Expert Systems With Applications 213 (2023) 119028

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Recent trends in mathematical expressions recognition: An LDA-based analysis

Sakshi, Vinay Kukreja *

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

ARTICLE INFO

Keywords:
Mathematical expressions
Research trends
Pattern recognition
Latent dirichlet allocation
Topic modelling

ABSTRACT

Context: Although recognition works on mathematical expressions have been explored for four decades, the current literature and trends are varied and frequently influenced by distinct emerging methods and technology. This situation instigates the necessity of an organized review to provide heedful insight into research trends and patterns currently prevailing in the domain of mathematical expression recognition (MER).
Objective: To identify and associate (semantic mapping) the leading research zones, core research areas, and research trends steering in the MER domain. Identifying prominent recognition models based on extracted research areas. To develop the development chart from extracted research trends for directing the future works in this direction.
Method: A manual and automatic search has been performed across the reputed digital libraries for corpus formation. The formulated corpus is used for topic modeling, and Latent Dirichlet Allocation is deployed for information modeling for achieving defined objectives.
Result: The corpus of 325 research papers published from 1967 to 2021 has been processed using LDA. The five major research areas and ten research trends are identified. Leading research area is "Segmentation and Classification Procedures", and the trend with the highest related publications is "Contextual and Graph-based recognition". "Attention and Deep Networks" has emerged as the newborn trend, and the identified newborn, young, and matured trends impetrate more exploration from the MER research community.

1. Introduction

Recognition of handwriting is one of the research areas that lie in the domain of pattern recognition (Shinde et al., 2018), image processing (Said et al., 2000), (Sen & Shah, 2017), computer vision (J. Hu et al., 1996), (Graves et al., 2009), and many more associated fields. The recognition of mathematical expressions and symbols has been a challenging domain for researchers since the 1960s. The hyping interest in recognizing handwritten mathematical characters and expressions has been captivated by the introduction of advanced recognition models and trends. Considering the present era where machine learning and deep learning techniques have eased the generation of identification and recognition models, this study aims to identify and portray the recognition solutions influenced by the varying ongoing research trends in this area. It has been vividly observed there have been comprehensive reviews and manual surveys presented in a way to depict the kind of technologies involved in this recognition process. Yet, there has been no successful evidence in the literature that could present a clear picturesque of recognition trends, particularly for MER. High research attention has been witnessed in this field of MER after 2011 (Mouchère et al., 2014). The reason for this emergence of revolutionary research, hitting for megatrends, is due to the orientation of the CROHME (Competition on Recognition of Online Handwritten Mathematical Expression) (Mouchère, 2011) series of competitions that majorly focused on achieving better recognition rates. This emerging literature on recognition problems has posed a challenge before the researchers to review and identify the right direction for their research. Thus, it needs to be exclusively addressed to draw the complete figure, illustrating the varied recognition techniques leading to the creation of a generation diagram that could depict the research trends more clearly for advanced direction in this domain.

The literature reviews and analysis presented to date have been manually compiled, leaving a decent scope for incompleteness, noise, bias, and uncertainty. The manual surveys and reports, though, are competent in providing a good insight into the literature, yet the stances of partiality, equivocation, and obscurity cannot be neglected.

* Corresponding author.
E-mail addresses: sakshi@chitkara.edu.in (Sakshi), vinay.kukreja@chitkara.edu.in (V. Kukreja).

https://doi.org/10.1016/j.eswa.2022.119028
Received 22 June 2021; Received in revised form 23 May 2022; Accepted 10 October 2022
Available online 13 October 2022

Meanwhile, the inevitable inclination towards the cited papers and bibliometric contents cannot be thoroughly exorable. To summarize, the manual review process is likely to be influenced by a distinct factor that can deteriorate the quality of the observation and extractions gathered. The advent of machine learning models and natural language processing algorithms has led to exclusive advancements in research patterns and methodologies. Specifically, natural language processing and its algorithms are potentially rich in identifying and extracting the unobserved trends and presenting the details in a more precise and careful manner, all just so uninfluenced by external uncertainties and partialities. This study has chosen to deploy the algorithm-based analysis rather than manual tagging. The manual tagging process involves intensive efforts yet fails to extract the results with all accuracy intact. The algorithm, i.e., chosen for the automated analysis to be performed, is reasonably popular in topic modeling (Canini et al., 2009), (Bird et al., 2015). The basic functioning of the model involves formulating the text corpora and feeding it to the model for pattern identification, and then performing semantic analysis on extracted patterns. Several clustering (Onan, 2019a), as well as topic analysis techniques, can be deployed along with topic modelling.

In contrast to clustering, the topic analysis is a more appropriate approach for recognizing research trends as per the findings of Evangelopoulos (Evangelopoulos et al., 2012). Ideally, there exist differences while carrying out topic analysis and clustering (Onan, 2017). In clustering, every document is assigned a fixed cluster of the topic, whereas, in topic analysis, a particular document can be allocated to a mixture of topics. In this automated review, topic analysis and labeling have been embodied in integration to identify and extract the latent patterns and trends in MER using the formulated corpus of relevant studies. The significantly popular trends in topic modeling are Latent Semantic Analysis or Latent Semantic Indexing (Deerwester et al., 1990) and LDA (Blei et al., 2003). This review will be concentrated on deploying the LDA model for performing analysis of trends in MER. For example, in the broad field of software engineering, LDA has been applied for mining software repositories (Thomas, 2011), bug localization (Lukins et al., 2008), defect prediction (B. Clark & Zubrow, 2001), software categorization (Tian et al., 2009), classification of change messages (Fu et al., 2015), and software evolution (Banitaan & Alenezi, 2015).

1.1. Significance of this study

This study reviews the current state of the art by implementing a topic modelling technique that has never been used in this domain of MER.

Being the first semi-automatic review, the focus of this study does not attempt to summarize the entire history of the mathematical expressions as done in the previous review studies. It aims to broadly identify the updated account of the state-of-the-art techniques deployed for mathematical text recognition based on the vocabulary and fetched key terms.

One of the popularly known surveys of this area has been performed in 2012 (Zanibbi & Blostein, 2012). Thus, the current trends in this domain have not yet been even covered by any study, and even the recent surveys (Zhelezniakov et al., 2021), (Sakshi & Kukreja, 2021) published lack the description of the ongoing trends. Consequently, this lengthy gap needs to be compiled wisely, and our study is evident in the kinds of recognition solutions prevailing the recent times.

The scope of the study gravitates around the application of LDA to the MER domain that assists in a semi-automated analysis for predicting the research zones, research areas, and recent trends. The study aims to instigate the overall status of the MER research domain by semi-automated means. The deployment of LDA equips not only the extractions of keywords and terms from the fed documents of literature (Onan et al., 2016b) but also added metadata details, culminating in providing journal-wise and year-wise publication sources analysis. Moreover, the extracted research zones, areas, and research trends have been semantically mapped for developing an indelve perspective of the domain.

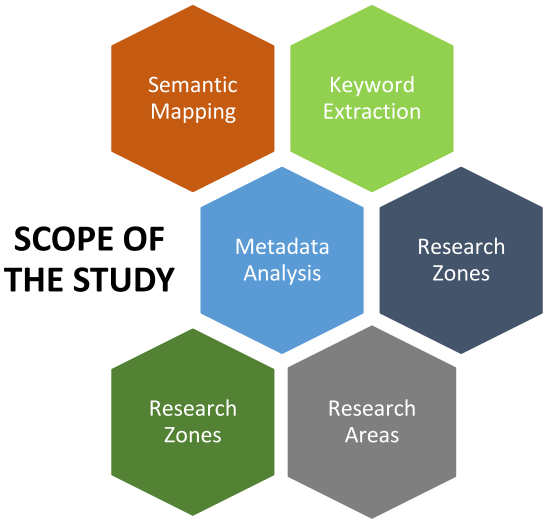

Fig. 1. Scope of the study.

Fig. 1 briefly displays the scope of the study.

1.2. Motivation for the work

- The primary concern that motivates this work is the lack of analysis; there is no recent research analysis that statistically compiles and reviews the contemporary literature and outcomes of the research patterns involved in this process of recognition.
- Though character recognition has been a widely popular discipline yet, the problem of MER needs distinct attention with concern to the level of challenges and difficulties involved in this domain.
- The modern era of artificial intelligence has drawn ways for several advanced machine learning and deep learning models to be wisely deployed in the identification and recognition processes to raise recognition accuracy to a potentially significant scale of comparison.
- Researchers have published many articles in the field of topic modeling (Onan, 2019c), (Onan, 2018b) and applied in various fields such as software engineering (Tamburri et al., 2020), political science (Grimmer et al., 2021), medical (Onari et al., 2021), and linguistic science (Toçoğlu & Onan, 2020), (Savin et al., 2021), etc. There are various methods for topic modelling; Latent Dirichlet Analysis (LDA) is one of the most popular in this field. So, the review team has decided to experiment with this topic modeling-based algorithm to extract the research trends in this field of MER.

1.3. The Highlights

- The study will be the crisp representation of research patterns in a semi-automated manner, irrespective of bias, noise, and uncertainty.
- LDA Model has been first used for the analysis of research trends in the field of MER.
- The automated analysis will offer statistically more accurate and correct results as the extractions here would be algorithm-dependent.
- This review report will provide an updated and mature state of the art with an emphasis on advances and neoteric recognition trends for mathematical expressions that have been uncovered by earlier surveys of this domain.

1.4. Preponderance of LDA (semi-automated analysis)

The study implements the LDA model (semi-automated analysis) that uses the mathematical model for interpreting the survey results and concluding key findings. On the contrary, the existing surveys have been compiled manually, involving the manual collection of the targeted literature. This study gets an edge over existing manual methods as its semi-automated. Thus it can target a sufficiently larger corpus. The manual reviews are subjective but, at times, may suffer from discriminatory aspects such as opinion biasness. This opinion bias can transpire due to erratic expertise, experiences, and investigative skills of the reviewers. This is a gap that can't be eradicated thoroughly (Barbara Kitchenham et al., 2009).

Moreover, manual reviews fail to explicitly distinguish and develop comparisons among inter and intra documents on the grounds of keywords (Hew et al., 2019), methods, and findings. Likewise, the elucidation of emerging research areas, trends, and prediction of future directions can be entirely dependent on reviewers' viewpoint and experience. Contrastingly, the semi-automated analyses of literature are all-inclusive and generic, with comparatively less biasness involved in interpretations. The fundamental grounds of differences that provide the preponderance of deployed analysis methodology over the existing ones are represented graphically in Fig. 2. It has also been observed that most of the available compiled reviews on MER literature seemingly neglected the prediction of research trends and categorizing the recognition techniques. This semi-automated analysis using LDA quantitatively analyzes the literature and equips in identifying and comprehending textual representation of documents under the MER domain (R. Rani & Lobiyal, 2021). There have already been many evidential instances of LDA being deployed in varied domains by different research groups (Onan, 2019b), (Onan & Korukouglu, 2017). To the best of our knowledge, this study is the first-ever implementation of LDA for trend analysis in the MER domain.

1.5. Review of reviews

Though the research in the field of recognition of mathematical notation has its roots in the 1960s (Anderson, 1967), (Anderson, 1977), (Chang, 1970), the popular survey has been witnessed in the year 2000, where Chan and Yueng (Chan & Yeung, 2000) surveyed the literature concerning two essential stages of recognition process that are symbol recognition and structural analysis. The primary emphasis of the survey was on similarities and differences among the systems. The review study constraining itself to a perspective of analysis from a stages point of view, partially addressed the issues related to the recognition problem. Thereafter, a survey has been presented by Tapia and Rojas (E Tapia & Rojas, 2007) 2007, which focused on architectures, symbol classification methods, and techniques for the structural analysis of mathematical expressions. The survey portrayed the techniques concerning the structural analysis, and again, the research trends have been manually addressed.

One of the last stances, a significant survey, has been accounted for in 2012, where Zanibbi and Blostein (Zanibbi & Blostein, 2012) made extensive efforts to compile the works of both recognition and retrieval. This study addressed the recognition and retrieval issues and thoroughly gravitated around four critical math retrieval and recognition problems. The maturity of aspects driven by this study was though fair. Still, again the proclivity for research pattern extraction was not incorporated, and the emphasis of the study was shared between two concepts of recognition and retrieval.

Thus, the prior literature analysis, reviews, and surveys have approached this domain with different conceptualizations. Apart from being manual, there have specific inclinations contemplated in the past reviews. The surveys and research trend analysis have already been rare and were active only a decade ago. The research undoubtedly has progressed, but a righteous trend analysis has been very timely. A few survey studies (Chan & Yeung, 2000), (E Tapia & Rojas, 2007), (J Zhang & Hong, 2008), (Zanibbi & Blostein, 2012), (Zhelezniakov et al., 2021), (Sakshi & Kukreja, 2021) have endeavored to compile the studies and works periodically. A mature survey highlighting the recent state of the art and depicting the current trends is still a need.

2. The review layout

The layout and organization of this paper are set as follows:

The first preliminary section encompasses all the prerequisite concepts associated with mathematical expressions and recognition problems. It also highlights some important definitions of the topic modeling process that will portray the outcomes in the coming sections of the paper. Section 2 briefs outline the entire workflow of this review analysis along with additional prerequisite information about MER. Section 3 will comprise the research questions targeted by this study. Section 4 will discuss the methodology of the review process, while section 5 will be incorporated the implementation of the LDA model. Section 6 will be constituted with the results of the deployed model and will also include discussions around the extracted outcomes. The threat to the validity of this study and the conclusions with future directions will be contained in sections 7 and 8, respectively. Last but not least, section 9 will comprise bibliometric contents that are references to the studies.

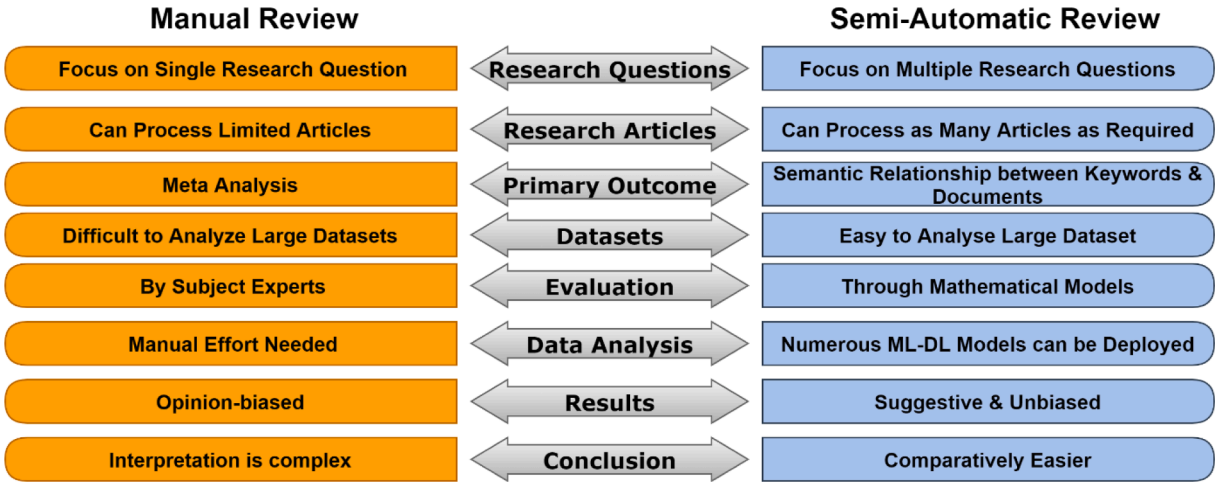

Fig. 2. Comparative Analysis of Manual Review and Semi-Automatic Review.

2.1. The review protocol and process

Considering the motivating factors and the present need of the hour, the reviewers have well planned all the protocols based on systematic literature review and PRISMA guidelines (Page et al., 2021) for literature selection. The foremost step includes the collection of appropriate corpora that include studies associated with mathematical expressions. The filtration process is carried out in the inclusion-exclusion phase, and thence, the research group formulated the corpus. This corpus is fed to the developed LDA model, and the resultant bag of words is grouped into topics and later labeled to extract the research patterns. Thus a level of semantic mapping has been achieved among research trends from extracted topics.

2.2. Mathematical expression recognition (MER)

The recognition problem of handwritten text or characters predominantly lies in the domain of pattern recognition. But the recognition of mathematical notation is very different than recognizing plain text or regular English characters. The two-dimensional nature of mathematical expressions makes the recognition process more challenging and technique-driven. Also, the inculcation of a large set of characters in the math dictionary of symbols calls for the requisite more elaborated symbol classifiers (Kukreja & Sakshi, 2022). Another difficulty relies on analyzing the intrinsic two-dimensional layout and ambiguities encountered in formulas, arrays of symbols, and diagrams, which require different treatment compared to plain handwritten text.

