pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220989

Structuring Research Directions in Medical Domain with Topic Modeling: Application for PhD Theses Synopses in Dentistry

Igor BABIKOV[a1], Sergey KOVALCHUK[a], Ivan SOLDATOV[b], Gennady GREBNEV[b]
[a]*ITMO University, Saint Petersburg, Russia*
[b]*Kirov Military Medical Academy, Saint Petersburg, Russia*

Abstract. The study analyzed scientific texts based on a manually created database of synopses of theses in dentistry. The main goal was to structure medical texts into various topics by means of natural language processing techniques (topic modeling). Furthermore, a dynamic topic modeling showed the most popular in the field of dentistry over almost the last thirty years.

Keywords. Machine Learning, Natural Language Processing, Topic Modeling, ARTM, Unstructured Data, Dentistry

Introduction

Lately many scientific spheres have started to use artificial intelligence and machine learning extensively including health care and dentistry. When machine learning practitioners work with texts, it is mostly about unstructured data.

As part of a larger ongoing study of structuring textual data, this research paper presents an idea of systemizing knowledge from unstructured medical texts by finding specialized topics in a carefully created corpus of dentistry synopses of theses.

1. Dataset Description

The dataset is a collection of 5779 Russian synopses of theses in the scientific specialty of dentistry from 1993 to 2020. Each one comprises the objectives and the conclusion for an actual thesis and the average text length is 209 words. It worth mentioning that each text has the corresponding dentistry specialty code either 14.01.14 (14.00.21) according to Higher Attestation Commission.

The distribution of the total number of theses synopses over the years is presented in Figure 1. We may conclude there was a rapid decline between 2014 and 2015 in terms a total number of theses defences due to the changes in Higher Attestation Commission.

[1] Corresponding Author. Igor Babikov, ITMO University Saint Petersburg, Saint Petersburg, Russia; Email:igorbabikov24@gmail.com

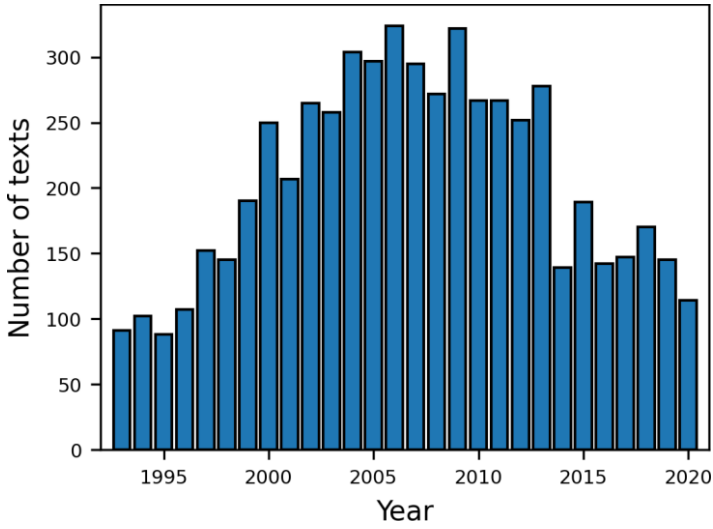

Figure 1. Distribution of documents across the years.

2. Understanding Dentistry Texts with Topic Modeling

2.1. Common Topic Modeling Techniques Application

The main task was to structure dentistry texts, especially distinguish different nosological forms and matching them with the mentioned Higher Attestation Commission "Dentistry" research area items. In the study, we tried various topic modeling options:

- Top2Vec [1]
- LDA [2]
- BERTopic [3]
- ARTM [4]

We tested each model against 6, 10 and 13 topics. These numbers were set via trial-and-error method. The number 6 was selected, since there are six available items in the "Dentistry" research area. We also opted for 10 and 13, since we had a hypothesis that there could be more meaning separation for nosological forms within the dataset. From the expert point of view, the most interpretable model was ARTM with six topics, which the results are provided in Table 1. The keywords in each topic in the table arehighlighted in bold.Second paragraph.

