Министерство связи и информатизации Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.3.049.77

НОВИКОВ
Алексей Алексеевич

МЕТОДЫ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ПРИ РАЗРАБОТКЕ ПРОЕКТНЫХ РЕШЕНИЙ

Диссертация
на соискание степени магистра
по специальности 1-45 80 01 «Системы и сети инфокоммуникаций»

Научный руководитель
кандидат технических наук, доцент
Стержанов Максим Валерьевич

_____
(подпись научного руководителя)

Минск 2024

ВВЕДЕНИЕ

В настоящее время, благодаря глобальному распространению сети Интернет большинство пользователей получили возможность в невиданных ранее объёмах получать и генерировать информацию. Это стало не только благом, но и породило ряд проблем, связанных со сбором, анализом и систематизацией информации. Аналитики, осуществляющие свою деятельность в различных областях, вынуждены использовать соответствующие этим условиям новые подходы к сбору, анализу и систематизации интересующих данных.

Для описания нового свойства таких данных, которые отличаются большим объёмом, высокой скоростью роста и огромным многообразием своих форм, был введён термин «большие данные» (big data). Феномен «больших данных» сформировал потребность в новых методах обработки и анализа, способных извлекать из этих, как правило, неструктурированных данных полезное «зерно», знание. Совокупность таких методов обозначается термином «интеллектуальный анализ данных» (data mining).

Из этой совокупности выделяется подмножество, специализирующееся на анализе текстовых данных - «интеллектуальный анализ текстовых данных» (text mining). Кроме того, выделяется группа методов, которые выполняют задачу тематического моделирования - построения статистических моделей, определяющих тематическую принадлежность каждого документа из корпуса.

Тематическое моделирование – это одна из современных технологий обработки естественного языка (англ. Natural language processing, NLP), активно развивающаяся с конца 90-х годов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие слова образуют каждую тему. Тематическое моделирование не претендует на полноценное понимание естественного языка (англ. natural language understanding, NLU), однако выявление тематики можно считать определённым шагом в этом направлении.

Тематическое моделирование принято относить к машинному обучению без учителя, поскольку темы строятся автоматически, но текстовым данным. Дня этого не требуется ни разметки, ни словарей, ни баз экспертных знаний. Существуют продвинутые тематические модели, способные учитывать такого рода данные дня улучшения тем и решения трудных задач текстовой аналитики. Такие модели тоже рассматриваются в данной книге.

Тематическая модель, как и нейросетевая, преобразует текст в векторное представление или эмбеддинг (embedding). Под эмбеддингом понимается возможность уменьшения размерности таких признаков ради повышения производительности модели. Прежде чем говорить о структурированных наборах

данных, полезно будет разобраться с тем, как обычно используются эмбеддинги.

Нейросетевые эмбеддинги не интерпретируемы, мы не понимаем смысла их координат. Тематический эмбеддинг – это вектор вероятностей тем. В каждой теме есть наиболее частотные слова, и если модель построена хорошо, то они оказываются связанными по смыслу. Глядя на них, можно сказать, о чём эта тема, составить её текстовое описание, дать ей название. Наиболее ценное свойство тематических моделей в том, что коллекция сама собой кластеризуется на интерпретируемые темы.

Тематическое моделирование похоже на кластеризацию документов. Отличие в том, что при обычной «жёсткой» кластеризации (hard clustering) документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет мягкую кластеризацию (soft clustering), распределяя содержимое документа по нескольким кластерам-темам. Тематическое моделирование называют также мягкой би-кластеризацией, поскольку каждое слово также распределяется по темам.

Вероятностная тематическая модель (англ. Probabilistic topic model, PTM) определяет вероятности тем в каждом документе и вероятности слов в каждой теме. Такие модели предсказывают вероятности появления слов в документах, но делают это не настолько хорошо, как глубокие нейронные сети типа BERT или GPT-3, но и намного проще и обладают свойством интерпретируемости.