The two-dimensional notation of mathematical expressions and symbols facilitates easy communication and transmission of mathematics and visualizations of concepts and ideas. While defining the mathematical expressions, the usual background relates to mathematical notation to scientific and engineering background. Still, there exists no formal definition that clarifies the syntax and semantics of mathematics. As a matter of fact, the mathematical notation is also not completely standardized, and many dialects are used by scientists. To solve the problem of typesetting, many researchers use subject-dependent special mathematical notation. It has also been observed that mathematics and its notation are varied for distinct fields: the notation used for aeronautics can be different for mechanical engineering. Myriad science and engineering disciplines have an extensive dictionary of symbols, signs, and notations; thus, the context of each symbol and notation could be eclectic in different contexts. Fortunately, pen-based mathematics has gained the attention of the research community over the last decade, and thus, the recognition of the handwritten source has been targeted by many research communities. Also, the present era is evident in the wide-ranging recognition of research works available. Here, this report aims to compile and depict all the advances and research trends in the discipline of mathematical notation recognition.

2.3. The recognition problem

The recognition process involves three standard problems: symbol segmentation, symbol recognition, and structural analysis. These are also considered the significant subprocesses engaged in the recognition. At the initial stage, the input fed to the system can be in the form of a group of strokes, which are further segmented into a set of hypothetical symbols. The recognition of a hypothetical set of symbols is accomplished by the symbol classifier. At the final stage, the structural analysis is performed on the recognized symbols, which considers the geometrical and spatial contexts and relations of the symbols in the formula, expression, or equation and completes the recognition process by determining the final symbol based on the likelihood of association by applying the recognition model. This entire process has definite scope for ambiguities to be involved. A part of local ambiguities can be resolved using contextual information, whereas others could be independent of context. Such instances demand the verification of semantics of the mathematical expression, and that too is a time-consuming activity.

2.4. Topic modelling

Natural Language Processing (NLP) is an emerging field used by various researchers, and in NLP, topic modeling gained more attention in the field of text mining. It is a powerful technique used for text mining in data mining (Onan et al., 2016a). This technique is used to find the relationship between the data and the corpus document. This technique is used by various researchers in different fields like engineering, medical, web analysis, sciences, etc. To perform topic modeling, multiple methods are used, and among all techniques, LDA is ranked as the most popular technique. State of the art by various researchers shows that LDA is a valuable technique for analyzing recent trends in different fields. Topic modeling is an efficient technique that is a sub-domain of NLP used to discover topics and semantic mining from unstructured data (Blei et al., 2003). Topic models are known as outstanding in representing the isolated data. This technique is a productive approach to extracting enormous information from the corpus. It is impossible to provide information on all research carried out using LDA. Some significant areas are cited here. The highlighted fields where the application of LDA are medical sciences (Y. Zhang et al., 2017c), (Hordri et al., 2017), (Y. Wu et al., 2012), software engineering (Thomas, 2011), (Linstead et al., 2007), geography (Cristani et al., 2008), (Tang et al., 2012), political science (Greene & Cross, 2015), (B. Chen et al., 2010), and social media websites (Arun et al., 2010), etc.

3. Research questions

Research patterns in the field of MER have been systematically identified and represented in this study by applying LDA to a corpus of 325 articles published during the period 1967 to 2021. As many as five associated research areas and ten research trends have emerged after analyzing the titles and abstracts of research articles. Semantic linking between ten specific research trends and five research areas has been identified and presented. The review has been undertaken systematically, keeping in view the guidelines proposed by (Kitchenham & Charters, 2007) (Petersen et al., 2015). This review is intended to find the answer to the following research questions:

Research Question 1. Which research areas have been explored mostly by the researchers?

Rationale: This research question intends to identify and extract the research zones that have been spotted and determined by the researchers in the field of MER. Though the first evident study on recognizing mathematical notation was witnessed in 1967 (Anderson, 1967), the researchers have been actively working for the last two decades. Thus, there is a need to identify and figure out the research zones that have been frequently espied and recognized.

Research Question 2. What research methods have been used for the recognition of mathematical expressions?

Rationale: The objective of this question is to recognize and discuss the research methods and the recognition techniques that have been deployed for MER. It has been observed that the discipline of recognition models has been influenced by different models at different times. As per the past literature extractions, the recent trends in recognition have been unexplored distinctly. So, this question becomes very valid and essential for bringing research extractions that could guide and direct future works in this direction.

Research Question 3. Which research areas demand greater attention from researchers?

Rationale: This research question aims to diagnose and ferret out the research zones that demand more attention and need to be distinctly explored. After identifying distinct research trends, it becomes essential to determine which areas have suffered ignorance or biases over the others. This again becomes a question of concern for highlighting those

parts of the domain that are still in a challenging zone and need to be researched competently and in the future.

4. Methodology

This section involves all the aspects of experimentation and activities undertaken to perform the MER analysis. The stepwise procedure of whatever tasks have been completed is prominently explained, which picturesquely defines our research methodology to predict the research trends of MER. The flowchart of the research methodology is depicted in Fig. 3.

4.1. Corpus collection

4.1.1. Identifying sources of corpus collection

The main sources of data collection and formation of research corpus are online digital libraries, journals, conference proceedings, and some parts of gray sources like thesis and technical reports. The process of corpus collection engrosses distinct stepwise procedures, and the involved steps are explained below:

4.1.2. Defining the search strategy

The search keywords are decided based on the research questions of the current study. The research works of (Sehra et al., 2017) have influenced the formulation of the research questions for this study. The search phrases identified are "mathematical expressions", "math expression", "handwritten mathematical expressions," and "recognition". The search string used for searching is ("Recognition" AND "handwritten mathematical expressions" "mathematical expression" OR "mathematical expressions" OR "math expression"). Search terms "mathematical formulas", and "math equations" are included in the search string to broaden the search space for required articles on MER. The search criteria conformed to relevancy and recency.

4.1.3. Extending the search to other sources

An automatic search has been accomplished through relevant sources of information using defined search criteria by open-source tools Publish and Perish and search engines of specific publishers. The bibliographic databases of ScienceDirect, IEEExplore, Wiley, and ACM are explored, and identified articles are added to the BibTeX database of Mendeley. The bibliographic database search is meant for searching specific keywords in the publication title, abstract, and keywords. As many as 535 articles are collected in the BibTeX database. However, 450 articles remained in the database after removing duplicate entries.

4.1.4. Applying inclusion and exclusion filters

Under the inclusion procedure, the authors have considered the research papers that are published in English only. Then, the studies concerning the mathematical expressions are only considered. The studies considered for inclusion must revolve around the recognition aspect of mathematical expressions. The studies concerning mathematical equations and formulas are also to be included. For a study to be considered, it must be published in a verified, and the authentic source is one of our inclusion principles. Full reviews have been included.

Under the exclusion procedure, the authors excluded and removed those studies that have been published in different native languages. The primary focus is on gathering the research articles published in recognized journals; other than that, all the studies are excluded. The studies with mathematical flowcharts and diagrams are excluded. However, the papers describing the same research in more than one publication are not excluded. The publications not conforming to the defined inclusion criteria are reviewed manually by analyzing the metadata of their BibTeX. After this step, studies have been considered for the purpose of the current research and review. The task of the document collection

Table 1
Corpus Pre-processing.

Pre-processing Steps	Results
Sample Abstract	We address the problem of handwritten symbol classification in the presence of distortion modeled by an affine transformation. We consider share rotation scaling and transformation since these type of transformation that occurs in practice and focuses on shear in this framework.
After Tokenization	'We', 'address', 'the', 'problem', 'of', 'handwritten', 'symbol', 'classification', 'in', 'the', 'presence', 'of', 'distortion', 'modeled', 'by', 'affine', 'transformation','.', 'We', 'consider', 'share', 'rotation', 'scaling', 'and', 'transformation', 'since', 'these','type', 'of', 'transformation', 'that', 'occur', 'in', 'practice', 'and', 'focus', 'on', 'shear', 'in', 'this','framework','.'
After Stop Word Removal	'address', 'problem', 'handwritten', 'symbol', 'classification', 'presence', 'distortion', 'modeled', 'affine', 'transformation','consider', 'share', 'rotation', 'scaling', 'transformation', 'type', 'transformation', 'practice', 'focus', 'shear','framework'
Stemming	'address', 'problem', 'handwrit', 'symbol', 'classif', 'presen', 'distor', 'model', 'affine', 'transform','consider', 'shear', 'rotat', 'scal', 'transform', 'type', 'transform', 'practice', 'focus', 'shear','framework'
Lemmatization	'address', 'problem', 'handwriting', 'symbol', 'classify', 'present', 'distort', 'model', 'affine', 'transform','consider', 'shear', 'rotat', 'scal', 'transform', 'type', 'transform', 'practice', 'focus', 'shear','framework'

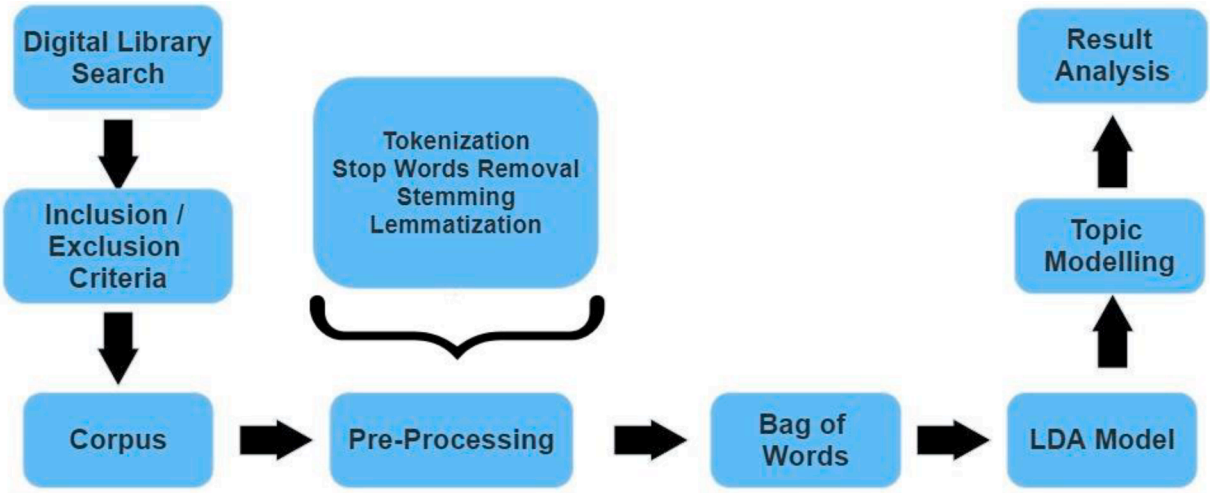

Fig. 3. Research Methodology.

accompanies the pre-processing part, summarized in Table 1. The whole of the research articles are included as a part of the corpus, but it becomes crucial to decide what contents of the article need to be loaded for processing of LDA.The first introductory research article (Blei et al., 2003) that proposes LDA for topic modelling is suggestive of using abstracts for document analysis. Also. Inspired and guided by other existing analyses on LDA (Griffiths & Steyvers, 2004), (Ponweiser et al., 2014), (Anupriya & Karpagavalli, 2015), (S.-W. Kim & Gil, 2019), (Chiyangwa et al., 2021), (Gil et al., 2021), where the input includes loading of abstracts to the model for topic modeling, the authors became convicted to input the abstracts for topic modeling. The purpose of deploying the LDA is to perform automated analysis of the bulk of studies which is otherwise be done manually.

When manual analysis of literature is completed, the readers screen the title of an article and parts of its abstract, and they try to determine whether or not to devote their scarce time to read on. Thus, the primary function of an abstract of a document is to signal its systematic methodology (Mo et al., 2015) and provide abstract knowledge of the article (Maier et al., 2018). For most readers, the findings described in the abstract are vital in determining the article's worthiness (Beller et al., 2013). Also, the coherence has been determined unaffected with consideration of abstracts over full texts when the document range is good (Syed & Spruit, 2017). So here, the abstracts of 325 publications are fed into the LDA model for analysis. A sample abstract undergoing preprocessing is illustrated in Table 1. The sample abstract is first tokenized, and the stop words are removed.

Further, stemming and lemmatization are accomplished. This preprocessed abstract is then fed to the LDA model. Similar to what is performed on sample abstract, pre-processing has been performed on all the 325 abstracts.

4.2. Preprocessing

It is a preliminary step that processes the dataset or the information collected. The objective of pre-processing is to discard the extraneous information present inside the gathered information. It systemizes the elimination of noisy words and characters from the dataset or the corpus collected and improves the dataset's quality. As a result, the profile of further processing becomes more accurate and acceptable.

The following steps have been incorporated for pre-processing the literature dataset (Table 1):

4.2.1. Loading the corpus

The corpus has been extracted from different sources and has been mustered up in a single.csv file. The manual filtration has already been performed in the inclusion-exclusion process. After the screening and selection, the duplicates have been demolished, and the final corpus consisting of 325 articles' abstracts is imported into the.csv file. Further, the.csv file is uploaded to the Google collab, where the experimentation is performed.

4.2.2. Tokenization

This is the step where lexical analysis is performed. All the abstracts per title are tokenized into tokens. The generated tokens are then transformed into lowercase letters for each document that has been considered. Thereafter we focus on the punctuation marks, characters, exclamation points, commas, apostrophes, question marks, quotations, semicolons, hyphens, and other punctuational marks. Further, any kind of equation or formula used in the abstract has been removed. Also, the numerical values are eradicated to get full-fledge textual tokens.

4.2.3. Stemming

Stemming is the process of reducing a word to its word stem. Stemming is essential in natural language understanding and natural language processing, which endeavors to extract the root or core word that is usually appended with the English suffixes and prefixes. It erases all the extraneous parts in the word and roots out the real, meaningful word. For example, "use" is the core word that can be extracted by stemming the word "useless", "useful", and "uses". To prepare an influential corpus, words are stemmed from their original form using the Snowball stemmer algorithm (Porter, 2001), and the resulting base keywords are stored in the cleansed corpus.

4.2.4. Stopwords removal

The stop words are the commonly used words such as "the", "if", "but", "a," or "an," etc. These words take up space in our corpus and consume valuable processing time. Thus, it becomes crucial to remove these stops removal, and here in our experimentation, we have used Natural Language Toolkit (NLTK) (Hardeniya et al., 2016). This toolkit has stopwords stored in more than sixteen languages. Here, the English language stopwords present in the NLTK library and other phrases used to build the corpus were removed from our cleansed corpus.

4.2.5. Lemmatization

The words which are previously stemmed need to be lemmatized. Lemmatization (Plisson et al., 2004) is when the context is considered and stemmed words are converted into more meaningful base words or lemmas. This phase targets removing inflected words and outputs the dictionary form of a word, as depicted in Table 1.

4.3. Phrase modelling: bi-gram and tri-gram model

The two words which occur together more frequently are named bigrams and the three words frequently occurring together in the document are termed trigrams. Here in this phase, the primary concentration is removing such combinations that frequently occur together, such as online_handwriting, offline_mode, etc. In this implementation, the gensim library has been used to remove such phrases. Gensim's Phrases model can build and identify these bigrams, trigrams, quadgrams, or even *n*-grams, and thus, we can make removal and make the data cleansing process better.

4.4. Applying lda model

The completion of pre-processing part, as suggested by the authors (Blei et al., 2003), (Mavridis & Symeonidis, 2014), led to the formulation of the processed corpus with the formulated dictionary called the bag of words with 2465 words(total vocab size). High-frequency words (with a frequency of more than 1000 occurrences) are removed from the bag of words. The corpus remained unaffected as the high-frequency words have to be less than 1000, and the maximum frequency ranged as high as 935. Then another filter is deployed for more qualitative results by removing the highest and least occurring words in the document. Both the extreme frequency range words are removed. Therefore, the words that are found to be occurring in more than ten documents and less than 50 percent of the total documents are also demolished. Finally, the bag of words formed after these filtrations for quality improvements is constituted 407 words. In other words, we can conclude that the bag of words is formulated after the most frequently and least frequently occurring are removed so that the corpus could become absolute.

This implementation of the LDA model has been performed using python programming (NLTK toolkit and NLP package named mallet). The LDA-based topic modelling has three input parameters that usually direct the entire experimentation process. Amidst the three input parameters, the authors have documented the number of topics and the hyperparameters α and β, and the number of iterations needed for the model to converge. α is the magnitude of the Dirichlet before the topic distribution of a document. This parameter is considered several "pseudowords", divided evenly between all topics present in every document, no matter how the other words are allocated to topics. β is the per-word-weight of Dirichlet prior over topic-word distributions.

Inspired by the first introductory work that proposed LDA, the Bayesian algorithm has been deployed to estimate the parameters for the LDA model. This algorithm for parameter estimation has been considerably supported by various LDA based works in the research community (Špeh et al., 2013), (Srihari, 2015), (Yang et al., 2021), (Zhao et al., 2021), (F. Wang et al., 2021), (M. Huang et al., 2021), (M. Wang et al., 2022). For identifying two, five, and ten topic solutions, as suggested by (Arun et al., 2010), the number of iterations considered is 200. The smoothing parameters can change the distribution over topics and words, respectively. Thus, the initialization of these parameters becomes a concern as the values can define the distribution of high-quality topic results. Fig. 4 represents LDA-based information modelling evidentially inspired by recent works of literature (Hidayatullah et al., 2019).