Topic 1 describes the issues of prevention of major dental diseases, high prevalence and intensity rate of tooth caries among children in particular as well as quality healthcare delivery to them.

Topic 2 addresses the general problems of maxillofacial surgery and dental implantation in various medical cases in particular.

Topic 3 illustrates the issues of dental orthopedics, dental prosthetic rehabilitation with arch deficiency.

The terms selection in topic 4 matches the problems of preventive therapeutic dental care dedicated to etiology, pathogenesis, epidemiology, preventive methods, diagnostics and treatment of periodontium diseases.

Topic 5 relates to the general problems of maxillofacial surgery and specifically jaw fractures in different cases.

Topic 6 as well as topic 4 also targets the issues of preventive therapeutic dental care. However, topic 6 particularly focuses on the chronic generalized periodontitis, its features, the development and course of the disease and its complications among distinct categories of population.

Table 1 The topics produced by ARTM

№	Topic Terms	Topic Name
1	***Стоматологический (dental), ребенок (child), кариес (tooth caries), помощь (help), распространенность (prevalence)***, зуб(tooth), пациент (patient), уровень (level), лечение (treatment), заболевание (disease), группа (group), рот (mouth), высокий (high),показатель (index), качество (quality), население (population), интенсивность (intensity)	Стоматологическая помощь детям (Child disease prevention)
2	***Костный (skeletal), имплант (implant), дефект (deficiency), челюсть (jaw),*** материал (material), метод (method), лечение (treatment), разработать (develop), пациент (patient), дефект (deficiency), исследование (research), результат (outcome)	Имплантация зубов (Dental implantation)
3	***Протез (denture), жевательный (masticatory), конструкция (structure), зубной (dental), дефект (deficiency),*** пациент (patient),лечение (treatment), группа (group), мышца (muscle), нижний (lower), ткань (tissue), показатель (index), функциональный (functional), материал (material)	Зубопротезирование (Dentofacial orthopedics)
4	***Пародонт (periodontium),ткань (tissue), воспалительный (inflammatory), хронический (chronic),*** лечение (treatment), зуб(tooth), применение (application), терапия (therapy), группа (group), слизистый (mucosal), оболочка (capping), хронический (chronic), пациент (patient)	Лечение заболеваний пародонта (Periodontiumdiseases treatment)
5	***Челюсть (jaw), нижний (lower), перелом (fracture), угол (corner), положение (location), верхний (upper),*** костный (skeletal), группа (group), метод (method), лицо (face), канал (channel), ряд (arch), случай (occasion), окклюзия (occlusion)	Переломы челюстей(Jaw fractures)
6	***Хронический (chronic), пародонтит (periodontitis), степень (extent), воспалительный (inflammatory),*** пациент (patient), заболевание (disease), ткань (tissue), уровень (level), активность (activity), изменение (change), тяжесть (severity), группа (group), процесс (process)	Лечение хронического генерализованного пародонтита (Chronic generalized periodontitis)

As with any topic modeling technique, there are general words within each topic, such as “treatment”, “patient”, “index” or “disease”, which cannot help provide a determined topic description.

The experiments with 10 and 13 topics gave similar results, though this higher number of topics could be generalized into six topics presented in Table 1.

Top2Vec and BERTopic are one of the state-of-the-art topic models, though they showed virtually similar results. We believe that the modern approaches such as

Top2Vec or BERTopic do not work with this scarce amount of data. Furthermore, these models take considerably longer time to train.

LDA did not perform well either producing uninformative topics.

2.2. Topic Visualization

In order to visually understand how research directions interconnect or differ we applied a dimensionality reduction technique with the UMAP [5] algorithm.

We considered the probability topics of ARTM as our features for each text document and reduced the dimensionality from the initial number of topics down to two. Figure 2 represents the UMAP 2D projection of the best ARTM instance with six topics (refer to Table 1 for the topic numbers).