Вероятностные тематические модели находят множество применений. Это выявление трендов в новостных потоках, патентных базах, научных публикациях, многоязычный информационный поиск, классификация и категоризация документов, тематическая сегментация текстов, суммаризация текстов, поиск тематических сообществ в социальных сетях, тегирование веб-страниц, обнаружение текстового спама. В том числе, тематическое моделирование может быть успешно использовано для решения социологических задач. С их помощью, например, можно определить разницу в освещении одних и тех же событий и фактов национальными и иностранными средствами массовой информации с использованием глобальной компьютерной сети Интернет (далее – интернет-СМИ) или увидеть, как менялось отношение интернет-СМИ к конкретному событию, объекту, субъекту и т.д., в определенный временной интервал.

Темой данного исследования является изучение разнообразия методов тематического моделирования при разработке проектных решений, перенос алгоритмов некоторых методов на мультипарадигменный язык программирования Python. В качестве примера показывается последовательность шагов, не-

обходимых для предварительной обработки данных, для построения тематического профиля национальных интернет-СМИ. Обосновывается выбор количества тем для построения модели тематического моделирования.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами (проектами) и темами

Основные положения магистерской диссертации разрабатывались в соответствии с:

Государственной программой «Цифровое развитие Беларуси» на 2021–2025 годы, утвержденной Постановлением Совета Министров Республики Беларусь 2 февраля 2021 г. № 66.

Законом Республики Беларусь от 28.12.2009 N 113-З «Об электронном документе и электронной цифровой подписи».

Законом Республики Беларусь от 10.11.2008 № 455-З «Об информации, информатизации и защите информации».

Цель и задачи исследования

Целью данной работы является изучение основных алгоритмов тематического моделирования при разработке проектных решений на примере автоматического исследования тематики сообщений в социальной сети Twitter в заданных хронологических рамках.

Основными задачами исследования являются:

анализ основных алгоритмов тематического моделирования;

обоснование выбора алгоритмов тематического моделирования в экспериментальной части исследования;

разработка типовой универсальной пошаговой модели исследования интернет-СМИ при помощи алгоритмов тематического моделирования.

Личный вклад магистранта

проведен системный анализ разнообразия методов тематического моделирования при разработке проектных решений, перенос алгоритмов некоторых методов на мультипарадигменный язык программирования Python;

проведен эксперимент по построению тематической модели корпуса сообщений пользователей сети Twitter;

разработана модель и программа оптимальной маршрутизации обмена данными подсистемы «Умный дом».

Опубликование результатов диссертации

По результатам исследований, представленных в диссертации, опубликовано 3 печатные работы, в том числе: 1 статья в научном журнале рекомендованных ВАК, общим объемом 2 авторских листа; 2 тезисов в сборниках и материалах конференций. По результатам исследований, представленных в диссертации.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения и библиографического списка.

Общий объем диссертационной работы 69 страниц, из них 55 основного текста, 15 рисунков и библиографический список из 45 наименований (в том числе 3 авторские публикации).

Проверка на уникальность

Проведена экспертиза диссертации Новикова Алексея Алексеевича «Методы тематического моделирования при разработке проектных решений» на корректность использования заимствованных материалов с применением сетевого ресурса «Проверка на уникальность» (адрес доступа: https://content-watch.ru/) в on-line режиме 06.06.2024 г. В результате проверки установлена корректность использования заимствованных материалов (оригинальность диссертационной работы составляет 82,4 %).

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе проведен анализ международного и отечественного опыта в области использования методов тематического моделирования, а также приведены области применения тематического моделирования.

Для описания нового свойства данных, которые отличаются большим объёмом, высокой скоростью роста и огромным многообразием своих форм, был введён термин «большие данные» (big data). Феномен «больших данных» сформировал потребность в новых методах обработки и анализа, способных извлекать из этих, как правило, неструктурированных данных полезное «зерно», знание. Совокупность таких методов обозначается термином «интеллектуальный анализ данных» (data mining).

Из этой совокупности выделяется подмножество, специализирующееся на анализе текстовых данных - «интеллектуальный анализ текстовых данных»

(text mining). Кроме того, выделяется группа методов, которые выполняют задачу тематического моделирования - построения статистических моделей, определяющих тематическую принадлежность каждого документа из корпуса.

Поскольку тематическое моделирование - многофункциональный, развивающийся метод, в настоящее время он используется во многих областях: от обработки естественного языка до научной литературы, программной инженерии, биоинформатики, гуманитарных наук и так далее. Тематическое моделирование находит широкое применение в решении задач информационного поиска, автоматического аннотирования и индексирования документов, выявления паттернов поведения, обнаружения и отслеживания событий в новостных потоках, построения профилей интересов пользователей в рекомендательных системах и т.д.