4.4.1. Algorithm-based choice of topics solutions

The α value has been kept as 1/T (Ethen, 2015), where T is the number of topics, and the β has been fixed as 0.01 for all topic solutions. For optimizing the hyperparameters, a java-based NLP package named mallet is used. The deployment of mallets accomplished the task of the number of topics. According to (Sehra et al., 2017), there is no established measure to defend the optimal number of solutions. However, heuristic parameters suggested by (Cao et al., 2009) and (Arun et al., 2010) can be applied to find the optimal range of topic solutions. The choice of the topic solution has been influenced by the heuristics and findings of the studies (Bradford, 2008), (Cao et al., 2009), (Arun et al., 2010), (Sehra et al., 2017). The choice of several topics is also performed algorithmically. An algorithm-based selection has been accomplished using the k means clustering algorithm; the optimal number of topic solutions for identifying research trends has been chosen. The deployment of k means clustering depicts the optimal choice of several topics for representing the core research areas is five. Thus, five topic solution has been selected optimistically. Also, the researchers have intended to explore a high-ranking value of ten so that the research trends could be more evidently discovered. The LDA model results have been used to analyze the metadata content and perform the *meta*-analysis. The *meta*-analysis thus performed facilitates to recognition of innovation literature (Leong et al., 2021). As a result of metadata analysis, the growth graph of research in this field over the years is depicted in Fig. 5. The dominating publication channels have been revisited and conclusively illustrated in Fig. 6.

4.5. Topic labeling

The authors have reviewed high-loading articles of all topic solutions in the current study. Further, the labeling of all the topic solutions has been performed individually to formulate the conclusive topic label. The task of labeling topic solutions has been carried out jointly by the two researchers, and the names of topic labels culminated after several rounds of brainstorming sessions and discussions. The task involved examining extracted terms and abstracts of documents related to a particular factor or topic, thus analyzing and interpreting the underlying research area or trend.

Determining a topic label is an arduous task that demands rigorous analytical ability and expertise. The two authors have consensually termed the research zones, areas, and trends based on the keywords fetched for every topic solution. According to the frequent practice in classical factor analysis, the authors have related each topic label to its key terms or high-loading terms and documents to assist in labeling every topic solution. Further, for every topic solution, a table listing has been maintained (refer to Table 2) that lists the prominent high-loading terms (key terms) and documents that load satisfactorily well (highly associated) on a particular topic or factor. Thus, the labeling has been accomplished manually, making this review semi-automatic. All the keyword extraction and fetching high loading terms and documents have been completed by the LDA model. Contrastingly the labeling of all topics or factors has been collectively done by the reviewers. Table 2 presents ten high loading terms as key terms and the high loading documents for each topic or factor in two, five, and ten topic solutions, along with a contribution value or factor-loading value. The contribution value or factor loading value for each topic specifies the extent of the relation of the related key term with a specific topic solution. The values shown in the "Contribution" column of Table 2 mean the probability values based on the estimated topic distributions. For example, "79.37″ in the first row indicates that the occurrence probability of the first topic in the topic distribution of paper with title 285 is 0.7937. The high loading terms and documents are those terms and studies, respectively, that load or relate sufficiently well to the identified factor or topic label. The several topic solutions, key terms (high loading terms), and high loading papers with their corresponding contribution value (factor analysis value) have been represented in chronological order in Table 2.

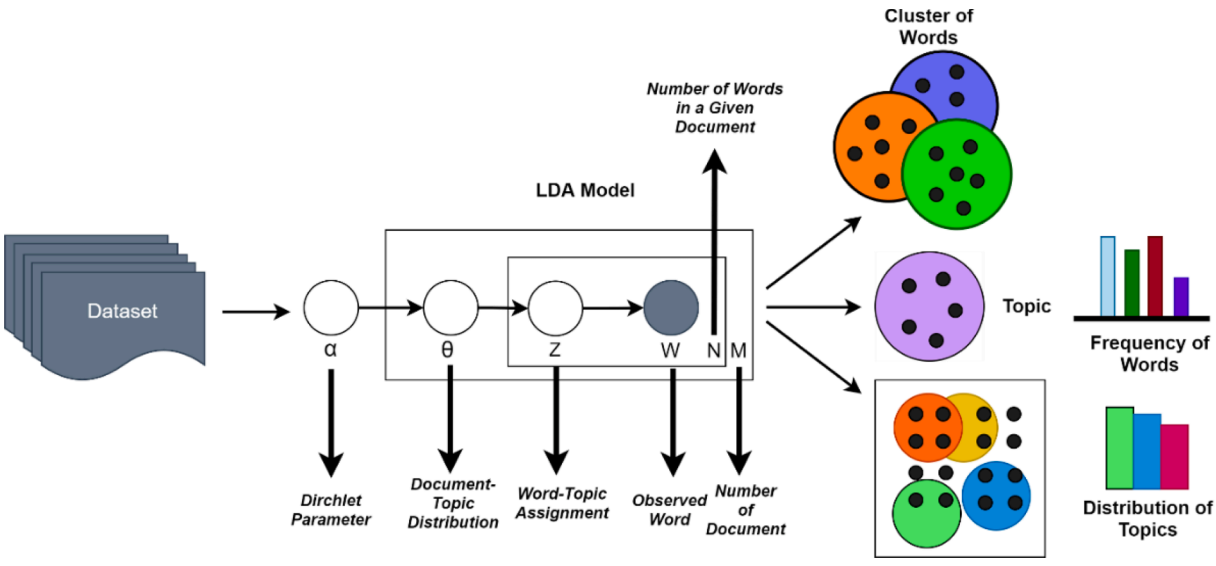

Fig. 4. LDA-based information modelling.

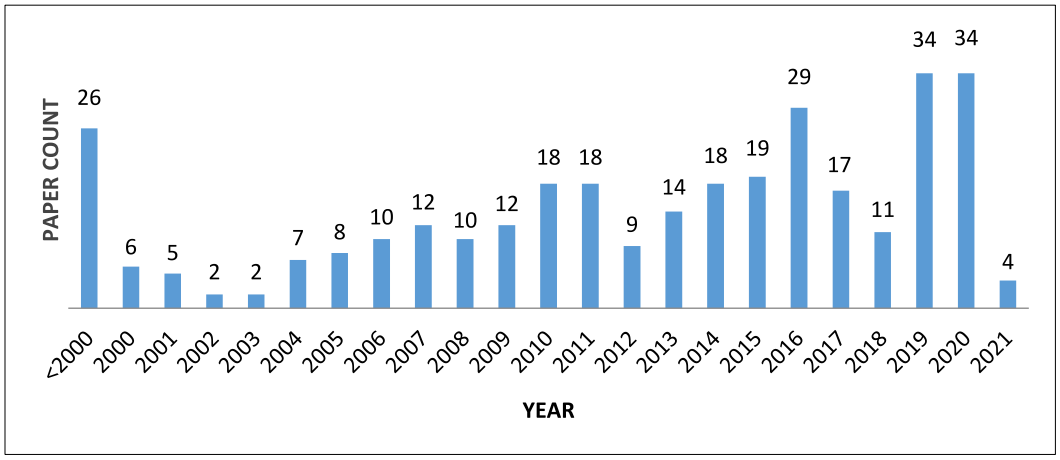

Fig. 5. Year-wise Publication Analysis.

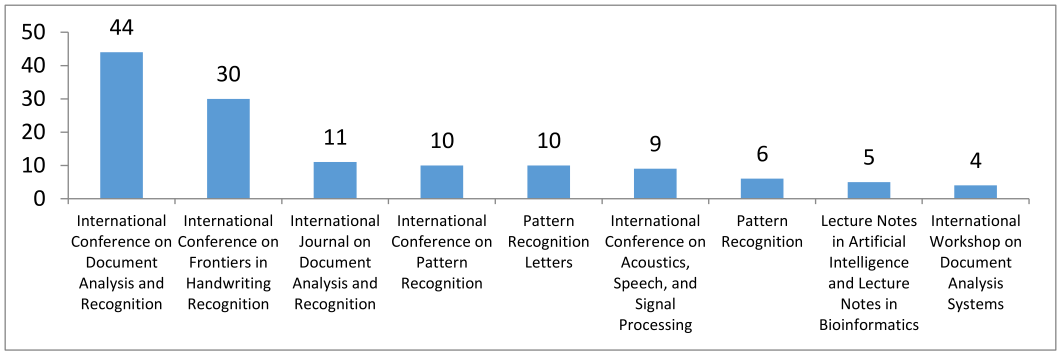

Fig. 6. Dominating Journals and Conferences.

5. Result analysis

5.1. Topic solutions in gist

The loadings for two, five, and ten topic solutions have been acquired by deploying the LDA model and are presented in Table 2. The selection of two, five, and ten topic solutions is based on k means clustering and is influenced by the previous studies (Arun et al., 2010), (Cao et al., 2009). Ideally, the coherence score ranging from the values of 0.3 to 0.6 (Röder et al., 2015) is considered good to have been used. The value of T (count of several topics) has been directed corresponding to the values of k and the coherence score graph. In the stances where high coherence scores have been achieved, the corresponding values of k have been used to choose several topic solutions. In the current scenario, the coherence score obtained for two topic solution is suggestive as 0.39, and the score obtained for five topics optimal solution is 0.41. And for the ten topic solution, the coherence value has been acquired as 0.36. Thus, five topic solution is the optimally best choice. The reviewers have also experimented with another number of topic solutions, but the values after ten topic solutions have declined. The dominance of each topic solution is also supported by the corresponding count of articles covered by it. Also, the different topic solutions could likely have the same research areas, but the number of studies covered by each topic has been wisely filtered for semantic mapping. Table 3 summarizes the count of publications corresponding to each topic solution.

5.2. Semantic mapping of extracted topic solutions

LDA model allows utilizing the number of factors extracted, for instance, the high loading key terms and high loading documents, to reach a level of aggregation at which common labels could be allotted to the extracted two, five, and ten topic solution. The structured arrangement of the topic solution, as shown in Fig. 7, states that at a high level of aggregation, core research zones and areas can be identified, whereas common research trends at the low level of aggregation. In a similar fashion, based on overlapping high loading documents and key terms among varied topic solutions, the authors have performed semantic mapping among the core research zones, areas, and trends. The detailed description of semantic mapping can be visualized in Table 3 and section 6.1. Apparently, the authors summarize the extracted topic solutions by plotting inter-relationships among them. The initial choice of two topics will broadly depict the core research zones that have been widely covered by the researchers in the compiled research literature. In five topic solutions, the authors have expressed the research areas that have been explored so far in detail.

Further, in the arrangement, as displayed in Fig. 7, the five topic solutions are widened into ten topic solutions with new trends emerging as the research trends in MER. This complete structured arrangement results from semantic mapping performed by the reviewers. Let's not confuse the LDA model being deployed hierarchically as it's certainly not true as the structured arrangement may look like a hierarchy, but it is accomplished as part of the semantic mapping process. The in-depth analysis of the process is vividly explained in section 6.1.

5.3. Two topic solution: Core research zones

The core research zones explored and discovered based on the two topic solutions have been depicted in the topics T2.1 and T2.2. Let us discuss how this labeling has been performed. While implementing LDA on two topic solution, the keywords and their loading has been

Table 2
High loading Research Articles for Topic Solutions.

Topic ID	Key Terms	Topic Label	Count	High Loading Documents	Contribution (%)
	Two Topic Solution				
2.1	method, analysis, result, problem, base, segmentation, stroke, structure, user, propose, recognize, approach, structural, input, character, process, document, present, technique, spatial	Recognition Techniques and Analysis Methods	96	Title 285 (Shan et al., 2021)Title 286 (Rhee et al., 2008)Title 97 (Le & Nakagawa, 2017)Title 126 (Dong & Liu, 2017)Title 108 (D.-H. Wang et al., 2020)	79.37 79.31 79.25 77.09 76.21
2.2	feature, model, propose, image, base, online, dataset, network, accuracy, classification, result, rate, neural, achieve, task, datum, attention, competition, present, method	Model Development and Functionality	229	Title 35 (Yingying et al., 2009)Title 291 (J. Fitzgerald et al., 2006)Title 19 (Taranta et al., 2016)Title 133 (Taranta and LaViola, 2015)Title 34 (Gong et al., 2015)	79.65 79.38 79.28 79.26 79.21
	Five Topic Solution				
5.1	structure, approach, analysis, tree, result, base, relationship, structural, propose, parse, spatial, graph, process, relation, problem, technique, present, recognize, interpretation, information	Parsing Techniques	45	Title 271 (Naderan, 2017)Title 123 (Ray Genoe et al., 2006)Title 215 (Celik & Yanikoglu, 2011a)Title 226 (H.-J. Winkler et al., 1995)Title 229 (J. Fitzgerald et al., 2006)	79.46 78.45 78.31 77.28 76.28
5.2	method, image, character, base, document, problem, segmentation, propose, present, provide, analysis, study, error, technique, performance, research, print, math, box, form	Character-Based Recognition methods	74	Title 166 (Chan & Yeung, 2000)Title 46 (Xiaorong & Chaoying, 2004) Title 35 (Yingying et al., 2009)Title 186 (J. Huang et al., 2020)Title 285 (Shan et al., 2021)	79.28 77.25 76.81 76.25 74.78
5.3	feature, model, propose, classification, base, rate, result, show, set, classifier, level, recognize, stroke, high, formula, achieve, accuracy, dataset, present, distance	Segmentation and classification procedures	92	Title 277 (Raymond Genoe, 2010)Title 285 (Shan et al., 2021)Title 286 (Rhee et al., 2008)Title 283 (H. Wang & Shan, 2020)Title 95 (L. Hu & Zanibbi, 2013)	79.59 79.21 78.59 77.31 73.31
5.4	online, task, dataset, neural, attention, network, competition, datum, propose, feature, model, result, image, performance, accuracy, base, approach, stroke, offline, crohme	CROHME and Neural Network Model	61	Title 143 (Álvaro, Sánchez, & Benedí, 2014b) Title 255 (Mahdavi et al., 2019)Title 88 (J. Wang et al., 2019)	79.69 78.47 70.05
5.5	user, stroke, interface, recognize, time, input, result, order, stage, single, method, line, online, enter, network, support, computer, write, combine, processing	Structural Analysis mechanism	53	Title 237 (L. Chen, 1992)Title 243 (Naderan & Zaychenko, 2013)Title 244 (S. J. Rani & Kumari, 2016)	79.73 79.63 76.51
	Ten Topic Solution				
10.1	problem, character, base, input, provide, math, develop, computer, mathematic, application, make, accuracy, equation, recognize, present, exist, research, dimensional, design, solve	Dimensional Model Construction and Offline recognition	24	Title 218 (Wells, 1976)Title 293 (Smirnova & Watt, 2010)Title 149 (Lee et al., 2018)Title 43 (Vuong et al., 2010)Title 185 (J Zhang & Hong, 2008)	74.39 73.22 66.28 64.18 63.28
10.2	method, document, image, propose, segmentation, print, problem, base, technique, work, extract, stage, accuracy, character, result, present, information, experiment, scientific, analysis	Parse Tree-Based recognition model	34	Title 188 (Phong et al., 2017)Title 140 (Álvaro & Sánchez, 2010)Title 221 (Pillay, 2014)Title 135 (Phong et al., 2020)	79.52 77.75 69.52 68.49
10.3	stroke, graph, order, parse, result, context, rate, model, grammar, time, base, score, parsing, recognize, algorithm, crohme, online, process, top, free	Contextual Mapping and Graph-based recognition	46	Title 215 (Celik & Yanikoglu, 2011b)Title 209 (Álvaro, Sánchez, & Benedí, 2014a)Title 197 (Shi et al., 2011)Title 172 (Le et al., 2016)Title 83 (L. Hu & Zanibbi, 2011)	79.61 78.61 77.14 72.13 70.46
10.4	user, interface, recognize, result, box, input, network, write, draw, matrix, enter, support, combine, output, stroke, neural, enable, mathematic, online, fuzzy	Input Methods	27	Title 217 (Büyükbayrak et al., 2007)Title 243 (Naderan & Zaychenko, 2013)Title 244 (S. J. Rani & Kumari, 2016)Title 101 (Ernesto Tapia & Rojas, 2005)Title 232 (Mahmoud et al., 2011)	99.79 99.79 99.79 99.57 99.53
10.5	method, approach, process, analysis, generate, present, reduce, propose, structural, segmentation, efficient, base, language, time, technique, problem, experiment, alternative, structure, strategy	Performance parameters and analysis	21	Title 233 (H.-J. Winkler et al., 1995)Title 48 (Kaplan, 2016)Title 29 (Phan et al., 2015)Title 164 (Phan et al., 2018)Title 179 (Xiangwei & Abaydulla, 2010)	79.59 77.58 61.44 59.32 43.13
10.6	propose, level, segmentation, stroke, spatial, online, set, information, classifier, structure, relationship, train, interpretation, offline, accuracy, label, input, result, global, evaluate	Segmentation and Spatial Constructs	33	Title 96 (A. M. Awal et al., 2009)Title 247 (G. Chen & Tang, 2013)Title 38 (A.-M. Awal et al., 2010a)Title 121 (Medjkoune et al., 2012)Title 201 (Rhee & Kim, 2009)	79.79 77.79 71.34 67.13 63.31
10.7	model, attention, network, propose, neural, end, base, latex, achieve, art, decoder, state, crohme, deep, structure, convolutional, dataset, accuracy, sequence, dimensional	Attention and Deep Networks	14	Title 277 (Raymond Genoe, 2010)Title 283 (H. Wang & Shan, 2020)Title 285 (Shan et al., 2021)Title 286 (Rhee et al., 2008)Title 27 (Shan et al., 2021)	79.77 78.71 76.35 73.26 72.49

(*continued on next page*)

Table 2 (*continued*)

Topic ID	Key Terms	Topic Label	Count	High Loading Documents	Contribution (%)
10.8	structure, analysis, tree, error, propose, approach, base, structural, spatial, relationship, problem, result, performance, method, evaluation, pen, construct, process, give, relation	Spatial relations and symbol identification	41	Title 118(A.-M. Awal et al., 2010c)Title 123 (Ray Genoe et al., 2006)Title 131 (Li & Tian, 2010)Title 34 (Gong et al., 2015)	79.44 78.44 77.62 76.32
10.9	feature, base, classification, dataset, model, result, image, method, present, set, approach, distance, match, math, datum, online, recognize, database, evaluation, show	Features based model development	43	Title 143 (Álvaro, Sánchez, & Benedí, 2014b) Title 127 (Yousefi et al., 2010)Title 204 (Álvaro et al., 2013)Title 265 (Okamoto & Higashi, 1995)Title 268 (Littin, 1995)	79.82 79.69 79.12 78.24 76.39
10.10	graph, feature, dataset, result, task, online, competition, datum, performance, accuracy, application, present, provide, test, evaluate, problem, research, base, experiment, offline	Online and Offline Recognition	42	Title 223(H. J. H.-J. H. J. Winkler & Lang, 1997)Title 224(Y. Hu et al., 2014)Title 295(H. J. H.-J. H. J. Winkler & Lang, 1997)Title 60 (Guo & Liu, 2018)	79.72 79.34 78.37 78.33

extracted. The extraction results of LDA depict the high loading articles per topic and the high loading terms or keywords per topic. The labeling process is based on the high-loading keywords that have been collected and have been performed consensually by the review pair. Thus, in Table 2, the labeling per topic solution has been achieved corresponding to the terms that have been extracted under the heads T2.1, T2.2, and so on; it goes for five and ten topic solutions.

