УДК 004.912

ПОИСК НАУЧНЫХ НАПРАВЛЕНИЙ В МЕДИЦИНЕ С ПОМОЩЬЮ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Бабиков И.А (Университет ИТМО, г. Санкт-Петербург), **Солдатов И.К.** (Военно-медицинская академия имени С. М. Кирова, г. Санкт-Петербург)
Научный руководитель – к.т.н., Ковальчук С.В.
(Университет ИТМО, г. Санкт-Петербург)

Введение. В различных научных сферах, в частности стоматологии, приходится работать с большим количеством неструктурированных текстовых документов. Основная задача данной работы состоит в том, чтобы систематизировать знания из неструктурированных медицинских текстов путем нахождения узконаправленных тематик на основе специализированного стоматологического корпуса.

Основная часть. С целью выделения научных тематик в стоматологии из текстов были использованы следующие методы (и фреймворки) тематического моделирования; каждая модель была протестирована на 6, 10 и 13 темах.:

1) LDA (Gensim);
2) ARTM (BigARTM);
3) Top2Vec;
4) BERTopic.

Набор данных состоял из текстовых авторефератов диссертаций по специальности «Стоматология» на русском языке, собранных с 1993 по 2020 гг. С точки зрения эксперта-стоматолога (Военно-медицинская академия имени С. М. Кирова), наиболее интерпретируемые темы произвела модель ARTM с 6 темами:

1) стоматологическая помощь детям;
2) имплантация зубов;
3) зубопротезирование;
4) лечение заболеваний пародонта;
5) переломы челюстей;
6) лечение хронического генерализованного пародонтита.

Современные модели Top2Vec и BERTopic произвели схожие темы, но их обучение заняло значительно большее время. Модель LDA не смогла справиться с задачей: эксперт не смог разделить темы на разные нозологические формы.

Заметим, что число защит диссертаций на темы по лечению заболеваний пародонта (и генеразилованного пародонтита) всегда оставалось на высоком уровне (90 в 2017 г., 98 в 2009 г., 85 в 2010 г.), в то время как после 2014 года защит на тему «Имплантация зубов» стало значительно меньше.

Для визуализации того, каким образом темы исследований в стоматологических диссертациях пересекаются или отличаются друг от друга, был применен метод сокращения размерности алгоритмом UMAP. Темы тематических моделей рассматривались в качестве признаков для каждого документа и затем было произведено сокращение размерность от исходного количества тем до двух. В двумерном пространстве кластеры оказались не с шарообразной формой, а с вытянутой, что может сигнализировать о взаимосвязи кластеров между собой.

Построение хронологической динамики стоматологических тематик за последние десятилетия позволяет предсказать популярность научных направлений, что, например, может помочь авторам при выборе темы диссертации. Относительные мощности каждой темы не превышали 28% в каждом году. Темы по лечению заболеваний пародонта и лечение хронического пародонтита имеют обратный характер по их «популярности»: например, в 1997 г. защиты по теме «Лечение хронического генерализованного пардонтита» занимали 27% среди всех защит, а защиты по теме «Лечение заболеваний пародонта» составляли лишь 3%; обратную картину можно пронаблюдать за такие годы, как 2003, 2016 или 2019-2020 гг.

Динамическое тематическое моделирование также отразило изменения по специальности «Стоматология», внесенные Высшей Аттестационной Комиссией. Возрастающий интерес к теме «Переломы челюстей» привел к тому, что в результате изменений Высшей Аттестационной Комиссии в 2020 г. она перешла в отдельный пункт области исследований специальности 3.1.2 «Челюстно-лицевая хирургия», которая официально была добавлена в классификацию ВАК 2022 г.

Выводы. В исследовании проведен провалидированный экспертом-стоматологом анализ русскоязычных текстов авторефератов диссертаций по стоматологии с помощью тематического моделирования. Полученные результаты могут быть применены для информационно-поисковой поддержки в научных работах по стоматологии, а также позволяют автоматически разделить смежные направления исследования.

Несмотря на то, что в данном исследовании акцент был сделан на конкретной медицинской специальности, методология может быть перенесена и на другие дисциплины.

Список использованных источников:

1. Angelov D. “Top2Vec: Distributed Representations of Topics.” arXiv, 2020. doi: 10.48550/ARXIV.2008.09470.
2. Blei D. M., Ng A. Y., Jordan M. I. “Latent Dirichlet Allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
3. Grootendorst M. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” arXiv, 2022. doi: 10.48550/ARXIV.2203.05794.
4. Vorontsov K. “Additive Regularization for Topic Models of Text Collections,” Dokl. Math., vol. 89, pp. 301–304, 2013, doi: 10.1134/S1064562414020185.
5. McInnes L., Healy J., Melville J. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” arXiv, 2018. doi: 10.48550/ARXIV.1802.03426.