Analysis of Natural Language Processing in Dental Informatics

Hsin-Hsuan Chung
University of North Texas
Hsin-HsuanChung@my.unt.edu

Preetham Reddy Yaramada
University of North Texas
PreethamReddyYaramada@my.unt.edu

Sai Varun Pinna
University of North Texas
SaiVarunPinna@my.unt.edu

Simi Maithani
University of North Texas
SimiMaithani@my.unt.edu

Surekha Maddi
University of North Texas
SurekhaMaddi@my.unt.edu

Thi Phan
University of North Texas
ThiPhan@my.unt.edu

ABSTRACT

Dental informatics is a subcategory of health informatics that focuses on using computers and information science to improve dental practice, research, education practice. Also, dental informatics has grown rapidly in recent years, and it affects people's daily lives. Plus, this field of students is progressing a lot. Thus, the main purpose of this study is to extract insights into the dental informatics research area. Our objective is to showcase the dental informatics research progress and growth over the years. Besides, the goal of this study is to quantify the academic output of NLP in the realm of the dental informatics research field. The data is collected from multiple databases related to dental information from 2001 to 2020 and performed data descriptive and statistical analysis, citation analysis, and author's collaborations analysis. Most articles related to dental informatics are published in 2013, 2011, and 2008. Maximum dental articles are published in fields of medicine, computer science, and psychology. Advances in dental research, the journal of dental education, and studies in health technology and informatics are the top journals published in dental articles. Our results show significant insights into the dental informatics area.

Github link of the project:
https://github.com/maddisurekha1234/Dental_Informatics_Project

Keywords

Dental Informatics, LDA Modeling, Frequency and occurrence analysis, massive literature data

1. INTRODUCTION

In Dentistry, Dental Informatics is concerned with the intersection of Health Informatics and Dentistry as a whole. The field is growing in academic circles as well as within the profession; more dental schools and practices are embracing it. More and more dental schools and dental practices are implementing Electronic Health Records (EHR) systems and Health Information Exchanges (HIEs) are beginning a transformation in health care communications. The demands of federal and state programs to promote EHR adoption among certain health care providers are also beginning to affect dentists across the country. As a result of these changes, dentistry is engaging information technology to meet its clinical, administrative, research, and educational needs more than ever.

This project is aimed at understanding the research progress on Dental Informatics through a bibliometric analysis of the related publications. The first step was to collect all the related publications from the past 20 years: 2001-2020. Different databases were used to extract the publications related to Dental Informatics using the keyword "Dental Informatics". The data was collected from the following resources: Google Scholar, Web for science, Scopus, Springer, Science Direct, PubMed, PubMed Central, Embase. These resources are some of the biggest places that are being used in the current time to extract information and hold a hub of information on Dental Informatics. Descriptive techniques are used to understand the data further and to make the data ready for Topic modeling.

1.1 Goals

In this paper, we wanted to determine the article's size, the number of authors, its distribution across publications, its growth rate, its general content, its trend, the relationship between citations and publications, characteristics of citations, etc. We also use topic modeling to gain an understanding of the topic frequency, term frequency, and topic coherence score.

- Analyze the total number of articles published each year.
- Find the top 10 venue publications based on the total number of articles.
- Find top keywords in the titles of Dental Informatics articles.
- Find which field has the greatest number of articles
- Find the most cited paper per year for Dental Informatics.
- Find the total citations per year.

- Find the top 10 authors based on their publications.
- Analyzing the article and finding their reference article count over the period.
- Using Intertopic Distance Map (via multidimensional scaling) and finding Most Relevant terms in each topic.
- Finding the Coherence and Perplexity score of articles using Topic Modeling.
- Analyzing author's collaboration for all articles from 2001-2020.

1.2 Research questions

- What are the major research areas of Dental Informatics and how have they changed over the years?
- What is the relationship between Dental Informatics and IT in Dentistry?
- How does the yearly publication output in Dental Informatics compare with IT Dentistry?
- What has been the highest impact of institutions and researchers based on the data?
- Is there any collaboration of authors based on the data collected (more than two authors)?
- What is the impact of dental informatics research (on other papers, how do they differ)?

To perform analysis on the data, different data cleaning operations have been performed. Basic pre-processing such as converting the data to lower case to maintain uniformity, stop words, lemmatization, stemming, etc. have been performed. Performing topic modeling using LDA is the focus of this project and major kinds of analyses will be made using this.