- ***Defining The Labels For Topic Solution***

The two-topic solution presents an abstract view of the literature dataset and divides it into "Recognition Techniques and Analysis Methods" (T2.1) and "Model Development and Datasets" (T2.2). These are two significant labels that depict the core research areas that have been extensively explored by the researchers. These areas encompass the analysis of different techniques, proposed models, development, and recognition solutions for mathematical expressions. The keywords and their corresponding labels have been depicted in Table 4. Also, note that the entire set of keywords extracted under one topic can't be entirely suggestive for the labeling process. Thus, the labeling is a widened and broadened process that counts the total cumulative indication rather than being specifically suggestive.

5.4. Five topic solution: Research areas

To get a better insight into the research areas, the core research areas depicted in two topic solutions are further widened, and five topic solutions have been explored. In the five-topic solution, the keywords again played a significant role in the nomenclature of the topic solution. As per the keywords extracted, the labeling is accomplished. After being labeled based on the corresponding keywords' loading values, the identified topics become ready in the absolute state to predict and depict the major research areas that have been extensively researched in the domain of MER. The identified research areas are presented in topics (T5.1), (T5.2), (T5.3), (T5.4), and (T5.5). In five topic solutions, the emerged research areas are "Parsing Techniques" (T5.1), "Character-Based Recognition methods" (T5.2), "Segmentation and classification procedures" (T5.3), "CROHME and Neural Network Model" (T5.4), and "Structural Analysis mechanism" (T5.5). The role of each identified research areas can be distinctly visualized in Fig. 8 based on the count of studies or documents fetched under each topic. As a result of Fig. 8, it can be clearly analyzed that the topic "Segmentation and classification procedures" (T5.3) holds the maximum value and, thus, can be considered the most significant research area among all the five identified topic solutions or labeled research areas. Further, the researchers have endeavored to map the identified stigmas under the appropriated core research zones (T2.1) and (T2.2) in Table 5.

5.5. Ten topic solution: Research trends

Quantifying the number of topics extracted by the LDA topic model becomes necessary to get a detailed analysis of what research patterns can be observed through the surfaced keywords. The ten topic solution further resulted in comprehensive research trends in identifying and recognizing mathematical text. The ten topic solution led to the emergence of prominent research trends, including articles based on 80 articles on contextual mapping, parsing techniques, and grammar-based approaches, and 74 articles focusing on segmentation and spatial relationships. The ten-topic solution helps in surfacing some known and expected topics and has discovered and raised those research trends that have never been considered. In ten topic solutions, more research trends appeared, namely, "Dimensional Model Construction and Offline recognition" (T10.1), "Parse Tree-Based recognition model" (T10.2), "Contextual Mapping and Graph-based recognition" (T10.3), "Attention and Deep Networks" (T10.7) and "Features based model development" (T10.9). Other trends in our ten topic solution are "Input Methods" (T10.4), "Performance parameters and analysis" (T10.5), "Segmentation and Spatial Constructs" (T10.6), "Offline and Online recognition" (T10.10), and "Spatial relations and symbol identification" (T10.8). The high-loading terms of ten topic solutions and their corresponding count of documents or studies are depicted in Fig. 7. High-loading terms, top five articles for ten topic solutions with their corresponding count of studies are shown in Table 2. Note: here, topic terms are the high loading terms for each solution, and the count refers to the number of studies extracted under every topic label.

5.6. Highlighting research trends

The highlighting research trends have been depicted in ten topic solutions, where ten varied research trends have been identified and displayed. The most cardinal research trends that have surfaced are the topics labeled "Contextual Mapping and Graph-based recognition" (T10.3), Feature-based model development (T10.9), Online and Offline recognition (T10.10), and Spatial relations and symbol identification (T10.8). The extracted highlighting labels are depictive in picking and posing the current scenario of the MER research domain. The most pronounced trend is contextual mapped structures and graph-based recognition models. The recent literature (Lods et al., 2019), (Julca-Aguilar et al., 2020), (J. Wu et al., 2021) witnesses graph-based methodologies can be incorporated with the other machine learning-based algorithms to achieve a competent accuracy rate. This ongoing trend of graph-based recognition has come into the picture with the occurrence found in forty-six articles collected for the study. The maximum count for the topic "Contextual Mapping and Graph-based recognition" (T10.3) is suggestive to reveal that there has been continuous research over years upon this trend, which makes it the prominent and one of the highlighting trends. This pursuit of continuity of research on T10.3 is

Table 3
Year-wise Publication Analysis for 2, 5, and 10 Topic Solution.

-ID	Topic Name		less than2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
2.1	**Recognition Techniques and Analysis Methods**	Count	3	0	0	0	0	0	0	2	0	2	3	2	3	1	6	6	3	8	9	5	19	22	2	96
		Topic Ratio	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.20	0.25	0.11	0.17	0.11	0.43	0.33	0.16	0.28	0.53	0.45	0.56	0.65	0.50	
2.2	**Model Development and Functionality**	Count	23	6	5	2	2	7	8	8	12	8	9	16	15	8	8	12	16	21	8	6	15	12	2	229
		Topic Ratio	0.88	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.80	0.75	0.89	0.83	0.89	0.57	0.67	0.84	0.72	0.47	0.55	0.44	0.35	0.50	
5.1	**Parsing Techniques**	Count	2	3	0	0	0	2	3	2	1	4	0	1	3	1	2	0	3	3	0	4	4	5	2	45
		Topic Ratio	0.08	0.50	0.00	0.00	0.00	0.29	0.38	0.20	0.08	0.40	0.00	0.06	0.17	0.11	0.14	0.00	0.16	0.10	0.00	0.36	0.12	0.15	0.50	
5.2	**Character-Based Recognition methods**	Count	5	2	0	0	1	0	0	2	3	2	1	6	3	2	4	5	6	11	6	3	8	3	1	74
		Topic Ratio	0.19	0.33	0.00	0.00	0.50	0.00	0.00	0.20	0.25	0.20	0.08	0.33	0.17	0.22	0.29	0.28	0.32	0.38	0.35	0.27	0.24	0.09	0.25	
5.3	**Segmentation and classification procedures**	Count	3	0	1	0	0	0	1	1	0	0	2	2	3	4	6	6	4	8	8	3	16	23	1	92
		Topic Ratio	0.12	0.00	0.20	0.00	0.00	0.00	0.13	0.10	0.00	0.00	0.17	0.11	0.17	0.44	0.43	0.33	0.21	0.28	0.47	0.27	0.47	0.68	0.25	
5.4	**CROHME and Neural Network Model**	Count	12	1	1	1	0	2	1	4	3	2	4	5	5	1	2	3	2	5	1	1	3	2	0	61
		Topic Ratio	0.46	0.17	0.20	0.50	0.00	0.29	0.13	0.40	0.25	0.20	0.33	0.28	0.28	0.11	0.14	0.17	0.11	0.17	0.06	0.09	0.09	0.06	0.00	
5.5	**Structural Analysis mechanism**	Count	4	0	3	1	1	3	3	1	5	2	5	4	4	1	0	4	4	2	2	0	3	1	0	53
		Topic Ratio	0.15	0.00	0.60	0.50	0.50	0.43	0.38	0.10	0.42	0.20	0.42	0.22	0.22	0.11	0.00	0.22	0.21	0.07	0.12	0.00	0.09	0.03	0.00	
10.1	**Dimensional Model Construction and Offline recognition**	Count	2	0	1	2	0	0	1	0	1	0	1	0	0	1	1	3	4	5	0	0	1	1	0	24
		Topic Ratio	0.08	0.00	0.20	1.00	0.00	0.00	0.13	0.00	0.08	0.00	0.08	0.00	0.00	0.11	0.07	0.17	0.21	0.17	0.00	0.00	0.03	0.03	0.00	
10.2	**Parse Tree-Based recognition model**	Count	6	0	1	0	1	1	0	5	5	1	4	3	1	0	1	2	0	2	0	0	1	0	0	34
		Topic Ratio	0.23	0.00	0.20	0.00	0.50	0.14	0.00	0.50	0.42	0.10	0.33	0.17	0.06	0.00	0.07	0.11	0.00	0.07	0.00	0.00	0.03	0.00	0.00	
10.3	**Contextual Mapping and Graph-based recognition**	Count	5	1	0	0	0	0	4	1	0	2	1	2	3	2	2	0	6	2	1	3	7	3	1	46
		Topic Ratio	0.19	0.17	0.00	0.00	0.00	0.00	0.50	0.10	0.00	0.20	0.08	0.11	0.17	0.22	0.14	0.00	0.32	0.07	0.06	0.27	0.21	0.09	0.25	
10.4	**Input Methods**	Count	2	0	0	0	0	0	0	0	0	2	0	2	2	1	1	4	0	3	2	1	4	3	0	27
		Topic Ratio	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.11	0.11	0.11	0.07	0.22	0.00	0.10	0.12	0.09	0.12	0.09	0.00	
10.5	**Performance parameters and analysis**	Count	6	2	1	0	0	0	0	0	0	0	0	1	3	0	0	0	2	4	1	0	1	0	0	21
		Topic Ratio	0.23	0.33	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.11	0.14	0.06	0.00	0.03	0.00	0.00	
10.6	**Segmentation and Spatial Constructs**	Count	0	0	0	0	0	0	0	1	1	1	2	1	0	0	3	1	2	6	3	2	2	7	1	33
		Topic Ratio	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.08	0.10	0.17	0.06	0.00	0.00	0.21	0.06	0.11	0.21	0.18	0.18	0.06	0.21	0.25	
10.7	**Attention and Deep Networks**	Count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	6	6	0	14
		Topic Ratio	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.06	0.00	0.18	0.18	0.00	
10.8	**Spatial relations and symbol identification**	Count	2	1	0	0	0	1	1	2	1	1	2	3	1	0	3	1	1	3	5	2	5	4	2	41

(*continued on next page*)

Table 3 (*continued*)

-ID	Topic Name		less than2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
		Topic Ratio	2:26	1:6	0:5	0:2	0:2	1:7	1:8	2:10	1:12	1:10	2:12	3:18	1:18	0:9	3:14	1:18	1:19	3:29	5:17	2:11	5:34	4:34	2:4	
10.9	**Features based model development**	Count	2	2	2	0	1	5	2	1	3	2	2	6	4	2	1	1	3	0	1	0	1	2	0	43
		Topic Ratio	2:26	2:6	2:5	0:2	1:2	5:7	2:8	1:10	3:12	2:10	2:12	6:18	4:18	2:9	1:14	1:18	3:19	0:29	1:17	0:11	1:34	2:34	0:4	
10.10.	**Online and Offline Recognition**	Count	1	0	0	0	0	0	0	0	1	1	0	0	4	3	2	6	1	3	3	3	6	8	0	42
		Topic Ratio	1:26	0:6	0:5	0:2	0:2	0:7	0:8	0:10	1:12	1:10	0:12	0:18	4:18	3:9	2:14	6:18	1:19	3:29	3:17	3:11	6:34	8:34	0:4	

also indicative that the research the continued research is being followed in right direction.

6. Research questions and discussions

In the current study, the researchers aim to summarize the research trends in MER based on the selected corpus of 325 articles. The corpus formed after the filtration process includes articles from bibliographic databases from the first paper in this domain was first found. We have tried to compile the entire research literature from 1967 to the day 2021. Analysis of the corpus for n topic solution has been performed by applying LDA to find, extract and identify the latent research patterns and trends. In this section, an extensive technical discussion has been carried out for each formulated research question. Every answer below presents the gist summary of all the extractions after applying LDA. Thus, all three research questions have been discussed, given the literature dataset findings, and further future research opportunities have been explored and identified.

6.1. Research question 1

Which research areas have been explored mostly by the researchers?

To answer this question crisp, the authors can conclude that since the inception of mathematical expressions, various research trends have tentatively dominated the research era. The focus of researchers has been influenced by many factors, events, and technologies. The core research zones identified by the LDA model are T2.1 and T2.2, labeled as "Recognition Techniques and Analysis Methods" and "Model Development and Functionality". The two labels provided to the topic terms under T2.1 and T2.2 reflected the research zones predicted and concluded from the extracted high loading terms after applying LDA topic analysis. The extensive research in the MER domain is dominated by the research on recognition techniques, models, functionality, and analysis criteria. As a result of LDA analysis, the research areas in MER are highly concentrated with the aim of developing recognition models and techniques that could perform and analyze the outcomes more efficiently.

The year-wise count of articles for topic T2.1 and, T2.2 is scaled to values of topic ratios. The extrapolations and conclusions are inferred from the evaluated topic ratio values. There have been observations that following the year 2010, the topic ratios for T2.1 and T2.2 appear to show a range of hyping values. This implies that the CROHME series could be one of the contributing factors. As it is factual in 2011, there began an advent of competitions on handwriting recognition, popular as the CROHME series. The research zone seems to have been influenced by this competition series.

The values of topic ratios in T.2.2 are comparatively more than the topic ratios values in T2.1. This is suggestive of concluding that T2.2, labeled as "Model Development and Functionality," is more concentrated and focused by the research community. One of the possible reasons for this occurrence could make the advent of model-based strategies in the past decade, where machine learning and deep learning-based models have been proactively deployed for recognition, identification, and classification. The studies (Khuong et al., 2019), (Firdaus & Vaidehi, 2020), (Savchenkov et al., 2018) evidential under topic T2.2 are suggestive and give a clear picture of the techniques used in their implementation process.

While probing the five topic solutions, the defined labels will portray the research areas more lucidly. "Parsing Techniques" (T5.1), "Character-Based Recognition" (T5.2), and "Structural Analysis Mechanism" (T5.5) emerge from the topic T2.1, where recognition techniques and their analysis methods have been primarily focused. The articles extracted under these topic labels can be semantically mapped to the T2.1 topic solution. In contrast, the topics (T5.3), and (T5.4) can be semantically mapped to the topic solution (T2.2) labeled as "Model Development and Functionality". Eventually, the trend of this research

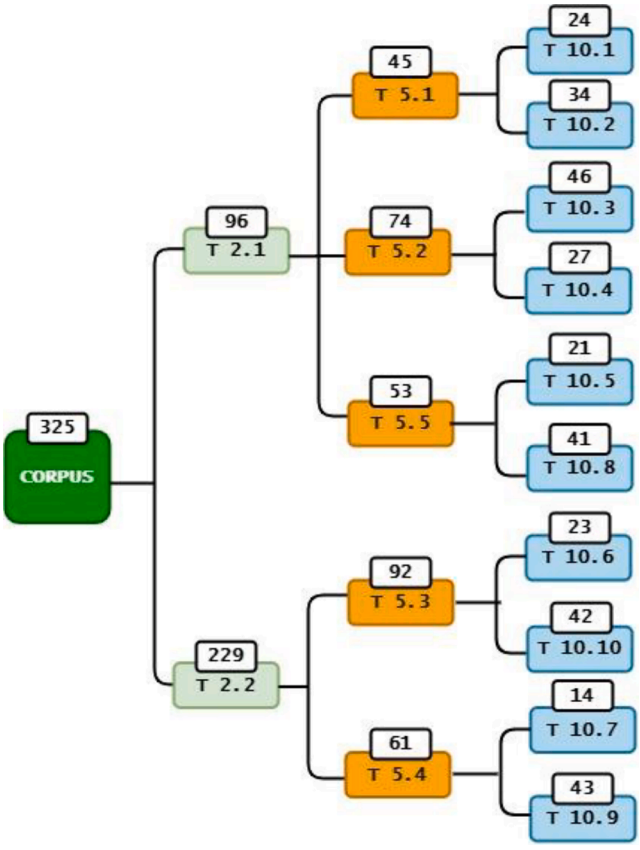

Fig. 7. Semantic Mapping for Topic Solution.

area is excellently visible in the topics (T5.3) and (T5.4). These topic solutions labeled as "Segmentation and classification procedures" (T5.3) and "Neural Network Model" (T5.4) also witnessed the evidently similar hype as observed in research trends on the topic (T2.2). The studies (Aguilar & Hirata, 2012) (Le & Nakagawa, 2013) apparently belonging to the topic (T5.4) also seem to be convicted, overlapping with the studies that emerged under the topic (T2.2). Thus, Segmentation and classification procedures" (T5.3) and the "Neural Network Model" (T5.4) have emerged from this topic solution. Similarly, there is also evidence of studies common in mapped topics like (T2.2) and (T5.3) (Golubitsky & Watt, 2010), (L. Hu & Zanibbi, 2013), and (T5.1) and (T10.1) (Anderson, 1967), (Y. Chen et al., 2000). The apprehensible semantic mapping of the topic solutions has been depicted in Table 4.