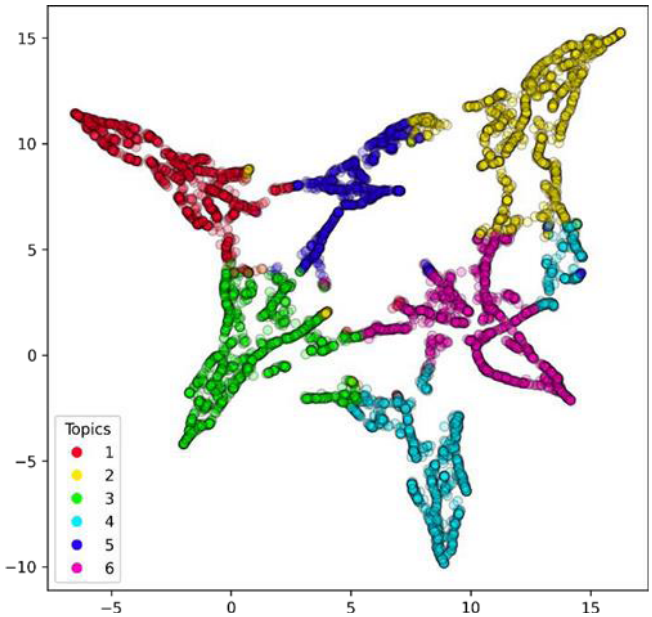

Figure 2. UMAP projection of the ARTM topics.

Considering Figure 2, there are several points that we could highlight. Firstly, in our dataset of dental theses synopses the topic of child dental health does not intersect with the works about periodontium diseases. Secondly, the theses which focused on maxillo-facial area injuries only partially or implicitly were linked to the topics of teeth implantation or dentofacial orthopedics. As a general fact, the topic "Chronic generalized periodontitis" (Table 1) is a sub-topic of treatments of various periodontium diseases. This is also shown in Figure 2.

2.3. Dynamic Topic Modeling

During the research, we also analyzed the chronological dynamic of the most popular topics within the dataset (Figure 3). For visualization purposes, Figure 4 shows the maximal topics for each year (refer to Table 1 for the topic numbers). We could see that

before 2010 the dominant topics for the dentistry theses were “Dental implantation”, “Periodontium diseases treatment” and “Chronic generalized periodontitis”.

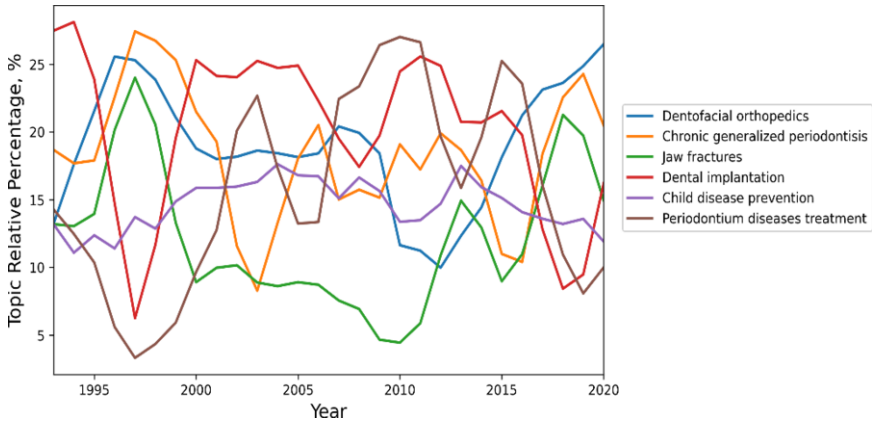

Figure 3. Chronological Dynamic of the most popular ARTM topics.