2. LITERATURE REVIEW

2.1 The importance of receiving dental information

In recent years, Dental Informatics has grown rapidly, and it is related to many fields, such as technology, information and computing science, and communications [13]. Dental informatics, a novel interdisciplinary between general dentistry and medical informatics, has become famous [18]. The field of dental informatics affects people's daily lives [23]. Many people search for dental information when they encounter dental problems or before they receive dental care [14]. Because people's dental knowledge level only depends on their education level and previously acquired dental information [27], people have to know how to select medical websites or professional organizations to obtain higher-quality information [3]. In addition, people know more about crucial dental information, they have better oral health status [7]. To be more specific, children are less likely to have fundamental dental background knowledge [16]. It is why about 80% of 5-15 years old children are impacted by negative dental issues, such as dental caries [11]. Fortunately, teenagers have better oral health status since they can maintain their teeth and take responsibility for taking care of them [2]. Moreover, many young adults are interested in different issues of dental health, such as the best toothpaste, dental injuries, and orthodontic appliances. Therefore, they desire to be more familiar with specific dental information [2] and prefer to receive dental information from oral health care professionals [27]. Namely, not only do non-professional people want to understand dental information but also dentists are required to update and improve their dental knowledge [24]. Dentists have to keep up with the developments of dental informatics because dental informatics brings plenty of clinical practices applications and tools [19]. Dental practice increasingly relies on electronic information systems [13]. Thus, clinical dental information systems are designed for dentists to fulfill their demand for dental information. The systems integrate clinical dental evidence, provide timely dental information, and demonstrate better visual dental problems [24]. Overall, whomever people treat or are treated for dental service, receiving dental information is vital [29].

2.2 Applications of Dental Informatics - Human Identification by dental information

When forensic investigators execute the human identification procedure, they can apply multidisciplinary methods, such as anthropology, medicine, dentistry, and genetics; among these approaches, dental examinations are a low-cost choice [21]. In other words, the identity of the victims can be determined by dental information that is a reliable and speedy identified method [20]. One example of the identified method is dental human identification which compares the antemortem and postmortem dental data [9]. Because most people have dental treatment before they die, forensic dental doctors can search for their dental records and use dental information to determine their identity [21]. Dental information typically contains x-ray-based radiographs, panoramic radiographs, and orthopantomography (OPGs). This information can establish the identity of human remains [26]. Forensic dental can also focus on several unique characteristics of dental information, such as dilacerations and bifurcations, in dental roots to make the identification procedure easier [9]

2.3 Dental Informatics improve the delivery of dental care

Most dentists in developing countries are unfamiliar with dental informatics, not to mention ordinary people [10]. More specifically, People in developing countries have fewer chances to receive dental services than the ones in developed countries [28]. Few dental clinics in developing countries, poor living conditions, and low-income lead to the low delivery of dental care [5], [17]. Having the internet is easier for developing countries than having dental services, so developing countries should utilize social media and suitable methods to convey health information to increase people's health literacy and health behavior [6]. Adopting computer applications to promote dental is the purpose of dental informatics [22]. Plus, developing countries should try to properly apply the resources for people to access oral and dental services [6]. Hence, a program was designed to deliver dental information. The program provides general dental information and regular health counseling [17]. Offering dental information for people in developing countries can boost their dental informatics knowledge [15].

2.4 Latent Dirichlet Allocation (LDA)

Researchers can apply various topic modeling methods such as Latent Dirichlet Allocation (LDA) when conducting a set of experimentations [1]. First, LDA is a topic model that collects discrete data and can give the topic of each document in the

document set in the form of a probability distribution [8]. Also, LDA is a hierarchical Bayesian probabilistic model of text documents and is an unsupervised learning algorithm [4]. It does not require a manually labeled training set during training. It just needs a document set and the number of specified topics k [12].

3. DATA COLLECTION

Data collection is a systematic approach of gathering and measuring information from various sources.

3.1 Multiple stages in the data collection process

Our goal is extracting the Dental informatics articles published between 2001 to 2020. Below are the different tasks performed in three stages in the data collection process.

3.1.1 Stage 1

We have searched in the below publication databases (refer table 1) with the keyword Dental Informatics and using advanced search and limited our search history to years 2001 to 2020 articles. Then extracted the articles information related to the subtopics (refer table 2) data into excel sheet with the column like database, year, title of the article, authors, venue, DOI. After the initial search we have retrieved about 200 articles by looking at the title of the article and abstract. A lot of them don't have DOI's and few of the articles are not retrievable.

Table 1. Referred Databases information

S. No	Databases
1.	PubMed
2.	PubMed central
3.	Google scholar
4.	Science direct
5.	Web of science
6.	Embase
7.	Scopus
8.	Springer
9.	Medline

3.1.2 Stage 2

In the second stage, we have filtered the articles further down by reading the abstract and full text. There are only a few articles that have access to the full text. We have divided the articles into Mesh – medical subject headings into 6 subtopics. In this stage we have filtered inaccessible, unrelated articles and the final count of the articles is 129. All these articles don't have proper DOI's and for the articles that don't have DOI's, there are PubMed or PubMed central ID's. We have noted the proper links to all the 129 articles where we can access the information of the article. We read the abstract and some full-text articles and we have written the summary of each article like aim or objective of the article, methods they have used, results, and conclusion of these 129 articles. In the end, we prepared the final file of the metadata.

3.1.3 Stage 3

Once high-level metadata is available in a csv file for 129 articles. Our next goal is to extract the in-depth metadata extraction of each article. We have used Semantic DOI API and PubMed API to extract the article metadata extraction in json files. We have implemented the python code which takes the article DOI or PMICD as an input hit the API and reads the metadata which is json file format. Below are the links.