Further, the ten-topic solution gave us a delved profound view of the research trends. The high loading terms and labels of the extracted ten topic solution coherently present its mapping with the labels identified in the five topic solution. The research trends "Dimensional Model Construction and Offline recognition" (T10.1) and "Parse Tree-Based recognition model" (T10.2) discuss majorly parsing based methods and other dimension-based models which has emerged from the "Parsing Technique" (T5.1) which further correlates with the research zone "Recognition techniques and Analysis methods" (T2.1). The studies (Raymond Genoe, 2010), (Julca-Aguilar et al., 2015), (D.-H. Wang et al., 2020a) extracted under the mapped labels are also found to be overlapping that constitute the base of mapping criteria. The research zones mapped to research areas that are further mapped to research trends are entirely accomplished considering the number of titles/studies common to the considered threads. The research trend "Contextual Mapping and Graph-based recognition" (T10.3) and "Input Methods" (10.4) are emerging from the research area labeled as "Character-based Recognition" (T5.2). Usually, the modes of inputs (Breiner et al., 2017), (Kanahori et al., 2000) and contextual mapping (Garain & Chaudhuri,

Table 4
Semantic Mapping of 2 Topic,5 Topic, and 10 Topic Solution.

Two topic solution		Five topic solution		10 topic solution	
Topic ID	Topic label	Topic ID	Topic label	Topic ID	Topic label
2.1	Recognition Techniques and Analysis Methods	5.1	Parsing Techniques	10.1	Dimensional Model Construction and Offline recognition
				10.2	Parse Tree-Based recognition model
		5.2	Character-Based Recognition methods	10.3	Contextual Mapping and Graph-based recognition
				10.4	Input Methods
		5.5	Structural Analysis mechanism	10.5	Performance parameters and analysis
				1 0.8	Spatial relations and symbol identification
2.2	Model Development and Functionality	5.3	Segmentation and classification procedures	10.6	Segmentation and Spatial Constructs
				10.10.	Online and Offline Recognition
		5.4	CROHME and Neural Network Model	10.7	Attention and Deep Networks
				10.9	Features based model development

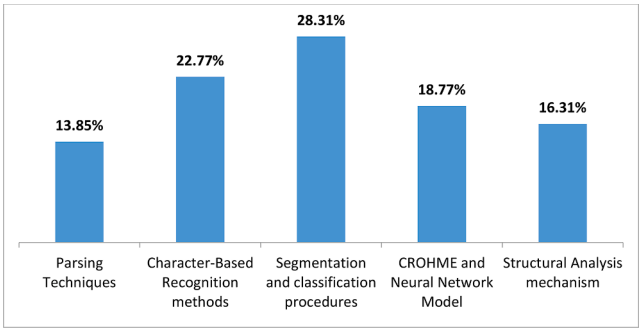

Fig. 8. Percentage of Article Loaded for 5-Topic Solution.

2003), (K. Kim et al., 2009), (A.-M. Awal et al., 2010b), (A. M. Awal et al., 2014) considered in the recognition process have been extensively mapped with the research area (T5.2). The core research area, "structural analysis mechanism" (T5.5), has engrossed the trend "Performance parameters and analysis" (T10.5), which describes the performance-related parameters used for assessing the performance of recognition models and their modification (Kaplan, 2016), (Phan et al., 2018) and "Spatial relations and symbol identification" (T10.8) (Mouchère et al., 2011), (Vuong et al., 2008). "Segmentation and Spatial Constructs" (T10.6), having a focus on segmented models (Basu et al., 2002), "Online and Offline Recognition" (T10.10) focuses on different modes of recognition (Hai et al., 2014), (Le et al., 2014), (Quiniou et al., 2011) has emerged from this research area. Similarly, the research trends

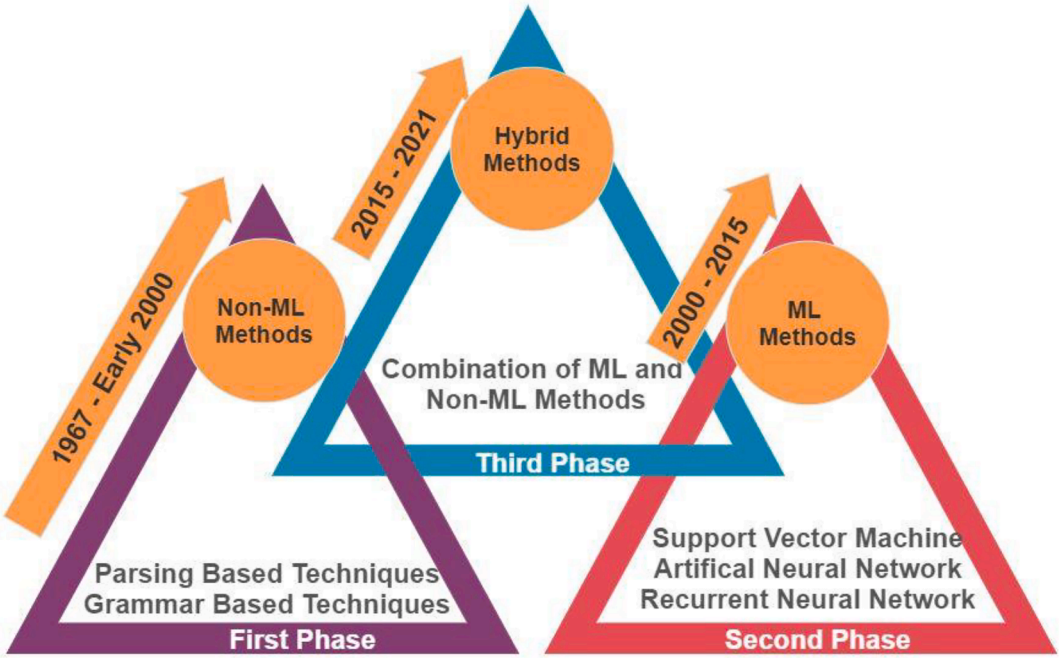

Fig. 9. Generation Diagram of MER.

"Attention and Deep Networks" (T10.7) and "Features based model development" (T10.9) covers the concepts of the attention-based mechanism (Zhang, et al., 2017b), (C. T. Nguyen et al., 2020), (J. W. Wu et al., 2020) and also feature-based recognition models (Onan, 2016), (Ramirez-Pina et al., 2018), (Onan, 2018a) that inherits the functionality from the machine learning algorithms (Savchenkov et al., 2018), (Gharde, 2012), (Zhu et al., 2013) can be coherently mapped to the research area labeled "CROHME and Neural Network" (T5.4). Certain other research trends have also been uncovered, such as "hybrid modeling patterns of recognition" (T60.54), which focus on calibration and combinational aspects of development (V. Nguyen et al., 2019), (A.-M. Awal et al., 2010a), (Mahmoud et al., 2011), and fuzzy logic-based concepts being reported by comparatively less number of studies (J. A. Fitzgerald et al., 2007), (Ray Genoe et al., 2006), (Lods et al., 2019) has been considered as an isolated topic.

Highlights (RQ1): The trends in the MER research domain have been influenced by several identified techniques and events. Almost 60 % of studies belonging to the topic "*Model development and Functionality*" (T2.2) evolved after 2010. Consequently, the inspiring and influencing event is found to be CROHME (started in the year 2011).

Summary and Challenges (RQ1): To understand the explored research areas, two, five, and ten topic solutions have been extracted and mapped. The significant research zones, areas, and culminating trends have been mapped semantically to display the likelihood of topics. This semantic mapping (manual) demands time and several brainstorming sessions to draw the structured mapped chart. The significant instances of overlapping articles and topics have been a substantial challenge in this process.

6.2. Research question 2

What research methods have been used for the recognition of mathematical expressions?

Based on the analysis of high-loaded articles, it has been investigated that various trends have kept on emerging periodically. The LDA model extractions have effectively displayed the occurrences and indicated the recognition methods that have been ever implemented to recognize mathematical expressions. The tentative investigations release the observations and results that major sub-domain of computer sciences steered the research race in MER. These major subdomains are namely character and pattern recognition, computer vision techniques, parsing methods, fuzzy logic, natural language processing, and machine learning. Currently, machine learning-based hybrid models have been frequently deployed for the recognition process. Thus, the extracted methods have been grouped into the categories termed non-machine learning-based models, machine learning-based methods, and hybrid recognition methods, as displayed in Fig. 9. The conventional approaches of recognition revolved around non-machine learning concepts such as "CYK algorithm" (Phan et al., 2016), "graph-grammar based models" (Julca-Aguilar et al., 2020), "stochastic context grammar" (Le & Nakagawa, 2016), and other parsing based algorithms. The research trends "Dimensional Model Construction" (T10.1), "Contextual Mapping and Graph-based recognition" (T10.3), and "Parse Tree-Based Recognition model" (T10.2) apparently depict and support this category research method. The researchers have independently chosen diverse computer vision and pattern recognition methods for the recognition process. The dataset considered has also been varied and insufficient. The second identified category of recognition method has been actively implemented since 2003 (Ernesto Tapia & Rojas, 2003). Here there have also been observed trends in the recognition methods identified. The early phase machine learning models were dominated by the SVM (R. Clark et al., 2013), (Keshari et al., 2008), (Shamim et al., 2018). The middle research era cleared the massive research implementation of ANN (Simistira et al., 2014) and neural network-based recognition models (Onan, 2015), (Onan & Toçouglu, 2016), (Onan, 2021). The late phase is ardently dominated by deep learning-based recognition methods (Onan, 2020), considering CNN (Sakhawat et al., 2018), (J. Wang et al., 2020b), and generative adversarial paired network (J. W. Wu et al., 2020). The research trends that are proposed for this purpose are namely "Feature-based Model Development" (T10.9), "Attention and Deep Networks" (T10.7), and the research area "CROHME and Neural Network Model also defends this recognition category of methods.

The present research trends reveal that the hybrid recognition methods have been proficiently deployed for recent MER. The studies (Zhang et al., 2017a),(Le et al., 2019),(Ramirez-Pina et al., 2018),(A.-M. Awal et al., 2010a) have been ardently picked, witnessing the context of the combinational model (hybrid methods) that uses the best of non-

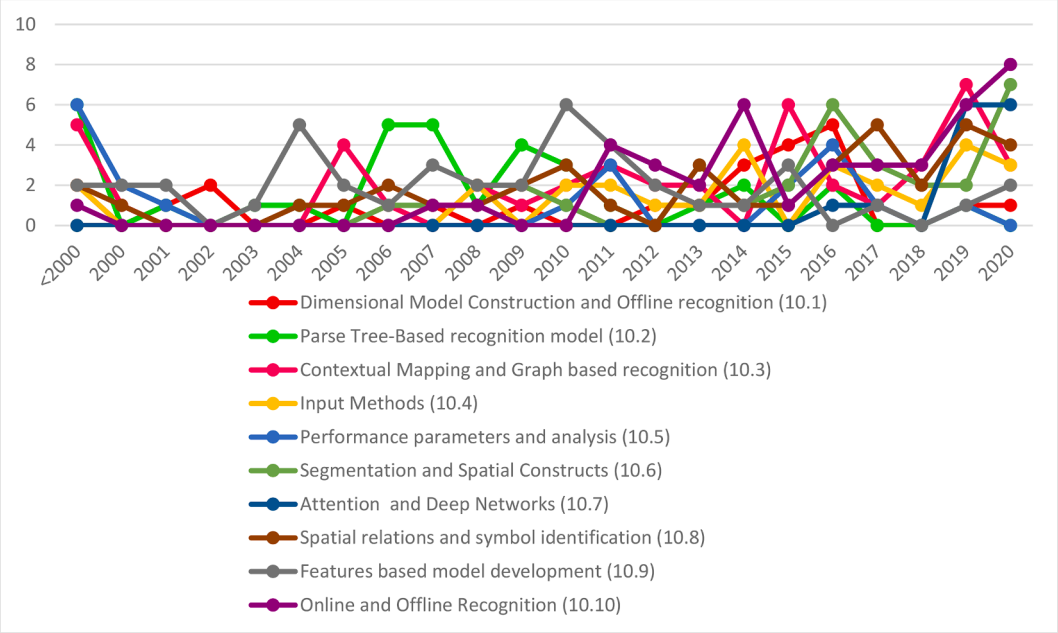

Fig. 10. Dynamics of MER Research Trends.

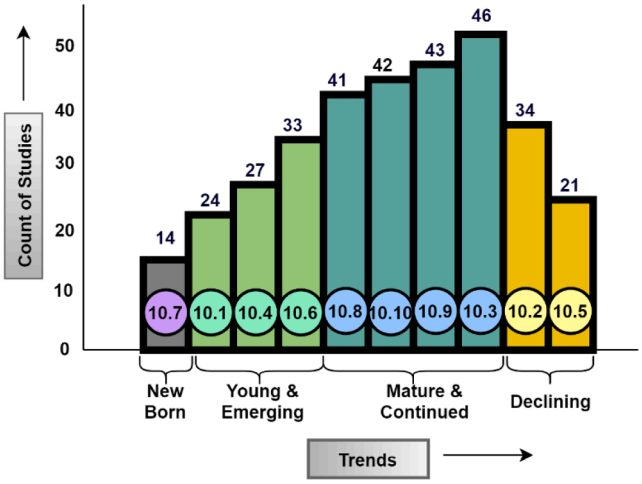

Fig. 11. Development Chart of Research Trends.

machine learning and machine learning algorithm. This hybrid recognition approach has optimized the capabilities of both the previously defined categories of recognition methods. The hybrid recognition approach has not been consistently researched and deployed and is the considered area of concern in terms of research trends.

Highlights (RQ2): The impact of three classified methods has been witnessed corresponding to different time ranges (Fig. 9). First-generation mapped to parsing and grammar-based methodologies, second phase dominated by SVM, ANN, and other neural network-based models. The third phase, or the recent times reflecting the trends of hybrid models, combines the two former phases.

Summary and Challenges (RQ2): The current recognition methods in MER are mainly centered around machine learning approaches. Thus, the classification has been compiled on the same grounds. The three different recognition methods (ML, Non-ML, and Hybrid) identified are impactfully mapped with distinct time zones. The fundamental challenge in the process of identifying these methodologies included the risk of missing out on some techniques that have been exiguously deployed.

6.3. *Research question 3*

Which research areas demand greater attention from researchers?

Research in the field of MER has been carried out for the last four decades. Yet, a completely automated and organized study on recognizing mathematical expressions remains very timely. The research displayed by the LDA model's deployment has extended to the horizons of the current state of the art. The trends that have been investigated show some research area has been widely explored, whereas few research trends have been scarcely researched and covered. The year-wise count of research trends is depicted in Fig. 10 lucidly. It presents a clear picture of what has been the overall journey of all research trends. Some of them have grown continuously with time and now seem to have reached a point of maturity, whereas others seem to have emerged newly. Thus, analyzing the development chart of the research trends that predict the dynamics of research trends in MER (refer to Fig. 10), the authors have categorized the trends under four phases; namely: (a) newborn trends, (b) young and emerging trends, (c) continued and matured trends and (d) declining trends. Fig. 10 shows the dynamics of MER research trends, representing the year-wise development of each trend. Note: The time frame considered for the dynamics of MER (Fig. 10) is up to the year 2020. As the publications for the ongoing year 2021 cannot be finalized currently. The development chart of research trends drawn in Fig. 11 clearly gives an idea about these four mentioned categories of trends. The x-axis in Fig. 11 marks the categories of identified trends, and the y-axis marks the count of studies under each trend category header. The vertical bars in Fig. 11 denote varied trends identified under ten topic solutions. According to the number of studies, each trend is placed under different trend categories. Each of which has been discussed in the later parts of the study.

6.3.1. *Newborn trends*

It has been witnessed that some research trends are in toddling phase. For instance, the research trend "Attention and Deep Networks" (T10.7) has the least count of studies, as depicted in Fig. 11. It certainly holds a decent scope for future research, but the factor for this scarcity in the research is the newness of the concept. Thus, the researchers have termed this trend a **newborn trend**. From the total count of studies collected for the trend, most of the studies only belong to the past five to seven years. Also, the recent collection of studies of the past five years

shows the good stances of studies using attention-based or encoder-decoder-based mechanisms. But the low overall count is suggestive of calling this trend emerging or newborn.

