Appl. 2020; 162: 113851. doi: 10.1016/j.eswa.2020.113851.
31. Shen S, Cheng C, Yang J, Yang S. Visualized analysis of developing trends and hot topics in natural disaster research. PLOS ONE. 2018; 13(1): e0191250. doi: 10.1371/journal.pone.0191250, PMID 29351350.
32. Papagiannopoulou E, Tsoumakas G. A review of keyphrase extraction. WIREs Data Min Knowl Discov. 2020; 10(2): e1339. doi: 10.1002/widm.1339.
33. Zhang C, Wang H, Liu Y, Wu D, Liao Y, Wang B. Automatic keyword extraction from documents using conditional random fields. J Comput Inf Syst. 2008; 4(3): 1169-80.
34. Hu X, Wu B. Automatic keyword extraction using linguistic features. In: Sixth IEEE International Conference on Data Mining- Workshops (ICDMW'06); 2006. p. 19-23. doi: 10.1109/ICDMW.2006.36.
35. Alguliev RM, Aliguliyev RM. Effective summarization method of text documents. In: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05); 2005. p. 264-71. doi: 10.1109/WI.2005.57.
36. Onoda T, Yumoto T, Sumiya K. Extracting and Clustering Related Keywords based on History of Query Frequency. In: Second International Symposium on Universal Communication. 2008;162-6. doi: 10.1109/ISUC.2008.22.
37. Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Inf Sci. 2020;509:257–89.

38. Kwon S. Characteristics of interdisciplinary research in author keywords appearing in Korean journals. Malays J Libr Inf Sci. 2018; 23(2): 77-93. doi: 10.22452/mjlis.vol23no2.5.
39. Chang YW, Huang MH. A study of the evolution of interdisciplinarity in library and information science: using three bibliometric methods. J Am Soc Inf Sci Technol. 2012; 63(1): 22-33. doi: 10.1002/asi.21649.
40. Sarica S, Luo J. Stopwords in technical language processing. PLOS ONE. 2021; 16(8): e0254937. doi: 10.1371/journal.pone.0254937, PMID 34351911.
41. Brillouin L. Science and information theory. 2nd ed. Oxford, England: Academic Press; 1962;xvii.:351. (Science and information theory, 2nd ed).
42. Leydesdorff L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. J Am Soc Inf Sci Technol. 2007; 58(9): 1303-19. doi: 10.1002/asi.20614.
43. Kim MC, Zhu Y, Chen C. How are they different? A quantitative domain comparison of information visualization and data visualization (2000-2014). Scientometrics. 2016; 107(1): 123-65. doi: 10.1007/s11192-015-1830-0.

Cite this article: Mukherjee B, Majhi D, Tiwari P, Chaudhary S. Comparing Research Topics through Metatags Analysis: A Multi-module Machine Algorithm Approaches Using Real World Data on Digital Humanities. J Scientometric Res. 2024;13(1):58-70.

ANNEXURE- I

Stopword list for articles on 'digital humanities'

['acceptable', 'amount', 'total', 'display', 'region', 'novel', 'combination', 'sheet', 'location', 'application', 'methodology', 'case', 'manifestation', 'related', 'comprises', 'provide', 'march', 'system', 'related', 'process', 'real', 'multiple', 'infection', 'report', 'single', 'center', 'hospitalized', 'ill', 'meta', 'analysis', 'long', 'term', 'review', 'cross', 'risk', 'factor', 'literature', 'papers', 'able', 'systematic', 'conference', 'case', 'study', 'accordingly', 'across', 'actually', meantime, 'nonetheless', 'obtain', 'omitted', 'beforehand', 'containing', 'contains', 'consist', 'detection' 'elsewhere', 'enough', 'following', 'furthermore', 'hardly', 'immediate', 'inc', 'indeed', 'instead', 'into', 'inward', 'identification', 'improve', 'include', 'just', 'regardless', 'regards', 'specifying', 'announce', 'information', 'novel', 'omitted', 'potentially', 'predominantly', 'presumably' 'reasonably', 'significant', 'significantly', 'substantially', 'successfully', 'sufficiently', 'suggest', 'thread', 'usefulness', 'unlike', 'sold', 'created', 'that', 'made', 'after', 'struggling', 'their', 'only', 'previously', 'leaving', 'acceptable', 'amount', 'reasonably']

ANNEXURE II

Scopus Code Broad domain

1201	Art and Humanities
1202	History
1203	Language and Linguistics
1204	Archaeology
1205	Classics
1206	Conservation
1207	History and Philosophy of Science
1208	Literature and Literary Theory
1210	Music
1211	Philosophy
1212	Religious Studies
1213	Visual Arts and Performing Arts
1300	Biochemistry, Genetics and Molecular Biology
1400	Business, Management, and Accounting
1700	Computer Science and Engineering
1900	Earth and Planetary Sciences
2000	Economics, Econometrics and Finance
2200	General Engineering (Tools and Techniques)
2216	Architecture
2300	Environmental Science
2600	Mathematics
2701	Medicine
3201	Psychology -General
3301	Social Sciences -General
3302	Archaeology
3304	Education
3305	Geography, Planning and Development
3308	Law
3309	Library and Information Sciences
3312	Sociology and Political Science
3314	Anthropology
3315	Communication
3316	Cultural Studies