Table 2. Medical Subtopics information

S. No	Subtopics
1.	General overview
2.	Patient centered health care providers
3.	Dental education and research
4.	Digital dentistry and information technology
5.	Electronic health records
6.	Innovations, challenges, and opportunities

Semantic DOI API link: https://api.semanticscholar.org/v1/paper/

PubMed article metadata extraction:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=22368089&tool=my_tool&email=my_email@example.com

3.2 Data Processing and Cleaning

As we have 129 articles metadata available in json files, the next step is to parse the json files and process and clean the data. We have written the python code which reads the json file based on the tags and stores the data in a data frame.

We have performed multiple data cleaning steps like replacing null values, removing stop words, converting them into lowercase, spelling correction. After performing these all steps, the clean data is available for data analysis.

Table 3. Statistical information of the data collection

Total Dental Informatics Paper	129
Total Authors	631
Total Citations	1921
Total number of Conferences	91

4. METHODOLOGY

We perform different data cleaning operations. In data cleaning different steps are used for converting all into lower case and replacing punctuation marks, stop words, lemmatization, and stemming to get basic word extension. We search for sequence databases for sequences that are like the query sequence is the most common type of database similarity search.

4.1 Topic Modeling using LDA

In communication research, topic models based on Latent Dirichlet allocation (LDA) are becoming more popular. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities. However, concerns about the approach's dependability and validity have gotten little attention so far. FASTA was the first rapid search method, which detected brief common patterns in the query and database sequences and aligned them. The Basic Local Alignment Search Tool (BLAST), which is comparable to FASTA but can search rarer and more significant sequences, is gaining prominence. The PAM250 Scoring Matrix is a scoring matrix based on an evolutionary model that predicts amino acid type changes through time. PSI-BLAST and PHI-BLAST are iterative search algorithms that outperform BLAST and FASTA in terms of detection rate. Protein domain family analysis is done with Pfam and SMART (Simple Modular Architecture Research Tool).

4.2 Data Mining

EDRs include a wealth of useful information and analyzing it can give insight into how dental practitioners can enhance patient health by recognizing dental illness early. This form of analysis allows dentists to learn from their own data as well as aggregated data from other practices to compare their practice to that of their peers. Researchers can also use this real-world data to better assess the quality of care provided to various groups. Small individual clinics, major dental service companies, and dental educational institutions may all benefit from dental data analytics and machine learning procedures.

4.3 Converting to .csv files

Because CSV files have such a simple format, they can be opened and exported by almost any application. It's also simple to convert an Excel file into a CSV file if you need to remove additional formatting and only store your data.

Visualization In this project, we are using python libraries like matplot, nltk etc for visualizing our outputs from LDA topic modeling.

4.4 Collaborative analysis

To build an author collaboration network, which counts and visualizes the authors' work and their collaboration between the articles. We analyze the data and extract the insights information of the author's work and create a network graph.

4.5 Word Count

Word Count is the technique to find the word with the frequency in the article titles and then we can normalize different forms of words. This technique helps to most frequently used words in the article titles in the Dental Informatics area. We visualize the word counts graph using Tableau.

4.6 Citation Analysis

Citation analysis plays a major role in the scientific disclosure of the health informatics area. It is basically finding the citation between the articles based on the authors' work. This is the key point to analyze the scientific paper and make decisions on the impact of the paper. Here, we combine citation count analysis and citation intent analysis (sometimes called citation content location, meaning where these references are mentioned in the citing article to analyze the impact of Dental informatics papers in and outside the IR community. Specifically, we first collect information including total citations received, total references, and citation intent then conduct statistical analysis regarding these aspects and visualize the results with Tableau.

5. DATA ANALYSIS

The process of inspecting, cleaning, transforming, and modeling data with the objective of identifying usable information and informing conclusions is known as data analysis. In our project we have also followed the steps and performed multiple tasks in each stage.

5.1 Data Inspection and Cleaning

Once we have the low-level metadata information of articles in a csv file, we have analyzed the data in every corner and drawn conclusions on the data extraction. Below are a few steps.

1. We have analyzed the CSV file and found out that how many articles have the proper DOI information and how many articles do not have the DOI. For the articles

which have the DOI, we extracted the metadata JSON files from semantic API, and for the rest of the article which doesn't have DOI, we have extracted using PMICD.

2. We have also analyzed how is the raw text data. Is the data noisy or not to draw a conclusion on data cleaning techniques?
3. In the third step, we have performed the data cleaning techniques to clean the data and make it ready for data transformations.

5.2 Data Transformations

Once the cleaned data is available our next goal is to apply data transformations. It is a process of converting raw data into a format or structure that would be more suitable for the model to discover the insights. In our project, we have used multiple data transforming techniques.

1. Data mapping: As our input data is in JSON file format, we have extracted the data from each key and value pair in JSON files and saved it as a python data frame format which will be very easy for further analysis.
2. Aggregations: As our goal of the analysis is to showcase the research process in different dental fields of studies, by journals, by over the year, etc. We have applied the python aggregations techniques to extract the insights from the data.
3. Visualization: A good visualization can aid in properly explaining the findings of the research and can also be used to better comprehend the dataset. We have also showcased our results using python libraries like matplot, nltk, etc for a better understanding of results.

6. RESULTS AND DISCUSSION

6.1 Analyzing the articles published each year

The number of articles published in dental informatics was the highest in 2013. There were 11 articles published in the year 2013. There were two other years where the same number of articles were published. They are 2008 and 2011.