Во второй главе рассмотрены базовые понятия тематического моделирования, определены задачи и основные алгоритмы тематического моделирования, приведен обзор алгоритмов BERT и LDA, а также описания процесс оценки качества тематических моделей.

Основные понятия для тематического моделирования, которые мы используем в данной работе. Определим эти понятия следующим образом:

1 *Терм* – основная единица дискретных данных. В роли термов могут выступать исходные слова, леммы, словосочетания.

2 *Документ* – это последовательность из N слов, обозначаемых как $w = (w_1, w_2, ..., w_n)$, где w_n – это n-ое слово в последовательности.

3 *Корпус* – это набор документов, обозначенных как $D = (w_1, w_2, ..., w_M)$.

Цель тематического моделирования – выявить эти скрытые переменные – темы, которые определяют смысл нашего документа и корпуса.

В настоящее время существует несколько методов тематического моделирования, которые в свою очередь делятся на алгебраические и вероятностные. Подходы к тематическому моделированию представляют собой разновидность методов машинного обучения без учителя, поскольку темы и параметры смеси неизвестны и выводятся исключительно на основе данных.

Один из наиболее распространенных методов построения тематических моделей - метод латентного размещения LDA (Latent Dirichlet Allocation). Алгоритм основывается на распределении Дирихле и модели bag-of-words. Данный метод осуществляет мягкую кластеризацию. Предполагается, что слова, содержащиеся в документе, порождены латентной темой, которая определяется вероятностным распределением на множестве всех слов текста. LDA позволяет работать с корпусами большого размера.

Другой фундаментальный метод тематического моделирования – это pLSA. Модель pLSA - это вероятностная модификация модели LSA. Алгоритм

работы в модели pLSA основан на методе Байеса. Также в данной модели учитывается многозначность слов.

BERT (Bidirectional Encoder Representations from Transformers) – это двунаправленная многоязычная модель. Данная модель представляет собой двунаправленную нейронную сеть, осуществляющую контекстно-зависимую обработку и обученную на архитектуре Transformer.

В третьей главе представлен эксперимент по построению тематической модели корпуса сообщений пользователей сети Twitter, описана предварительная обработка корпуса, построена тематическая модель корпуса, проведен анализ результатов эксперимента.

Результаты работы алгоритма LDA представлены на рисунке 1.

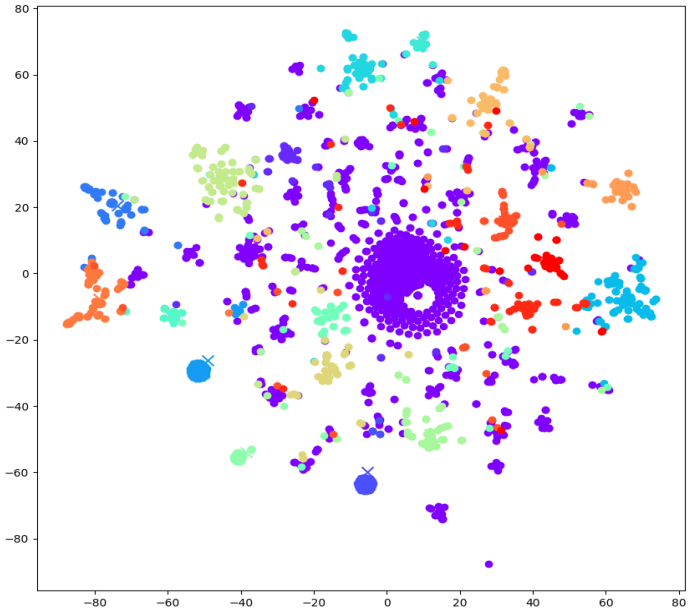

Рисунок 1 – Результат кластеризации с помощью алгоритма LDA

Обратившись к визуализированным кластерам, можно сказать, что алгоритм действительно четко выделяет определенные темы в корпусе. Однако самый многочисленный кластер находится среди других кластеров и на относительно небольшом расстоянии от них. Вероятно, это выбросы в виде шума в корпусе, которые алгоритм не смог отнести ни к одной из более узких тем. На изображении мы можем увидеть, что кластеры, выделенные с помощью метода К-средних, образуют достаточно связные и чаще всего пространственно разделенные между собой множества. Однако, здесь же мы видим, что несмотря на выделенные темы везде прослеживаются вкрапления одной и той же более обширной темы. В нашем случае это тема “пандемия COVID-19”.