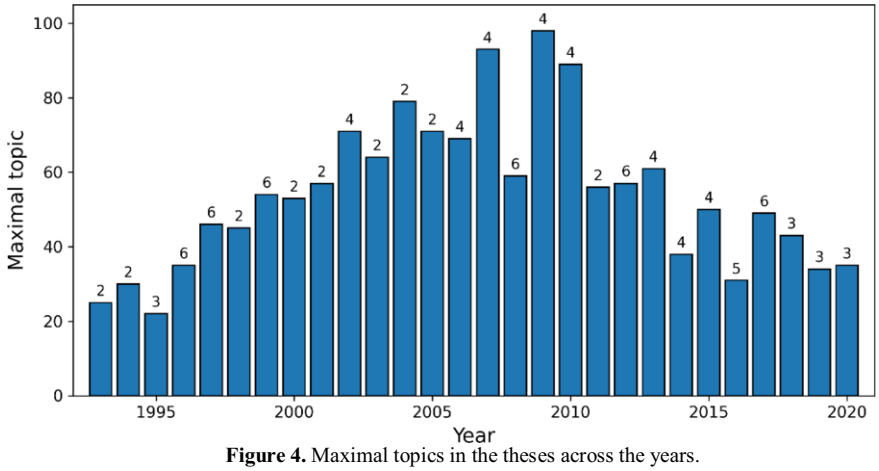

Figure 4. Maximal topics in the theses across the years.

Another important observation could be that before 2010 the research area “Jaw fractures” was underrepresented but since 2014 it has been receiving more popularity in scientific community. These results could propose that the changes in Higher Attestation Commission scientific specialties classification in 2020 had this area moved into a new specialty 3.1.2 “Maxillofacial Surgery” in 2022.

3. Discussion and Future Works

The results of the study infer a number of discussion items regarding data science and machine learning, scientometrics as well as dentistry research topics structuring.

In spite of the fact that the study focused only on a narrow medical domain, dentistry in particular, the machine learning techniques for textual data analysis used showed promising results. We managed to produce appropriate topics, which the subject matter

expert could interpret and match them with the corresponding dental scientific specialties classification. There are not many machine learning production-ready solutions for this and other medical domains. One can consider the accomplished analysis as a perspective future data science tool for processing different existing medical domain text corpora; including such common natural language processing tasks as topic modeling and evaluating interconnections of produced topics or assessing topic popularity throughout a given period. We believe that the methodology and practical applications of the techniques used in the study are not limited to the dentistry domain and can be applied to obtaining a dynamic structure of knowledge from textual data in various areas. As a result, our pipeline provides research and educational personnel with a tool for searching scientific knowledge in selected medical areas and correctly categorize between similar scientific research directions.

In the view of scientometrics, it is important to determine main research areas in a certain scientific domain. It is also significant to predict which research area is going to be dominant in the upcoming decades. Within this paper, we managed to discover certain structure and features of the dynamic in the scientific studies of the dentistry domain by means of the machine learning techniques.

A significant validation fact of our study is reflecting the upcoming changes in the dentistry specialty code from 14.01.14 to 14.00.21 by embedding a new specialty "Maxillofacial Surgery" and expanding knowledge domain according to the new specialties classification introduced by Higher Attestation Commission responsible for thesis defense in Russian Federation.

4. Conclusion

In the study, we carried out the analysis of Russian dentistry synopses of theses with topic modeling. One of the models, ARTM, gave us the most interpretable topics that we could treat as various nosological units. In addition, we performed a dynamic topic modeling to observe which topics were the most popular throughout the last thirty years in the scientific area of dentistry and predict which areas will likely be prevalent in the years to come.

As part of further research, we aim to perform a more detailed analysis of various models' outputs (resulting topics and their probabilities for documents) as well as topic interpretation. Furthermore, as of machine learning, we plan to build a pipeline that could automatically classify scientific knowledge of a given subject.

References

[1] Angelov D. Top2Vec: Distributed Representations of Topics. arXiv: 2008.09470 (2020). doi: 10.48550/ARXIV.2008.09470.

[2] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003; 3: 993-1022.

[3] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv: 2203.05794 (2022). doi: 10.48550/ARXIV.2203.05794.

[4] Vorontsov K. Additive Regularization for Topic Models of Text Collections. Dokl. Math. 2014; 89: 301–304. doi: 10.1134/S1064562414020185.

[5] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426 (2018). doi: 10.48550/ARXIV.1802.03426.