	Year	No.of articles
0	2013.0	11
1	2008.0	11
2	2011.0	11
3	2003.0	10
4	2014.0	10
5	2017.0	10
6	2019.0	9
7	2007.0	9
8	2015.0	6
9	2016.0	6
10	2009.0	6
11	2020.0	6
12	2012.0	5
13	2018.0	5
14	2010.0	5
15	2002.0	4
16	2005.0	2
17	2006.0	1
18	2001.0	1
19	2004.0	1

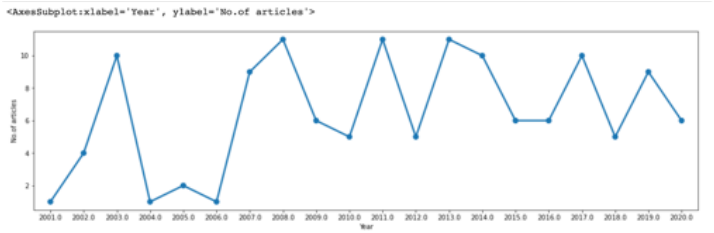


Figure 1. The number of dental informatics articles published in each year from 2001 to 2020

6.2 Top 10 cited venues

Advances in Dental Research (ADR) is the venue that includes the most dental informatics articles.

Reasonable explanations:

- ADR is a journal covering the technologies and fields related to Medicine.
- ADR was founded in 1987.

	venue	count
0	Advances in Dental Research	8
1	Journal of dental education	6
2	Studies in Health Technology and Informatics	5
3	Journal of Dental Education	4
4	International Journal of Medical Informatics	4
5	Journal of American Dental Association (1939)	4
6	The journal of evidence-based dental practice	2
7	Journal of the American Medical Informatics As...	2
8	British Dental Journal	2
9	Journal of dental education	2

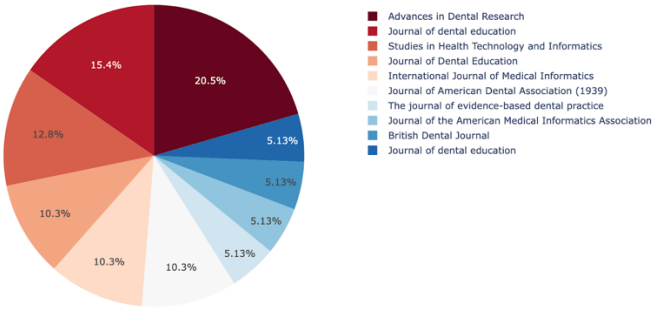


Figure 2. Top 10 cited venues

6.3 Top keywords in the titles of Dental Informatics articles

The image shows the top keywords in the title of dental informatics articles from 2001 to 2020.

This is a word cloud that shows the words that appeared more frequently in the articles. The size of the word in the word cloud is directly proportional to the frequency of the word in the articles. In this word cloud, we can see that the word dental has the highest frequency among all the other words. Some of the other words are health record, electronic resources, informatics, etc.

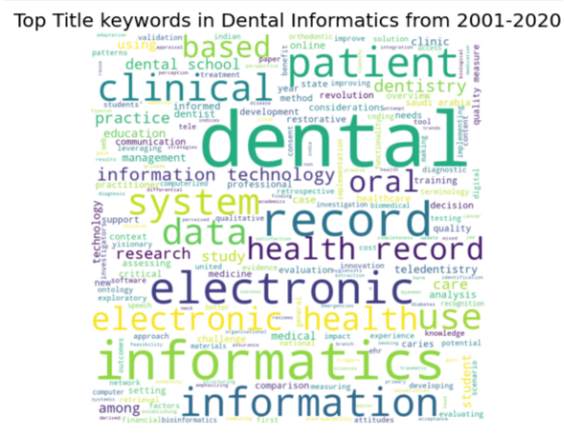


Figure 3. Top keywords in the titles of dental informatics articles

6.4 Which field has the most number of articles?

The field with the largest number of dental informatics articles is medicine, followed by computer science.

Reasonable explanations:

- Dental informatics is a subspecialty of health informatics, so it is most related to medicine.
- Several dental informatics articles apply deep learning, artificial intelligence methods, and Convolutional Neural Network (CNN) to conduct their research.

	Field	Articles
0	Medicine	119
1	Computer Science	40
2	Psychology	5
3	Business	5
4	Engineering	4
5	Political Science	1
6	Materials Science	1

Figure 4. Top fields with the most dental informatics articles

6.5 The most cited articles per year

The result presents the most cited dental informatics articles from 2001 to 2020.