Несмотря на то, что такие слова, как “коронавирус”, ”ковид” и т.д. были отсортированы на этапе токенизации корпуса, частотными являются такие слова, как названия симптомов, слова, связанные с темой вакцинации, наименования членов семьи и т.д. Эти слова не составляют отдельную тему, так как появляются в большей части пользовательских сообщений на разные более частные темы. Ключевые слова для каждой темы приведены в Приложении Г. На основании полученных ключевых слов мы можем выделить следующие темы:

– Ежедневные новостные сводки по заболевшим на территории России/Санкт-Петербурга/Европы;

– Вакцинирование в России;

– Новый штамм вируса COVID-19 и исследования в Беларуси;

– Ограничение авиасообщения между Россией и Турцией;

– Тестирование на предмет вируса COVID-19.

Темы распределены достаточно четко, однако, некоторые из них практически не поддаются интерпретации, некоторые из них, очевидно, содержат ключевые слова, относящиеся к другому кластеру. Вероятно, это связано с тем, что алгоритм не учитывает контекст слов, что является помехой для составления качественной тематической модели на таком неоднородном корпусе текстов представленным на рисунке 2.

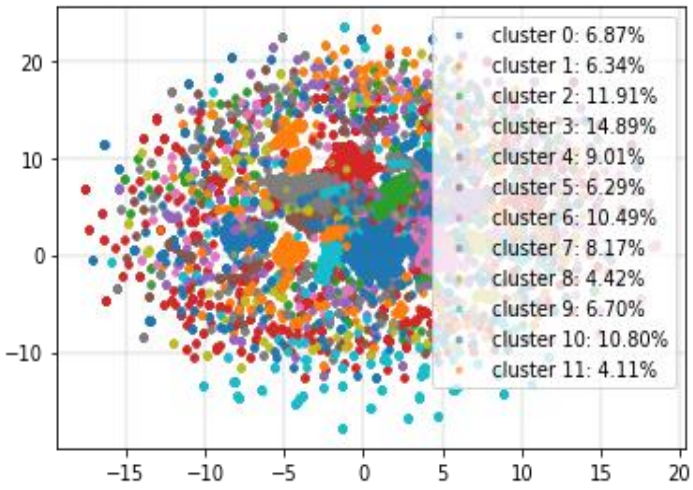

Рисунок 2 – Результат кластеризации с помощью алгоритма LDA+BERT

Кластеры, полученные с помощью алгоритма LDA+BERT выделены отчетливо, однако, мы можем видеть, что в большинстве случаев они пересекаются. Это связано с тем, что, как уже было описано выше, корпус очень неоднородный с большим количеством шума, а также многие сообщения не относятся конкретно к одной теме, а могут охватывать сразу несколько тем/собы-

тий. Тем не менее, абсолютно все документы были распределены к интерпретируемым темам. Ключевые слова для каждой темы приведены в Приложении Д. Основываясь на них, можно выделить основные темы:

– Туризм в условиях пандемии COVID-19 (11,24 %);

– Ежедневная статистика о заболевших вирусом COVID-19 (19,34 %);

– Вакцинация в разных странах (8,30 %);

– Новая волна вируса COVID-19 в России (4,27 %);

– События, связанные с COVID-19 в Беларуси (8,53 %);

– Международные новости (5,21 %);

– Симптоматика/тестирование на наличие вируса (8,62 %);

– События в Республике Беларусь, связанные, в том числе, с деятельностью Президента Республики Беларусь (6,81 %).

Также два кластера, выделенные моделью, содержат достаточно специфические ключевые слова, не составляющие одну конкретную тему. Скорее всего, к этим кластерам относятся сообщения на разобщенные/очень узкие темы, опечатки, а также слова, отсутствующие в словаре. На эти темы приходится 7,77 % и 7,40 % от общего массива кластеров.