Addressing a newborn trend: (T10.7) Attention and Deep Networks

The sparse articles shadowed under this research trend are also probable as it has recently emerged and experimented with. The practical consequential effect on the accuracy of different hybrid techniques clamors for more contemplation and exploration. The least explored topic is found to be the "Attention and Deep Networks," which have been currently dawned and actualized. This trend depicts the scarcity of hybrid models, and one of the biggest challenges is the formalization of this fact. This study has portrayed the research advancements and trends that demand concentration, and now it has been justified assuredly through this semi-automated analysis.

Potential Points for future work: The need and scarcity of this research trend reveal the potential for further research on the following questions:

- Which hybrid techniques are more potent for pre-processing and segmentation procedures?
- What will be an effective combination of techniques, parsing, or NLP-based criteria in deploying neural network-based models?

6.3.2. Declining trends

One of the reasons for the scarcity of studies extracted under specific trends is the newness of the concept. But this is not true for all the trends with a lesser number of studies. According to the dynamics diagram of research trends, some trends diminish with time. Over the years, these trends, namely "Parse-Tree based Recognition" (T10.2) and "Performance and Parametric analysis" (T10.5), seem to extinct and decline with the reduced count of associated studies. The dynamic diagram depicts the growth and works witnessed on these trends over the years. The declining pace of studies suggests these two trends the declining trends (refer to Fig. 11).

Addressing a declining trend: (T10.2) Performance and Parametric Analysis

The analysis and estimation metrics-related issues need to be extensively addressed and looked for. One of the most occurred challenges here is the fact that different recognition techniques engage different parameters for accuracy evaluation. Still, a justified analysis could be performed on the analysis metrics involved in the recognition process. Different recognition models use varied parameters for the evaluation of the systems. This varied choice of parameters by various studies makes it challenging to evaluate, grade, and compare the other recognition systems.

Potential Points for Future Works: There is decent potential for the probation and exploration for the following setbacks found under this trend:

- What are the accuracy metrics and parameters involved in recognizing based on varied models?
- Can there be standardization of the metric aspect of the studies to launch a universal parameter of accuracy estimation to facilitate the comparative performance analysis?

6.3.3. Young and emerging trends

Examination of the count of studies over the years reveals that some trends have emerged in the past decade, or we can say that these trends have been explored sufficiently well in the last quarter of the chosen time frame. The trends Input Methods (T10.4), Segmentation and Spatial Constructs (T10.6), and Dimensional Model Reduction (T10.1) witnesses a dormant role in the first decade of the dynamics chart of trends, whereas these trends seem to have geared up the frequency in the last decade, particularly in a recent couple of years. These trends have comparatively less count than some majorly active trends like "Contextual Mapping and Graph-based recognition" (T10.3). Yet, these trends hold the potential to rise with time and reaching a level of maturity in coming times. These trends call for more attention from the research community to achieve a stage or point of maturation in the approaching years.

Addressing a young and emerging trend: T(10.4) Input Methods

The observations from the recent publications depict the evolving trend of various input modes for inputting handwritten data into a machine. Initially counted, two significant modes of input were online and offline (Davila et al., 2014), where the online mode of inputting data by drawing on a touchscreen surface, and the offline involves feeding a scanned copy of the image. The reviewers have observed the four input forms: online input (on touch surface), offline-handwritten (scanned image of handwritten text), offline-printed (scanned image of printed text), and voice-based input (spoken form). These input methods are developing and demand more attention. The current scenario depicts blooming studies that work with multi-modalities (Medjkoune et al., 2013), (Jiang et al., 2010), and speech-based input methods (Medjkoune et al., 2012). The dataset named HAMEX (Quiniou et al., 2011) needs to have experimented with enough new input methods introduced to the research community could be a stage of well-exploration and maturation.

Potential Points for Future Works: The less explored dimension of the multimodal and other emerging input methods under this research trend reveals the potential for further research on the following questions.

- Can research on varied input methods be extended to contribute to mathematical learning for the disabled?
- What could be more dimensions where multimodalities and other input methods are explored?

6.3.4. Matured and continued trends

This category holds those trends that have been constantly explored over the years. This pursuit of continuity depicts that the research is going in the right direction. It includes trends like Contextual Mapping and Graph-based recognition" (T10.3), Spatial Relations and Symbol Identification (T10.8), Feature-based model development (T10.9), and Online and Offline Recognition (T10.10) are the topics with a reasonably fair count of studies. The count of studies under these trends is contemplated to be perpetual and continued in the overall dynamics of trends. These trends have decent subsistence and seem to have attained the point of maturity and deserve a considerable amount of attention from the research community. The term "mature" in the label should not be confused with saturation. The matured trends here depict ongoing ongoing trends being explored and attended satisfactorily well. Indeed, these promising and deserving trends clamor for more attention and exploration by the research community. A significant number of studies under these trends have been witnessed under these topic labels (refer to Fig. 11).

Addressing a matured and continued trend: (T10.3) Contextual Mapping and Graph-based recognition

The graph-based approach could prove to be a pretty promising approach for future research. The perpetually researched domain named "Contextual Mapping and Graph-based recognition" witnesses most of the recent articles (T. Zhang et al., 2018), (Lods et al., 2019), (Julca-Aguilar et al., 2020), (J. Wu et al., 2021), where contextual mapping and graph-based recognition have been performed competently. The graph-based methodologies, when experimented with along with machine-learning models, are achieving a constant rise in popularity among MER research groups. The continued rise in popularity characterizes this trend as one of the most promising trends extracted by the LDA model.

Potential Points for Future Work: The fairly explored and continually researched trend signifies the trend "Contextual Mapping and Graph-based recognition "as a promising trend that deserves attention from the MER research community. Further, it reveals a potential scope to work on the following questions in the future.

- What impact do graph-based recognition models make in the regular machine learning recognition process?
- Do the hybrid models developed using a graph or associated tree-based structures outcome the state of art models presented in CROHME?

7. Threats to validity

This analysis is based on LDA topic modeling and has bounded to the limitations of this topic modeling technique. Moreover, some other threats to the validity of this study, along with their corresponding mitigation strategies, have been discussed in this section:

7.1. Search string bias

This threat deals with the effectiveness of the empiric search string practiced for retrieving valuable articles in this domain. A sufficient count has been achieved, yet the risk of missing out is a concern. The bibliographic material has also been inferred; the search string insufficiency has been eradicated appropriately due to the limitations of selected search terms, synonyms, string formulation, and search engines' variedness resulting in imperfect retrieval selection of literature corpus. The relevancy of the selected articles is though cross-checked by applying a two-step review process. Yet, the choice of keywords and limitation of search string may lead to the omission of some functional studies.

7.2. Subjectivity of the topic labeling

The labeling of topics is a significant concern due to the subjectivity and biasing involved in it. To overcome this limitation, both the authors of the paper conducted several rounds of meetings. They did topic labeling individually, and later on, it is combined to generate a conclusive label.

8. Conclusion

This research is based on mathematical foundations and discovers the research trends in MER literature. It uncovers the research trends by analyzing documents published by the researchers. The applied approach has proven its significance by depicting the research areas and the corresponding research trends very proximate to the extracted topic solution. The trend analysis, being semi-automated and algorithm-dependent, finely predict the potential topic solutions or trends that need to be exclusively addressed by the research community. Moreover, researchers can analyze any of the research trends out of ten research trends based on the status in the development chart of trends (Fig. 11).

In this study, the authors have attempted to introduce trend identification in the field of MER discipline by processing and investigating a sufficient good portion of the corpus (formed of literature publications) and empirically recognizing key research zones, areas, and trends. The study sublimates and interprets the intellectual core of MER research decently by postulating five research areas as "Parsing Techniques", "Character-Based Recognition methods", "Segmentation and Classification Procedures", and "CROHME and Neural Network Model", and "Structural Analysis Mechanism".The extracted research areas depict in detail the recognition methods. These recognition models are placed according to their popularity in different time frames. This report of existing categories of techniques is vividly depicted through the generation diagram (Fig. 9) in the field of MER, which defines the baseline of deployed recognition models.

From the high diversity in topics being examined on MER, the authors have posed the ten major research trends, and that's the principle contribution of the study. The results of the analysis provide an empirical basis for future debates about identity and diversity in the MER field. The most significant extractions of the survey are postulated as follows:

- One of the most dominating publication channels is the "*International Conference on Document Analysis and Recognition*" (#44). Being the organizer of the CROHME series, it has wholly bridged the research rift and boosted the potential research among the research community.
- The prominent years which witnessed most of the publications on MER are 2019 (#34) and 2020 (#34).
- The analysis also predicts the significant research area is "Segmentation and Classification Procedures," holding 28.31 % of selected studies, as displayed in Fig. 8.
- In the year wise-count analysis, when scaled to topic ratios, every topic apparently exhibited a rising scale of ratio values after the year 2010, which is suggestive that the domain of mathematical expression and its associated recognition tasks have been an explored direction in the evolving era of the past decade.
- The categorization of techniques reveals three phases—the first phase as non-ML, the second phase as ML, and the third phase as a hybrid. The phase of hybrid techniques
- seems to have evolved recently in the last quarter of the past decade.
- The trend with the highest number of publications (#46) is "Contextual Mapping and Graph-based Recognition" (T10.3).
- "Attention and Deep Networks" is identified as the newborn trend, and "Contextual and Graph-based recognition" emerged as the matured research trend.
- The extracted research patterns reveal that certain trends demand varied attention. The analysis shows that "young and emerging trends" and "matured and continued trends" are fairly explored yet deserve continued attention. Also, the recent focus of the research community seems to target the "newborn trends," which has a long way to reach a point of maturity.

There have been several observations regarding the research trends, and the following are the points that need to be focused on for future research:

- Advancement in the deployment of hybrid concepts of recognition.
- Hybrid techniques using machine learning, deep learning, and the NLP model demand substantial attention.
- Implementation of the latest technological recognition trends should be openly accepted, explored, and compared for refinements and innovativeness.
- The advent of the trend of varied input methods clamors for more exploration of multimodal recognition systems.
- Graph-based recognition comes up as a matured and continued trend that deserves continuous attention.
- Concentration on standardization of datasets and accuracy metrics needs to be exclusively addressed.
- Automated and systematic surveys remain to be covered as scarcity has been discovered.

Overall, the application of LDA has helped the researchers enable them to face new challenges and meet the alignment of their work with the contemporary research trends. They may also use other topic modeling techniques to identify the hidden patterns and trends from a sizeable bibliographic dataset. Also, considering abstracts instead of full texts is a limitation as it indeed introduces a loss of information. On the contrary, using full text for processing would have led to considerable noise, specifically for the methodology sections. Nevertheless, this could be a limitation and potential concern for the study. Some limitations have been identified during research on MER trends, yet there is enough scope for future research in varied specified dimensions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The current study has not received any funding for performing the research work.

References

Aguilar, F. D. J., & Hirata, N. S. T. (2012). ExpressMatch: A system for creating ground-truthed datasets of online mathematical expressions. *IAPR International Workshop on Document Analysis Systems, 155–159*. https://doi.org/10.1109/DAS.2012.38

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, & J. S.. (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(5), 14. https://doi.org/10.1117/12.2216384

Álvaro, F., & Sánchez, J. A. (2010). Comparing several techniques for offline recognition of printed mathematical symbols. *International Conference on Pattern Recognition, 1953–1956*. https://doi.org/10.1109/ICPR.2010.481

Álvaro, F., Sánchez, J. A., & Benedí, J. M. (2014a). Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models. *Pattern Recognition Letters, 35*(1), 58–67. https://doi.org/10.1016/j.patrec.2012.09.023

Álvaro, F., Sánchez, J. A., & Benedí, J. M. (2013). An image-based measure for evaluation of mathematical expression recognition. *Iberian Conference on Pattern Recognition and Image Analysis, 682–690*. https://doi.org/10.1007/978-3-642-38628-2_81

F. Álvaro J.A. Sánchez J.M. Benedí Offline features for classifying handwritten math symbols with recurrent neural networks 22nd International Conference on Pattern Recognition 2014 Stockholm, Sweden 2944 2949 10.1109/ICPR.2014.507.

R.H. Anderson Two-dimensional mathematical notation Syntactic Pattern Recognition, Applications 1977 Springer 147 177.

Anderson, R. H. (1967). Syntax-Directed Recognition of Hand-Printed Two-Dimensional Mathematics. *Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium.*, 436–459. https://doi.org/10.1145/2402536.2402585.

P. Anupriya S. Karpagavalli LDA based topic modeling of journal abstracts 2015 International Conference on Advanced Computing and Communication Systems 2015 Coimbatore, India 1 5 10.1109/ICACCS.2015.7324058.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murty, M. N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 391–402, Hyderabad, India. https://doi.org/10.1007/978-3-642-13657-3_43.

Awal, A.-M., Mouchère, H., & Viard-Gaudin, C. (2010a). A hybrid classifier for handwritten mathematical expression recognition. *Document Recognition and Retrieval XVII, 7534*, Article 753410. https://doi.org/10.1117/12.840023

A.-M. Awal H. Mouchère C. Viard-Gaudin Improving online handwritten mathematical expressions recognition with contextual modeling Twelveth International Conference on Frontiers in Handwriting Recognition 2010 Kolkata, India 427 432 10.1109/ICFHR.2010.73.

A.-M. Awal H. Mouchère C. Viard-Gaudin The problem of handwritten mathematical expression recognition evaluation 12th International Conference on Frontiers in Handwriting Recognition 2010 Kolkata, India 646 651 10.1109/ICFHR.2010.106.

Awal, A. M., Mouchère, H., & Viard-Gaudin, C. (2014). A global learning approach for an online handwritten mathematical expression recognition system. *Pattern Recognition Letters, 35*(1), 68–77. https://doi.org/10.1016/j.patrec.2012.10.024

A.M. Awal H. Mouchère C. Viard-Gaudin Towards handwritten mathematical expression recognition 2009, 10th International Conference on Document Analysis and Recognition, 1046–1050 2009 Barcelona, Spain 10.1109/ICDAR.2009.71.

Banitaan, S., & Alenezi, M. (2015). Software evolution via topic modeling: An analytic study. *International Journal of Software Engineering and Its Applications, 9*(5), 43–52. https://doi.org/10.14257/ijseia.2015.9.5.05.

Basu, S., Chaudhuri, C., Kundu, M., Nasipuri, M., & Basu, D. (2002). Segmentation of Offline Handwritten Bengali Script. *Proceedings- 28th IEEE ACE*, 171–174. https://doi.org/10.48550/arXiv.1202.3046.

Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I., … Tovey, D. (2013). PRISMA for abstracts: Reporting systematic reviews in journal and conference abstracts. *PLoS Medicine, 10*(4), e1001419.

Bird, C., Menzies, T., & Zimmermann, T. (2015). The art and science of analyzing software data. *Elsevier*. https://doi.org/10.5555/2886235

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bradford, R. B. (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 153–162, California, USA. https://doi.org/10.1145/1458082.1458105.

Breiner, T., Nguyen, C., Esch, D. Van, & Brien, J. O. (2017). *Automatic Keyboard Layout Design for Low-Resource Latin-Script Languages*.

Büyükbayrak, H., Yanikoglu, B., & Erçil, A. (2007). Online handwritten mathematical expression recognition. *Document Recognition and Retrieval XIV, 6500*, 65000F. https://doi.org/10.1117/12.704043

K. Canini L. Shi T. Griffiths Online inference of topics with latent Dirichlet allocation Artificial Intelligence and Statistics 2009 65 72 https://doi.org/10.1.1.187.6726.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Celik, M., & Yanikoglu, B. (2011a). Handwritten Mathematical Formula Recognition using a Statistical Approach. *IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, 498–501. https://doi.org/10.1109/SIU.2011.5929696.

M. Celik B. Yanikoglu Probabilistic mathematical formula recognition using a 2D context-free graph grammar International Conference on Document Analysis and Recognition 2011 Beijing, China 161 166 10.1109/ICDAR.2011.41.

Chan, K. F., & Yeung, D. Y. (2000). Mathematical expression recognition: A survey. *International Journal on Document Analysis and Recognition(IJDAR), 3*(1), 3–15. https://doi.org/10.1007/PL00013549

Chang, S.-K. (1970). A method for the structural analysis of two-dimensional mathematical expressions. *Information Sciences, 2*(3), 253–272. https://doi.org/10.1016/S0020-0255(70)80052-4

Chen, B., Zhu, L., Kifer, D., & Lee, D. (2010). What is an opinion about? exploring political standpoints using opinion scoring model. *Proceedings of the AAAI Conference on Artificial Intelligence, 24*(1), 1007-1012,Georgia, USA. https://doi.org/10.5555/2898607.2898768.

Chen, G., & Tang, Y. (2013). Baseline based multi-candidate mathematical expression recognition. *Jisuanji Gongcheng Yu Yingyong(Computer Engineering and Applications), 49*(1).