	Unnamed: 0	Title	Year_y	TotalCitationCount_y
5	5	Using big data to promote precision oral heal...	2020	10
14	14	Health Data in Dentistry: An Attempt to Master...	2019	16
19	19	Feasibility of Electronic Health Record–Based ...	2018	9
29	29	Perceived critical success factors of electron...	2017	26
35	35	Measuring up: Implementing a dental quality me...	2016	20
41	41	Does use of an electronic health record with d...	2015	11
51	51	Exploring design requirements for repurposing ...	2014	46
62	62	Electronic dental record use and clinical info...	2013	35
67	67	From information technology to informatics: th...	2012	58
78	78	Evaluating a dental diagnostic terminology in ...	2011	43
83	83	Consortium for oral health-related informatics...	2010	40
89	89	Using teledentistry to improve access to denta...	2009	115
100	100	Potential of information technology in dental ...	2008	116
109	109	A qualitative investigation of the content of ...	2007	43
110	110	Design of tele-educational system for interpr...	2006	0
112	112	Use of information and communication technolog...	2005	85
113	113	Should dentistry be part of the National Healt...	2004	18
123	123	Teledentistry and its use in dental education	2003	123
127	127	Electronic patient records for dental school c...	2002	48
128	128	Dental informatics. A cornerstone of dental pr...	2001	77

Figure 5. The most cited dental informatics articles per year

6.6 Total citations per year

In 2003, dental informatics articles had the highest total number of citations.

Reasonable explanations:

- Chen et al. (2003) article has been cited 123 times. Also, there were 9 dental informatics articles published in 2003.
- Chen, J. W., Hobdell, M. H., Dunn, K., Johnson, K. A., & Zhang, J. (2003). Teledentistry and its use in dental education. The Journal of the American Dental Association, 134(3), 342-346.

	Year	TotalCitationCount
0	2001	77
1	2002	75
2	2003	278
3	2004	18
4	2005	99
5	2006	0
6	2007	116
7	2008	233
8	2009	223
9	2010	104
10	2011	152
11	2012	82
12	2013	91
13	2014	108
14	2015	17
15	2016	37
16	2017	117
17	2018	29
18	2019	33
19	2020	32

Figure 6. Total citations of dental informatics articles per year

6.7 LDA Topic Modeling

We use the LDA topic modeling algorithm to see the top topics in the dataset and visualize the topic keyword. We picked the column "Abstract" which summarizes the articles' content. We build an LDA model with 20 different topics where each topic is a combination of keywords, and each keyword contributes a certain weightage to the topic.

```
[(0,
  '0.045*"systems" + 0.044*"usability" + 0.039*"four" + 0.034*"speech" + '
  '0.030*"lack" + 0.025*"functionality" + 0.020*"recognition" + 0.018*"system" '
  '+ 0.018*"show" + 0.016*"dental"'),
 (1,
  '0.053*"authors" + 0.039*"common" + 0.035*"users" + 0.027*"identified" + '
  '0.027*"index" + 0.021*"participants" + 0.015*"holistic" + '
  '0.015*"documentation" + 0.015*"documented" + 0.015*"nine"'),
 (2,
  '0.045*"dental" + 0.034*"study" + 0.032*"clinical" + 0.031*"health" + '
  '0.029*"care" + 0.020*"patients" + 0.018*"electronic" + 0.017*"dentists" + '
  '0.016*"patient" + 0.016*"using"'),
 (3,
  '0.023*"ai" + 0.023*"presents" + 0.010*"coi" + 0.010*"diagnosed" + '
  '0.010*"genuine" + 0.010*"tmd" + 0.010*"tmdmimicking" + 0.010*"disorders" + '
  '0.010*"conditions" + 0.005*"conceptual"'),
 (4,
  '0.041*"problems" + 0.029*"part" + 0.028*"standard" + 0.026*"area" + '
  '0.023*"standards" + 0.020*"patient" + 0.015*"discusses" + 0.014*"goals" + '
  '0.014*"wide" + 0.014*"guidelines"'),
 (5,
  '0.039*"edrs" + 0.027*"oral" + 0.026*"status" + 0.018*"smoking" + '
  '0.018*"authors" + 0.017*"user" + 0.016*"using" + 0.016*"patients" + '
  '0.015*"successfully" + 0.013*"designed"'),
 (6,
  '0.057*"new" + 0.043*"oral" + 0.029*"system" + 0.025*"healthcare" + '
  '0.021*"completed" + 0.017*"implementation" + 0.014*"literature" + '
  '0.012*"managing" + 0.012*"research" + 0.011*"project"'),
 (7,
  '0.022*"cds" + 0.011*"subdivision" + 0.011*"development" + 0.006*"creating" '
  '+ 0.006*"association" + 0.006*"committee" + 0.006*"ada" + '
  '0.006*"infomatics" + 0.006*"american" + 0.006*"standars"'),
 (8,
```

Figure 7. The top 8 topics keywords

Model perplexity and topic coherence provide a convenient measure to judge how good a given topic model is. In our experience, the topic coherence score has been more helpful because it helps in determining the relative distance between different words within a topic.

	num topics	Coherence	Perplexity
0	5	0.341898	-7.236973
1	6	0.379693	-7.274034
2	7	0.370659	-7.303521
3	8	0.360547	-7.331016
4	9	0.340588	-7.339491
...	...	...	...
90	95	0.561665	-7.786447
91	96	0.567628	-7.793448
92	97	0.531173	-7.811646
93	98	0.521034	-7.801680
94	99	0.592094	-7.804403

95 rows × 3 columns

Figure 8. Model perplexity and topic coherence

Examine the product topics and the associated keywords: each bubble on the left plot represents a topic. The larger the bubble, the more prevalent that topic is.

The topic keyword in our first experiment is “dental.”

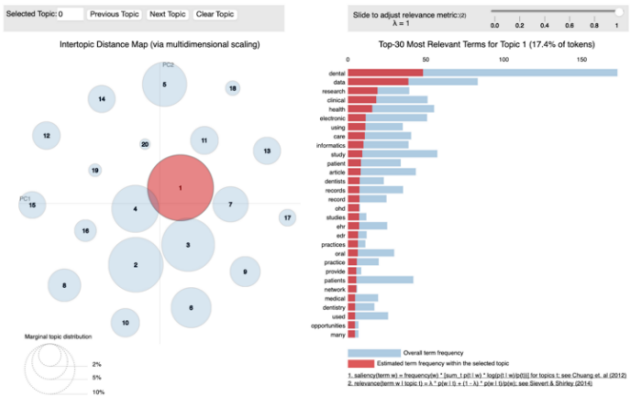


Figure 9. The initial result of LDA topic modeling

To optimize the number of topics for LDA, we build many LDA models with different values of the number of topics and pick the one that gives the highest coherence value.

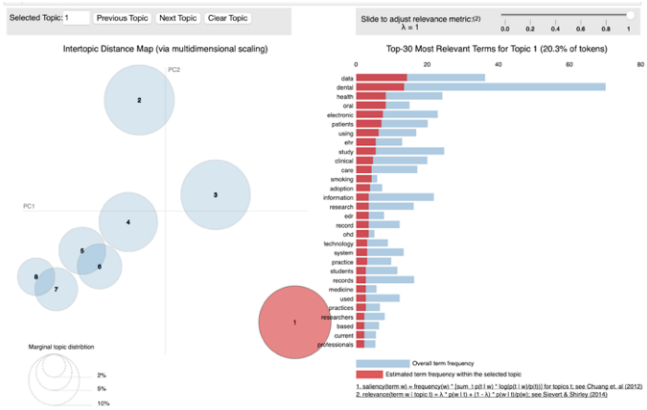


Figure 10. The result of LDA topic modeling with k=8

In our conclusion, the good topic model has fairly big and non-overlapping bubbles. The coherence scores (from k = 8 to k = 20) keep increasing and it is better to pick the model that gives the highest C_V. Therefore, we pick a model with 20 topics and the keyword is “dental.”

6.8 Term occurrence in the articles

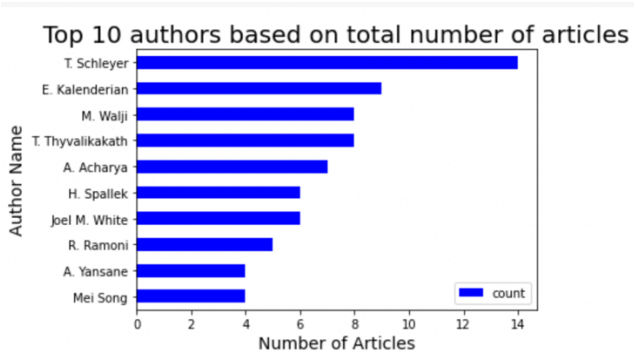
This shows the term and its occurrence frequency in the articles. In this study, the term has the highest frequency of 8 in article number 734.

	term	occurrences
734	studi	8
728	student	7
195	dental care	7
24	among	7
243	develop	6
480	medic	6
294	evalu	6
592	practic	6
627	qualiti	6
529	oral health	5
387	improv	4
570	patient record	4
99	challeng	4
642	record data	4
219	dental school	4
270	educ	4
703	school	4
236	design	4

Figure 11. Term occurrence in the dental informatics articles

6.9 Top 10 authors

The below graph shows the top 10 authors of the publications based on the total number of articles they have written. We can see that e T. Schleyer wrote the maximum number of articles with a total count of 14.



	name	count
29	T. Schleyer	14
23	E. Kalenderian	9
11	M. Walji	8
24	T. Thyvalikakath	8
40	A. Acharya	7
54	H. Spallek	6
55	Joel M. White	6
132	R. Ramoni	5
12	A. Yansane	4
39	Mei Song	4

Figure 12. Top 10 authors in the dental informatics articles

6.10 Topics for each year and ranking the years based on the number of topics for each year

The table shows the main topic for that year based on the number of topics for each year and the ranking the years based on the topics of that year. In the year 2019, there were two articles published on Oral Health followed by Silo(dataset), precipitating factors, deciduous tooth, etc.

	topics	Rank	Year
2	Oral health	1.0	2019
40	Silo (dataset)	2.0	2013
11	Precipitating Factors	3.0	2018
28	Deciduous tooth	4.0	2016
0	Clinical Data	5.0	2020
48	Head and neck structure	6.0	2012
73	Contain (action)	7.0	2005
33	General Practice, Dental	8.0	2015

Figure 13. Top 8 topics in the dental informatics articles

6.11 Comparison of citations and references each year for the venue "Journal of the American Dental Association"

This graph shows the comparison between the citations and references for the venue journal of the American dental association. From this graph, we can say that the citations and references for the journal have increased in popularity from the year 2010. As expected, the number of citations per year would be more than the number of references per year because there can be more than one citation for one paper.