Chen, L. (1992). A system for on-line recognition of handwritten mathematical expressions. *Computer Processing of Chinese and Oriental Languages, 6*(1), 19–39. https://doi.org/10.1109/ICFHR.2012.172

Y. Chen T. Shimizu K. Yamauchi M. Okada Ambiguous problem investigation in off-line mathematical expression understanding Smc 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions 4 2000 Nashville, TN, USA 2917 2922 10.1109/ICSMC.2000.884443.

Chiyangwa, T. B., van Biljon, J., & Renaud, K. (2021). Natural language processing techniques to reveal human-computer interaction for development research topics. *Proceedings of the International Conference on Artificial Intelligence and Its Applications*, 1–7. https://doi.org/10.1145/3487923.3487932.

Clark, B., & Zubrow, D. (2001). How good is the software: a review of defect prediction techniques. *Software Engineering Symposium, Carreige Mellon University*.

Clark, R., Kung, Q., & Van Wyk, A. (2013). System for the recognition of online handwritten mathematical expressions. *EuroCon, 2013*, 2029–2035. https://doi.org/10.1109/EUROCON.2013.6625259

M. Cristani A. Perina U. Castellani V. Murino Geo-located image analysis using latent representations 2008 IEEE Conference on Computer Vision and Pattern Recognition 2008 Anchorage, AK, USA 1 8 10.1109/CVPR.2008.4587390.

K. Davila S. Ludi R. Zanibbi Using Off-Line Features and Synthetic Data for On-Line Handwritten Math Symbol Recognition Fourteenth International Conference on Frontiers in Handwriting Recognition 2014 Hersonissos, Greece 323 328 10.1109/ICFHR.2014.61.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

L. Dong H. Liu Recognition of offline handwritten mathematical symbols using convolutional neural networks International Conference on Image and Graphics 2017 Shanghai, China 149 161 10.1007/978-3-319-71607-7_14.

Ethen. (2015). *Topic Modeling*. http://ethen8181.github.io/machine-learning/clustering/topic_model/LDA.html%0A.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems, 21*(1), 70–86. https://doi.org/10.1057/ejis.2010.61

Firdaus, S. A., & Vaidehi, K. (2020). Handwritten Mathematical Symbol Recognition Using Machine Learning Techniques: Review. *Advances in Decision Sciences, Image Processing, Security and Computer Vision, 658–671*. https://doi.org/10.1007/978-3-030-24318-0_75

Fitzgerald, J. A., Geiselbrechtinger, F., & Kechadi, T. (2007). Mathpad: A fuzzy logic-based recognition system for handwritten mathematics. *Ninth International Conference on Document Analysis and Recognition, 2*, 694–698, Curitiba, Brazil. https://doi.org/10.1109/ICDAR.2007.4377004.

Fitzgerald, J., Geiselbrechtinger, F., & Kechadi, M. (2006). Structural analysis of handwritten mathematical expressions through fuzzy parsing. *ACST, 6*, 151–156.

Fu, Y., Yan, M., Zhang, X., Xu, L., Yang, D., & Kymer, J. D. (2015). Automated classification of software change messages by semi-supervised latent dirichlet allocation. *Information and Software Technology, 57*, 369–377. https://doi.org/10.1016/j.infsof.2014.05.017

U. Garain B.B. Chaudhuri On machine understanding of online handwritten mathematical expressions Seventh International Conference on Document Analysis and Recognition 2003 Edinburgh, UK 349 353 10.1109/ICDAR.2003.1227687.

R. Genoe J.A. Fitzgerald T. Kechadi An online fuzzy approach to the structural analysis of handwritten mathematical expressions IEEE International Conference on Fuzzy Systems 2006 Vancouver, BC, Canada 244 250 10.1109/FUZZY.2006.1681721.

Genoe, R. (2010). *Real-time Structural Analysis of Online Handwritten Mathematical Expressions*. University College Dubin.

Gharde, S. S. (2012). Evaluation of Classification and Feature Extraction Techniques for Simple Mathematical Equations. *International Journal of Applied Information Systems, 1*(5), 34–38.
Gil, W.-J., Kim, J.-W., Park, K.-R., & Cho, H.-J. (2021). An Analysis of Research Trends in AI Education based on LDA. *Review of International Geographical Education Online, 11* (2), 254–262.
Golubitsky, O., & Watt, S. M. (2010). Distance-based classification of handwritten symbols. *International Journal on Document Analysis and Recognition, 13*(2), 133–146. https://doi.org/10.1007/s10032-009-0107-7
Gong, Y., Li, S., Wang, X., & Wang, X. (2015). Real-time recognition method of understanding on-line handwritten mathematical expression. *Computer Engineering and Applications Journal, 7*, 43.
D. Greene J.P. Cross Unveiling the political agenda of the european parliament plenary: A topical analysis Proceedings of the ACM Web Science Conference 2015 1 10 10.1145/2786451.2786464.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235. https://doi.org/10.1073/pnas.0307752101
Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science, 24*. https://doi.org/10.1146/annurev-polisci-053119-015921
Guo, Z., & Liu, Y. (2018). Research on Mathematical Formula Knowledge Base for Formula Recognition. *IEEE/WIC/ACM International Conference on Web Intelligence, 2018*, 619–622. https://doi.org/10.1109/WI.2018.00-27
N.D. Hai A.L. Duc M. Nakagawa Combination of LSTM and CNN for Recognizing Handwritten Online Mathematical Symbols The 17th Information-Based Induction Sciences Workshop 2014 10.1146/annurev-polisci-053119-015921.
Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). *Natural language processing: Python and NLTK*. Packt Publishing Ltd.
Hew, J.-J., Lee, V.-H., Ooi, K.-B., & Lin, B. (2019). Computer science in ASEAN: A ten-year bibliometric analysis (2009–2018). *Journal of Computer Information Systems*. https://doi.org/10.1080/08874417.2019.1601538
Hidayatullah, A. F., Aditya, S. K., Karimah, & Gardini, S. T. (2019). Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). *IOP Conference Series: Materials Science and Engineering*, 1–8, Manila City, Philippines. https://doi.org/10.1088/1757-899X/482/1/012033.
Hordri, N. F., Samar, A., Yuhaniz, S. S., & Shamsuddin, S. M. (2017). A systematic literature review on features of deep learning in big data analytics. *Proceedings of International Journal of Advances in Soft Computing and Its Applications, 9*(1), 32–49.
Hu, J., Brown, M. K., & Turin, W. (1996). HMM based online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(10), 1039–1045.
L. Hu R. Zanibbi HMM-based recognition of online handwritten mathematical symbols using segmental K-means initialization and a modified pen-up/down feature 2011 457–462 Beijing, China 10.1109/ICDAR.2011.98.
L. Hu R. Zanibbi Segmenting handwritten math symbols using adaboost and multi-scale shape context features 2013 Washington, DC, USA 1180 1184 10.1109/ICDAR.2013.239.
Y. Hu L. Peng Y. Tang On-line handwritten mathematical expression recognition method based on statistical and semantic analysis 11th IAPR International Workshop on Document Analysis Systems 2014 171 175 10.1109/DAS.2014.47.
J. Huang J. Tan N. Bi Overview of Mathematical Expression Recognition 2020 Zhongshan city China 10.1007/978-3-030-59830-3_4.
Huang, M., Wen, S., Jiang, M., & Yao, Y. (2021). LDA Topic Mining of Light Food Customer Reviews on the Meituan Platform. *International Conference on Data Mining and Big Data*, 108–121.
Y. Jiang F. Tian H. Wang X. Zhang X. Wang G. Dai Intelligent Understanding of Handwritten Geometry Theorem Proving 2010 119–128 Hong Kong, China 10.1145/1719970.1719988.
Julca-Aguilar, F., Mouchère, H., Viard-Gaudin, C., & Hirata, N. S. T. (2020). A general framework for the recognition of online handwritten graphics. *International Journal on Document Analysis and Recognition, 23*, 143–160. https://doi.org/10.1007/s10032-019-00349-6
F. Julca-Aguilar H. Mouchère C. Viard-Gaudin H. Mouchere V.-G. Christian N.S.T. Hirata … C. Viard-Gaudin Top-Down Online Handwritten Mathematical Expression Parsing with Graph Grammar IberoAmerican Congress on Pattern Recognition 2 2015 444 451 https://doi.org/https://doi.org/10.1007/978-3-319-25751-8_53.
Kanahori, T., Tabata, K., Cong, W., Tamari, F., & Suzuki, M. (2000). On-line recognition of mathematical expressions using automatic rewriting method. *International Conference on Multimodal Interfaces, 394–401*. https://doi.org/10.1007/3-540-40063-x_52
Kaplan, V. (2016). A New Algorithm to Parse a Mathematical Expression and its Application to Create a Customizable Programming Language. *ICSEA, 2016*, 285.
Keshari, B., Watt, S. M. S., Keshari, Birendra and Watt, S. M., Keshari, B., & Watt, S. M. S. (2008). Online mathematical symbol recognition using svms with features from functional approximation. *Electronic Proceedings of Mathematical User-Interfaces Workshop*, 1–5. http://www.cecm.sfu.ca/~pborwein/MITACS/papers/OnlineMathSymb.pdf.
Khuong, V. T. M. V., Phan, M., Tran, V., Khuong, M., Phan, K. M., Nakagawa, M., Khuong, V. T. M. V., Phan, M., Phan, K. M., & Nakagawa, M. (2019). Interactive User Interface for Recognizing Online Handwritten Mathematical Expressions and Correcting Misrecognition. *Proceedings of International Conference on Document Analysis and Recognition Workshops (ICDARW), IEEE, 2*, 26–30. https://doi.org/10.1109/ICDARW.2019.10034.
Kim, K., Rhee, T. H., Lee, J. S., & Kim, J. H. (2009). Utilizing consistency context for handwritten mathematical expression recognition. *International Conference on Document Analysis and Recognition*, 1051–1055, Barcelona, Spain. https://doi.org/10.1109/ICDAR.2009.140.
Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences, 9*(1), 1–21. https://doi.org/10.1186/s13673-019-0192-7
Kitchenham, B., & S. Charters. (2007). *Guidelines for performing systematic literature reviews in software engineering*.
Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., Linkman, S., … Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology, 51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009
Kukreja, & Sakshi., V. (2022). Machine learning models for mathematical symbol recognition: A stem to stern literature analysis. *Multimedia Tools and Applications, 1–37*. https://doi.org/10.1007/s11042-022-12644-2
Le, A. D., Indurkhya, B., & Nakagawa, M. (2019). Pattern generation strategies for improving recognition of Handwritten Mathematical Expressions. *Pattern Recognition Letters, 128*, 255–262. https://doi.org/10.1016/j.patrec.2019.09.002
Le, A. D., & Nakagawa, M. (2016). A system for recognizing online handwritten mathematical expressions by using improved structural analysis. *International Journal on Document Analysis and Recognition, 19*(4), 305–319. https://doi.org/10.1007/s10032-016-0272-4
Le, A. D., & Nakagawa, M. (2013). A ground-truthing tool for making a database of online handwritten mathematical expressions. *Pattern Recognition and Media Understanding*, *112*(495), 147–150, Tokyo, Japan.
A.D. Le M. Nakagawa Training an end-to-end system for handwritten mathematical expression recognition by generated patterns 1 2017 Kyoto, Japan 1056 1061 10.1109/ICDAR.2017.175.
A.D. Le H.D. Nguyen M. Nakagawa Modified X-Y Cut for Re-Ordering Strokes of Online Handwritten Mathematical Expressions 12th IAPR International Workshop on Document Analysis Systems 2016 233 238 10.1109/DAS.2016.19.
A.D. Le T.V. Phan M. Nakagawa A system for recognizing online handwritten mathematical expressions and improvement of structure analysis 11th IAPR International Workshop on Document Analysis Systems 2014 51 55 10.1109/DAS.2014.52.
J. Lee B.W. Yogatama H. Christian Optical Character Recognition for Handwritten Mathematical Expressions in Educational Humanoid Robots 2018 Bandung, Indonesia 10.1109/ICSEngT.2018.8606374.
Leong, L.-Y., Hew, T.-S., Ooi, K.-B., & Lin, B. (2021). A meta-analysis of consumer innovation resistance: Is there a cultural invariance? *Industrial Management & Data Systems.*. https://doi.org/10.1108/IMDS-12-2020-0741
Z. Li X. Tian An improved analysis approach of overbrace/underbrace structure in printed mathematical expressions 2010 International Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering 2010 10.1109/CICC-ITOE.2010.22.
E. Linstead P. Rigor S. Bajracharya C. Lopes P. Baldi Mining concepts from code with probabilistic topic models 2007 461–464 Georgia, USA 10.1145/1321631.1321709.
Littin, R. H. (1995). *Mathematical expression recognition: Parsing pen/tablet input in real-time using LR techniques*. University of Waikato.
Lods, A., Anquetil, E., & Mace, S. (2019). Fuzzy visibility graph for structural analysis of online handwritten mathematical expressions. *International Conference on Document Analysis and Recognition*, 641–646, Sydney, NSW, Australia. https://doi.org/10.1109/ICDAR.2019.00108.
Lukins, S. K., Kraft, N. A., & Etzkorn, L. H. (2008). Source code retrieval for bug localization using latent dirichlet allocation. *15th Working Conference on Reverse Engineering*, 155–164, Antwerp, Belgium. https://doi.org/10.1109/WCRE.2008.33.
Mahdavi, M., Zanibbi, R., Mouchere, H., Viard-Gaudin, C., & Garain, U. (2019). ICDAR 2019 CROHME + TFD: Competition on recognition of handwritten mathematical expressions and typeset formula detection. *International Conference on Document Analysis and Recognition*, 1533–1538, Sydney, NSW,Australia. https://doi.org/10.1109/ICDAR.2019.00247.
Mahmoud, K., BingRu, Y., et al. (2011). A Hybrid Segmentation System of Offline Arabic Mathematical Expression Recognition. *International Symposium on Information Engineering and Electronic Commerce, 1–4*. https://doi.org/10.1115/1.859759
Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., , … Häussler, T., et al. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12* (2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754
Mavridis, T., & Symeonidis, A. L. (2014). Semantic analysis of web documents for the generation of optimal content. *Engineering Applications of Artificial Intelligence, 35*, 114–130.
S. Medjkoune H. Mouchère H. Mouchere S. Petitrenaud C. Viard-gaudin Using Speech for Handwritten Mathematical Expression Recognition Disambiguation 2012 187–192 Bari, Italy 10.1016/j.engappai.2014.06.008.
Medjkoune, S., Mouchère, H., & Petitrenaud, S. (2013). Multimodal mathematical expressions recognition: Case of speech and handwriting. *International Conference on Human-Computer Interaction. Springer, Berlin, Heidelberg.*, 77–86, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39330-3_9.
Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews, 4*(1), 1–12. https://doi.org/10.1186/s13643-015-0117-0
Mouchère, H. (2011). *CROHME*. https://www/isical.ac.in/~crohme/.
Mouchère, H., Viard-Gaudin, C., Kim, D. H., Kim, J. H., & Garain, U. (2011). CROHME2011: Competition on recognition of online handwritten mathematical expressions. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1497–1500, Beijing, China. https://doi.org/10.1109/ICDAR.2011.297.
Mouchère, H., Zanibbi, R., Garain, U., & Viard-Gaudin, C. (2014). Advancing the State-of-the-Art for Handwritten Math Recognition: The CROHME competitions,