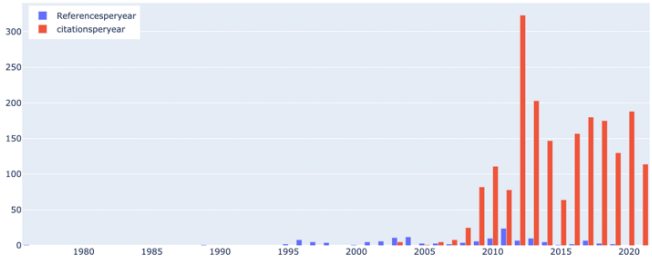


Figure 14. Comparison of citations and references each year for "Journal of the American Dental Association"

6.12 Author collaboration

The below table shows the article_id and the different authors that have collaborated for the article.

article_id	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
10.1002/J.0022-0337.2005.69.3.TB03927.X	L. Rajab	Z. Baqain																
10.1002/J.0022-0337.2011.75.5.TB05086.X	Joel M. White	E. Kalenderian	P. Stark	Rachel L Ramoni	R. Vaderhobli	M. Walji												
10.1002/J.0022-0337.2012.76.1.TB05241.X	T. Schleyer	T. Thyvalikakath	H. Spallek	Michael P. Dziabiak	L. Johnson													
10.1007/978-3-319-17272-9_9	E. Kalenderian	M. Walji	R. Ramoni															
10.1007/978-3-319-98298-4_11	M. Torres-Urquidy	V. J. H. Powell	S. Geist	Sushma Mishra	Monica Chaudhari	Mureen Allen												
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10.4338/ACI-2015-11-RA-0150	K. Schwei	R. Cooper	Andrea Mahnke	Zhan Ye	A. Acharya													
10.5210/ojphi.v11i2.10131	S. Swanik																	
10.5455/aim.2012.20.47-55	Fedja Masic																	

Figure 15. Author collaboration for the dental informatics articles

7. CONCLUSION AND LIMITATIONS

7.1 Conclusion

Dental Informatics is a novel interdisciplinary between general dentistry and medical informatics. It is one of the rapidly growing fields of science. The paper aimed at understanding the research progress on Dental Informatics by performing bibliometrics on related published articles. A total of 129 published articles from 2001-2020 were scrapped from different medical databases such as Google Scholar, PubMed, Web for Science, etc. Descriptive analytics and topic modeling were performed to analyze the articles and develop a better understanding of the published articles in dental informatics. After performing the analysis, it was found out that the maximum number of articles were published in the years 2008, 2011, and 2013. The venue in the articles that appeared the maximum number of times is Advances in Dental Research. Advances in Dental Research acts as a forum and holds all the explorations and discussions related to dentistry and other related fields. For understanding the content of the articles automatic keyword extraction from the titles was used to find out the most commonly occurring words in the titles.

We found out that a lot of articles discussed health records, dentistry, using information technology, and electronic resources in dentistry. There are different fields of study applied in dental information. Dental informatics is applied the most in the field of medicine. There were nine dental informatics articles published in the year 2003 because of which dental informatics articles have the highest number of citations that year. LDA topic modeling algorithms were used to understand the content of the articles better by generating different topics from the abstract of all the articles. Model perplexity and topic coherence values helped to judge how good a given topic model is. We also performed analysis on the authors of the articles and were able to extract the top 10 authors based on the number of articles they have written. Among the 129 articles, 14 articles were written by T,Shleyer. This research project can be used to understand the relevance of dental informatics and the pace with which it is growing in the current world.

7.2 Limitations

First, the main limitations of this study come from the selection of databases because the study only used 9 databases. However, ERIC, IEEE Xplore, ScienceDirect, Directory of Open Access Journals (DOAJ), and JSTOR are also on the top list of academic research databases. Besides, the articles in the field of dental

informatics are few, plus there are many articles that do not have DOI, so the study only has 129 suitable articles to conduct the analysis. Second, we do not have good experience in tableau, so we did not visualize the author citation and collaboration results in the network graph. Also, we do not have a good knowledge of the dental informatics area, in other words, we were not able to validate the results and did not know how to interpret the results and give reasonable explanations. Therefore, we needed to depend on the support and assistance of health informatics department students.

8. AUTHOR CONTRIBUTIONS

- Hsin-Hsuan Chung: Literature review, Data Collection, Results and Discussion, Data descriptive Analysis, created and synthesized the document
- Preetham Reddy Yaramada: Methodology, Make PPT
- Sai Varun Pinna: Make PPT
- Simi Maithani: Introduction, Data descriptive Analysis, Conclusion and Limitations
- Surekha Maddi (Team leader): Data collection and cleaning, Data Analysis, Data extraction and author collaboration, Data Collection, Make PPT
- Thi Phan: Experiment and Data analysis plan, Topic Modeling, Results, and Discussion

9. ACKNOWLEDGMENTS

- Naveen Katragadda: Data collection

10. REFERENCES

[1] Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, *189*, 191-194.

[2] Adekoya-Sofowora, C. A., Lee, G. T., & Humphris, G. M. (1996). Needs for dental information of adolescents from an inner-city area of Liverpool. *British dental journal*, *180*(9), 339-343.

[3] Allegue-Magaz, S., Peñarrocha-Diago, M., Chamorro-Petronacci, C. M., Lorenzo-Pouso, A. I., Blanco-Carrión, A., García-García, A., & Pérez-Sayáns, M. (2020). Quality and readability of the dental information obtained by patients on Internet: criteria for selecting medical consultation websites. *Medicina oral, patologia oral y cirugia bucal*, *25*(4), e508.

[4] Arora, R., & Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data* (pp. 91-97).

[5] Auluck A. (2005). Oral health of poor people in rural areas of developing countries. *J Can Dent Assoc, 71*(10):753-755.

[6] Bastani, P., Mohammadpour, M., Mehraliain, G., Delavari, S., & Edirippulige, S. (2021). What makes inequality in the area of dental and oral health in developing countries? A scoping review. *Cost Effectiveness and Resource Allocation*, *19*(1), 1-12.

[7] Benadof, D., Hajishengallis, E., Cole, A., & Vidal, C. (2017). Oral literacy demand in the pediatric dental clinic: a pilot study. *International journal of pediatric dentistry*, *27*(5), 326-333.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research*, *3*, 993-1022.

[9] Capeletti, L. R., Franco, A., Reges, R. V., & Silva, R. F. (2017). Intra-alveolar morphology assessed in empty dental sockets of teeth missing postmortem. *Forensic science international*, *277*, 161-165.

[10] Chhabra, K. G., Mulla, S. H., Deolia, S. G., Chhabra, C., Singh, J., & Marwaha, B. S. (2016). Dental informatics in India: time to embrace the change. *Journal of clinical and diagnostic research: JCDR*, *10*(3), ZE12.

[11] Gilchrist, F., Marshman, Z., Deery, C., & Rodd, H. D. (2015). The impact of dental caries on children and young people: what they have to say? *International journal of paediatric dentistry*, *25*(5), 327-338.

[12] Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, *23*, 856-864.

[13] Hou, A. Y., & Barnard, P. D. (1994). Dental informatics for Australia. *Australian dental journal*, *39*(5), 321-323.

[14] Hu, J., Luo, E., Song, E., Xu, X., Tan, H., Zhao, Y., ... & Li, Z. (2009). Patients' attitudes towards online dental information and a web-based virtual reality program for clinical dentistry: A pilot investigation in China. *International journal of medical informatics*, *78*(3), 208-215.

[15] Jampani, N. D., Nutalapati, R., Dontula, B. S. K., & Boyapati, R. (2011). Applications of teledentistry: A literature review and update. *Journal of International Society of Preventive & Community Dentistry*, *1*(2), 37.

[16] Murshid, E. Z. (2015). Dental knowledge of educators and healthcare providers working with children with autism spectrum disorders. *Saudi medical journal*, *36*(12), 1477.

[17] Pack, A. R. (1998). Dental services and needs in developing countries. *International dental journal*, *48*(S3), 239-247.

[18] Qi, S., Yan, Y., Luo, E., & Hu, J. (2015). The development of dental informatics and dental information technology in China: A systematic study. *Journal of Dental Sciences*, *10*(2), 176-184.

[19] Schleyer, T., & Spallek, H. (2001). Dental informatics: a cornerstone of dental practice. *The Journal of the American Dental Association*, *132*(5), 605-613.

[20] Schuller-Götzburg, P., & Suchanek, J. (2007). Forensic odontologists successfully identify tsunami victims in Phuket, Thailand. *Forensic science international*, *171*(2-3), 204-207.

[21] Silva, R. F., Franco, A., de Souza, J. B., Picoli, F. F., Mendes, S. D. S. C., & Nunes, F. G. (2015). Human identification through the analysis of smile photographs. *The*

American journal of forensic medicine and pathology, *36*(2), 71-74.

[22] Singaraju, S., Prasad, H., & Singaraju, M. (2012). Evolution of dental informatics as a major research tool in oral pathology. *Journal of oral and maxillofacial pathology: JOMFP*, *16*(1), 83.

[23] Sittig, D. F., Kirshner, M., & Maupome, G. (2003). Grand challenges in dental informatics. *Advances in dental research*, *17*(1), 16-19.

[24] Song, M., Spallek, H., Polk, D., Schleyer, T., & Wali, T. (2010). How information systems should support the information needs of general dentists in clinical settings: suggestions from a qualitative study. *BMC medical informatics and decision making*, *10*(1), 1-9.

[25] StackPath. (n.d.). Retrieved October 31, 2021, from https://www.dentaleconomics.com/macro-op-ed/technology-trends/article/14185481/datadriven-dentistry-a-new-perspective.

[26] Tohnak, S., Mehnert, A. J. H., Mahoney, M., & Crozier, S. (2007). Synthesizing dental radiographs for human identification. *Journal of dental research*, *86*(11), 1057-1062.

[27] Verploegen, V. J. N., & Schuller, A. A. (2019). Erosive tooth wear: knowledge among young adults and their preferred information sources. *International journal of dental hygiene*, *17*(1), 85-92.

[28] Westerman, B. (1993). A preventive dental care programme at the workplace. *Australian dental journal*, *38*(3), 210-215.

[29] Zylan, R. (2020). Everyone has a'voice'in achieving better outcomes in oral health for people with an intellectual disability. *Intellectual Disability Australasia*, *41*(1), 5-6.