2011–2014. *International Journal on Document Analysis and Recognition, 19*(2), 173–189. https://doi.org/10.1007/s10032-016-0263-5
E. Naderan Online Handwritten Mathematical Expressions Recognition System Using Fuzzy Neural Network ArXiv Preprint 2017 ArXiv:1707.03088.
Naderan, E., & Zaychenko, Y. P. (2013). An Approach to Structural Analysis of Handwritten Mathematical Expressions in Real Time. *Visnyk NTUU" KPI": Informatics, Operation and Computer Science*, *2013*(58).
Nguyen, C. T., Khuong, V. T. M., Nguyen, H. T., & Nakagawa, M. (2020). CNN based spatial classification features for clustering offline handwritten mathematical expressions. *Pattern Recognition Letters, 131*, 113–120. https://doi.org/10.1016/j.patrec.2019.12.015
Nguyen, V., Cai, J., & Chu, J. (2019). Hybrid CNN-GRU model for high efficient handwritten digit recognition. *ACM International Conference Proceeding Series*, *2*, 66–71, Beijing, China. https://doi.org/10.1145/3357254.3357276.
Okamoto, M., & Higashi, H. (1995). Mathematical expression recognition by the layout of symbols. *Trans. IEIEC*, 474–482.
Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications, 42*(20), 6844–6852. https://doi.org/10.1016/j.eswa.2015.05.006
Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science, 42*(2), 150–165. https://doi.org/10.1177/0165551515591724
Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*. https://doi.org/10.1108/K-10-2016-0300
Onan, A. (2018a). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science, 44*(1), 28–47. https://doi.org/10.1177/0165551516677911
Onan, A. (2018b). Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine, 2018*. https://doi.org/10.1155/2018/2497471
Onan, A. (2019a). Consensus clustering-based undersampling approach to imbalanced learning. *Scientific Programming, 2019*. https://doi.org/10.1155/2019/5901087
Onan, A. (2019b). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access, 7*, 145614–145633. https://doi.org/10.1109/ACCESS.2019.2945911
Onan, A. (2020). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education, 28*(1), 117–138. https://doi.org/10.1002/cae.22179
Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience, 33*(23), e5909.
Onan, A. (2019c). Topic-enriched word embeddings for sarcasm identification. *Computer Science On-Line Conference*, 293–304, Czech Republic, Hlavni Mesto Praha. https://doi.org/10.1007/978-3-030-19807-7_29.
Onan, A., Bal, V., & Yanar Bayam, B. (2016a). The use of data mining for strategic management: A case study on mining association rules in student information system. *Croatian Journal of Education: Hrvatski Časopis Za Odgoj i Obrazovanje, 18*(1), 41–70. https://doi.org/10.15516/cje.v18i1.1471.
Onan, A., & Korukouglu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science, 43*(1), 25–38. https://doi.org/10.1177/0165551515613226
Onan, A., Korukouglu, S., & Bulut, H. (2016b). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications, 57*, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045
Onan, A., & Toçouglu, M. A. (2016). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access, 9*, 7701–7722. https://doi.org/10.1109/ACCESS.2021.3049734
Onari, M. A., Yousefi, S., Rabieepour, M., Alizadeh, A., & Rezaee, M. J. (2021). A medical decision support system for predicting the severity level of COVID-19. *Complex & Intelligent Systems*, 1–15. https://doi.org/10.1007/s40747-021-00312-1
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*.
Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology, 64*, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007
Phan, K. M., Le, A. D., Indurkhya, B., & Nakagawa, M. (2018). Augmented incremental recognition of online handwritten mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR), 21*(4), 253–268. https://doi.org/10.1007/s10032-018-0306-1
K.M. Phan A.D. Le M. Nakagawa Semi-incremental recognition of online handwritten mathematical expressions 15th International Conference on Frontiers in Handwriting Recognition 2016 Shenzhen, China 258 264 10.1109/ICFHR.2016.0057.
K.M. Phan C.T. Nguyen A.D. Le M. Nakagawa An incremental recognition method for online handwritten mathematical expressions 3rd IAPR Asian Conference on Pattern Recognition 2015 Kuala Lumpur, Malaysia 171 175 10.1109/ACPR.2015.7486488.
B.H. Phong L.T. Dat N.T. Yen T.M. Hoang T.-L. Le A deep learning based system for mathematical expression detection and recognition in document images 12th International Conference on Knowledge and Systems Engineering 2020 Can Tho, Vietnam 85 90 10.1109/KSE50997.2020.9287693.
Phong, B. H., Hoang, T. M., & Le, T.-L. (2017). A new method for displayed mathematical expression detection based on FFT and SVM. *Proceedings of 4th NAFOSTED Conference on Information and Computer Science*, 90–95, Hanoi, Vietnam. https://doi.org/10.1109/NAFOSTED.2017.8108044.
Pillay, A. (2014). *Intelligent Combination of Structural Analysis Algorithms: Application to Mathematical Expression Recognition*. Rochester Institute of Technology.
Plisson, J., Lavrac, N., Mladenic, D., et al. (2004). A rule based approach to word lemmatization. *Proceedings of IS, 3*, 83–86.
Ponweiser, M., Grün, B., & Hornik, K. (2014). Finding scientific topics revisited. In *Advances in latent variables* (pp. 93–100). Springer. https://doi.org/10.1007/10104_2014_11.
Porter, M. F. (2001). *Snowball: A language for stemming algorithms*.
Quiniou, S., Mouchère, H., Saldarriaga, S. P., Viard-gaudin, C., Morin, E., Petitrenaud, S., & Medjkoune, S. (2011). HAMEX – A handwritten and audio dataset of mathematical expressions. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 452–456, Beijing, China. https://doi.org/10.1109/ICDAR.2011.97.
Ramirez-Pina, C., Sanchez, J. S., Valdovinos-Rosas, R. M., & J. a. h.-s.. (2018). A Hybrid Feature Extraction Method for Offline Handwritten Math Symbol Recognition. *Iberoamerican Congress on Pattern Recognition, 1*, 893–901. https://doi.org/10.1007/978-3-030-13469-3
Rani, R., & Lobiyal, D. K. (2021). An extractive text summarization approach using tagged-LDA based topic modeling. *Multimedia Tools and Applications, 80*(3), 3275–3305. https://doi.org/10.1007/s11042-020-09549-3
Rani, S. J., & Kumari, V. V. (2016). An effective mechanism of feature based retrieval of mathematical expression from documents. *International Journal of Applied Engineering Research, 11*(5), 3462–3468.
Rhee, T. H., & Kim, J. H. (2009). Efficient search strategy in structural analysis for handwritten mathematical expression recognition. *Pattern Recognition, 42*(12), 3192–3201. https://doi.org/10.1016/j.patcog.2008.10.036
Rhee, T. H., Kim, J. H., & Kim, J. H. (2008). Robust recognition of handwritten mathematical expressions using search-based structure analysis. *Proceedings of International Conference on Frontier in Handwriting Recognition (ICFHR)*, 19–24.
Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408, Shanghai, China. https://doi.org/10.1145/2684822.2685324.
Said, H. E. S., Tan, T. N., & Baker, K. D. (2000). Personal identification based on handwriting. *Pattern Recognition, 33*(1), 149–160. https://doi.org/10.1016/S0031-3203(99)00006-0
Z. Sakhawat S. Ali L. Hongzhi Handwritten digits recognition based on deep learning4J ACM International Conference Proceeding Series 2018 Espoo, Finland 21 25 10.1145/3268866.3268888.
Sakshi, & Kukreja, V. (2021). A retrospective study on handwritten mathematical symbols and expressions : Classification and recognition. *Engineering Applications of Artificial Intelligence, 103*, Article 104292. https://doi.org/10.1016/j.engappai.2021.104292
Savchenkov, P. ;, Savinov, E. ;, Mikhail, T. ;, Kiyan, S. ;, & Esin, A. (2018). Neural Network Based Recognition of Mathematical Expressions (Patent No. 15/187 , 723). In *United States Patent* (15/187 , 723). Google Patents.
Savin, I., Drews, S., & van den Bergh, J. (2021). Free associations of citizens and scientists with economic and green growth: A computational-linguistics analysis. *Ecological Economics, 180*, Article 106878. https://doi.org/10.1016/j.ecolecon.2020.106878
Sehra, S. K., Brar, Y. S., Kaur, N., & Sehra, S. S. (2017). Research patterns and trends in software effort estimation. *Information and Software Technology, 91*, 1–21. https://doi.org/10.1016/j.infsof.2017.06.002
A. Sen H. Shah Automated handwriting analysis system using principles of graphology and image processing 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) 2017 Coimbatore, India 1 6 10.1109/ICIIECS.2017.8276061.
Shamim, S. M., Miah, M. B. A., Angona Sarker, M. R., Al Jobair, A., Sarker, A., Rana, M., & Jobair, A. A. (2018). Handwritten digit recognition using machine learning algorithms. *Indonesian Journal of Science and Technology, 3*(1), 29–39. https://doi.org/10.17509/ijost.v3i1.10795.
Shan, G., Wang, H., Liang, W., & Chen, K. (2021). Robust Encoder-Decoder Learning Framework towards Offline Handwritten Mathematical Expression Recognition Based on Multi-Scale Deep Neural Network. *Science China Information Sciences, 64*(3), 1–12. https://doi.org/10.1007/s11432-018-9824-9
Shi, Y., Soong, F., & Zhou, J. (2011). Symbol graph generation in handwritten mathematical expression recognition. In *U.S. Patent No. 7,885,456*. https://doi.org/10.1109/ICPR.2008.4761542.
S. Shinde R.B. Waghulade D.S. Bormane A new neural network based algorithm for identifying handwritten mathematical equations International Conference on Trends in Electronics and Informatics 2018 10.1109/ICOEI.2017.8300916.
F. Simistira V. Papavassiliou V. Katsouros G. Carayannis Recognition of Spatial Relations in Mathematical Formulas 14th International Conference on Frontiers in Handwriting Recognition 2014 10.1109/ICFHR.2014.35.
Smirnova, E., & Watt, S. M. (2010). Survey on Methods for Mathematical Expression Analysis in Arabic Handwriting. *Citeseer*.
J. Špeh A. Muhič J. Rupnik Parameter Estimation for the Latent Dirichlet Allocation Proceedings of the Conference on Data Mining and Data Warehouses 2013.
Srihari, S. (2015). Bayesian Parameter Estimation in Bayesian Networks. *International Conference on Computer Information Systems and Industrial Applications (CISIA 2015) Bayesian*, 1–21.
S. Syed M. Spruit Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation 2017 Tokyo, Japan 165 174 10.1109/DSAA.2017.61.
Tamburri, D. A., Palomba, F., & Kazman, R. (2020). Success and Failure in Software Engineering: A Followup Systematic Literature Review. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2020.2976642

Tang, H., Shen, L., Qi, Y., Chen, Y., Shu, Y., Li, J., & Clausi, D. A. (2012). A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images. *IEEE Transactions on Geoscience and Remote Sensing, 51*(3), 1680–1692. https://doi.org/10.1109/TGRS.2012.2205579

Tapia, E, & Rojas, R. (2007). A Survey on Recognition of on Line Handwritten Mathematical Notation. In *Technical Report B-07-01 Freie Universitat Berlin, Institut fur Informatik Takustr. 9, 14195 Berlin, Germany*. https://doi.org/10.17169/refubium-23077.

Tapia, Ernesto, & Rojas, R. (2005). Recognition of on-line handwritten mathematical expressions in the e-chalk system-an extension. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 1206–1210, Seoul, Korea (South). https://doi.org/10.1109/ICDAR.2005.197.

Tapia, Ernesto, & Rojas, R. (2003). Recognition of on-line handwritten mathematical formulas in the e-chalk system. *Seventh International Conference on Document Analysis and Recognition*, *3*, 980–984, Georgia, USA. https://doi.org/10.1109/ICDAR.2003.1227805.

E.M. Taranta J.J. LaViola Jr Math boxes: A pen-based user interface for writing difficult mathematical expressions Proceedings of the 20th International Conference on Intelligent User Interfaces 2015 87 96 10.1145/2678025.2701400.

Taranta, E. M., Vargas, A. N., Compton, S. P., & Laviola, J. J., Jr (2016). A Dynamic Pen-Based Interface for Writing and Editing Complex Mathematical Expressions With Math Boxes. *ACM Transactions on Interactive Intelligent Systems (TiiS), 6*(2), 1–25. https://doi.org/10.1145/2946795

S.W. Thomas Mining software repositories using topic models 2011 1138–1139 Honolulu, HI, USA 10.1145/1985793.1986020.

K. Tian M. Revelle D. Poshyvanyk Using latent dirichlet allocation for automatic categorization of software 2009 Vancouver, BC, Canada 163 166 10.1109/MSR.2009.5069496.

M.A. Toçoğlu A. Onan Sentiment analysis on students' evaluation of higher educational institutions International Conference on Intelligent and Fuzzy Systems 2020 10.1007/978-3-030-51156-2_197.

Vuong, B.-Q., He, Y., & Hui, S. C. (2010). Towards a web-based progressive handwriting recognition environment for mathematical problem solving. *Expert Systems with Applications, 37*(1), 886–893. https://doi.org/10.1016/j.eswa.2009.05.091

Vuong, B.-Q., Hui, S. C., & He, Y. (2008). Progressive structural analysis for dynamic recognition of on-line handwritten mathematical expressions. *Pattern Recognition Letters, 29*(5), 647–655. https://doi.org/10.1016/j.patrec.2007.11.017

Wang, D.-H., Yin, F., Wu, J.-W., Yan, Y.-P., Huang, Z.-C., Chen, G.-Y., Wang, Y., & Liu, C.-L. (2020). ICFHR 2020 Competition on Offline Recognition and Spotting of Handwritten Mathematical Expressions-OffRaSHME. *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 211–215, Dortmund, Germany.

Wang, F., Zhang, J. L., Li, Y., Deng, K., & Liu, J. S. (2021). Bayesian text classification and summarization via a class-specified topic model. *Journal of Machine Learning Research, 22*(89), 1–48.

Wang, H., & Shan, G. (2020). *Recognizing Handwritten Mathematical Expressions as LaTex Sequences Using a Multiscale Robust Neural Network*. Computer Vision and Pattern Recognition.

Wang, J., Du, J., & Zhang, J. (2020b). Stroke Constrained Attention Network for Online Handwritten Mathematical Expression Recognition. *Pattern Recognition, 119*, 1–29. https://doi.org/10.48550/arXiv.2002.08670.

J. Wang J. Du J. Zhang Z.R. Wang Multi-modal attention network for handwritten mathematical expression recognition 2019 Sydney, NSW, Australia 10.1109/ICDAR.2019.00191.

Wang, M., Gao, S., Gui, W., Ye, J., & Mi, S. (2022). Investigation of Pre-service Teachers' Conceptions of the Nature of Science Based on the LDA Model. *Science & Education*, 1–27.

Wells, M. B. (1976). Preprocessing of typed two-dimensional mathematical expressions. *ACM SIGPLAN Notices, 11*(9), 25–37. https://doi.org/10.1145/987500.987505

Winkler, H.-J., Fahrner, H., & Lang, M. (1995). A soft-decision approach for structural analysis of handwritten mathematical expressions. *1995 International Conference on Acoustics, Speech, and Signal Processing*, *4*, 2459–2462, Detroit, MI, USA. https://doi.org/10.1109/ICASSP.1995.480046.

Winkler, H. J. H.-J. H. J., & Lang, M. (1997). On-line symbol segmentation and recognition in handwritten mathematical expressions. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *4*, 3377–3380. https://doi.org/10.1109/icassp.1997.595518.

Wu, J. W., Yin, F., Zhang, Y. M., Zhang, X. Y., & Liu, C. L. (2020). Handwritten Mathematical Expression Recognition via Paired Adversarial Learning. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-020-01291-5

Wu, J., Yin, F., Zhang, Y., Zhang, X., & Liu, C. (2021). Graph-to-Graph : Towards Accurate and Interpretable Online Handwritten Mathematical Expression Recognition. *AAAI Conference on Artificial Intelligence, 35*, 2925–2933.

Wu, Y., Liu, M., Zheng, W. J., Zhao, Z., & Xu, H. (2012). Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. *Pacific Symposium on Biocomputing, 2012*, 422–433. https://doi.org/10.1142/9789814366496_0041

Xiangwei, Q., & Abaydulla, Y. (2010). The study of mathematical expression recognition and the embedded system design. *Journal of Software, 5*(1), 44–53. https://doi.org/10.4304/jsw.5.1.44-53

C.H.W.Q.X. Xiaorong Z. Chaoying A Survey of Mathematical Expression Auto-recognition 2004 Guangxi Sciences.

Yang, B., Wang, X., & Ding, Z. (2021). Understanding Service Providers' Competency in Knowledge-Intensive Crowdsourcing Platforms: An LDA Approach. *Complexity*, *2021*.

Yingying, J., Xiang, A., Feng, T., Xugang, W., & Guozhong, D. (2009). Error Correction for Handwritten Mathematical Expression Recognition by Pen and Speech. *Journal of Computer Research and Development, 46*(4), 689.

S. Yousefi M.P. Nguyen N. Kehtarnavaz Y. Cao Facial expression recognition based on diffeomorphic matching 2010 4549–4552 Hong Kong, China 10.1109/ICIP.2010.5650670.

Zanibbi, R., & Blostein, D. (2012). Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition, 15*(4), 331–357. https://doi.org/10.1007/s10032-011-0174-4

Zhang, J., & Hong, L. (2008). A survey on recognition of on-line handwritten mathematical expression. *Journal of Huaibei Coal Industry Teachers College (Natural Science Edition), 29*(3). https://doi.org/10.17169/refubium-23077.

Zhang, J., Du, J., & Dai, L. (2017a). Track, Attend, and Parse (TAP): An End-to-End Framework for Online Handwritten Mathematical Expression Recognition. *IEEE Transactions on Multimedia, 21*(1), 221–233. https://doi.org/10.1109/TMM.2018.2844689

Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., … Dai, L. (2017b). Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition Letters, 71*, 196–206. https://doi.org/10.1016/j.patcog.2017.06.017

Zhang, T., Mouchère, H., & Viard-Gaudin, C. (2018). A tree-BLSTM-based recognition system for online handwritten mathematical expressions. *Neural Computing and Applications, 2*(1). https://doi.org/10.1007/s00521-018-3817-2

Zhang, Y., Chen, M., Huang, D., Wu, D., & Li, Y. (2017c). iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems, 66*, 30–35. https://doi.org/10.1016/j.future.2015.12.001

Zhao, J., Huang, J. X., Deng, H., Chang, Y., & Xia, L. (2021). Are topics interesting or not? An LDA-based topic-graph probabilistic model for web search personalization. *ACM Transactions on Information Systems (TOIS), 40*(3), 1–24.

Zheleznia kov, D., Zaytsev, V., & Radyvonenko, O. (2021). Online Handwritten Mathematical Expression Recognition and Applications: A Survey. *IEEE Access, 9*, 1–24. https://doi.org/10.1109/ACCESS.2021.3063413

Zhu, S., Hu, L., & Zanibbi, R. (2013). Rotation-robust math symbol recognition and retrieval using outer contours and image subsampling. *Document Recognition and Retrieval XX, 8658*, 1–12. https://doi.org/10.1117/12.2008383