Анализ полученных результатов показал, что в целом обе тематические модели достаточно успешно определяют темы и образуют видимые кластеры. Рассмотрим работу алгоритмов с помощью выбранных метрик и сравним полученные результаты.

Для оценки результатов обратимся к использованным нами метрикам: когерентность для оценки интерпретируемости тем и silhouette score для оценки качества полученных кластеров.

Полученные нами результаты по двум метрикам приведены в таблице 1.

Таблица 1 – Результаты работы алгоритмов тематического моделирования

	LDA	LDA+BERT
Когерентность	0,246290346799	0,45401289590625554
Silhouette Score	0,2065349931	0,2572047

Обращаясь к полученным данным, мы можем сделать вывод, что качество тематической модели улучшилось, но значение Silhouette Score по-прежнему низкое. Это может быть связано с тем, что многие темы пересекаются. Количество более узких, частных тем очень небольшое, что, в свою очередь, связано с размером корпуса, использованного для данной работы. Также к возможным причинам можно отнести выбросы в виде опечаток и неформального

стиля общения в социальных сетях, где многие токены при использовании моделью BERT метода WordPiece при лемматизации теряют свою семантическую ценность.

В четвертой главе определены основные этапы необходимые для построения тематического профиля национальных интернет-СМИ.

Основные этапы необходимые для построения тематического профиля национальных интернет-СМИ:

– сбор данных;
– предварительная обработка данных;
– токенизация;
– лемматизация;
– удаление стоп-слов;
– тематическое моделирование;
– определение оптимального количества тем;
– популярность тем у читателей.

ЗАКЛЮЧЕНИЕ

Таким образом, исследование множества вероятностных тематических моделей показало высокий потенциал разнообразных вариантов их применения в практической (прикладной) плоскости, в том числе при разработке проектных решений в различных областях науки. При этом тематические модели могут выступать универсальным базисом с гибкой структурой, адаптивной к конкретному проекту, либо задаче в различных областях (СМИ, социологические, маркетинговые и иные исследования).

В рамках выполнения экспериментальной части данное исследование показывает возможные варианты использования алгоритмов тематического моделирования с помощью алгоритма LDA и дополнительно полученных векторов с помощью модели BERT, обученной для русского языка на корпусе русскоязычных сообщений из социальной сети Twitter.

Было проведено два эксперимента по построению тематической модели:

1 Построение модели с помощью алгоритма LDA. Для улучшения базового алгоритма была проведена дополнительная фильтрация корпуса, эксперименты по подбору оптимального количества тем и предварительное обучение модели на собранном корпусе;

2 Построение тематической модели с конкатенированными векторами, полученными с помощью алгоритма LDA для информации о вероятностном распределении и векторами, полученными с помощью модели BERT для русского языка для информации о контексте слов и их распределении в документах.

Результат экспериментов оценивался с помощью метрик когерентности для оценки качеств полученных тем и silhouette score для оценки качества полученных кластеров.

Алгоритм LDA показал относительно низкие результаты, даже при дополнительных настройках. При максимальных значениях метрик 1, в данной работе этот результат не был достигнут. Мы получили интерпретируемые темы, отражающие некоторые из происходящих событий, но, рассматривая результаты кластеризации, информация из многих сообщений была извлечена неверно/не была учтена в исследовании. По самой распространенной теме в корпусе (“Ежедневные новостные сводки по заболевшим на территории Беларуси/Минске/Европы”) можно сделать вывод о том, что большая часть сообщений, обработанных алгоритмом - посты новостных ресурсов, не содержащие шум в виде неформальной/разговорной лексики/опечатки.

Второй эксперимент показал более высокие результаты, что доказывает, что гипотеза о том, что дополнительная информация о контексте слова, а именно дополнительно полученные векторы предложений, а не слов, помогает улучшить качество тематической модели и дает более интерпретируемые результаты и, следовательно, на выходе дает более качественные кластеры, о чём свидетельствуют увеличившиеся значения метрик качества.

Мы получили большее количество тем, что говорит о том, что модель справилась с обработкой тех документов, которые в первом эксперименте, либо относились к шуму, либо были ошибочно отнесены к тем или иным кластерам.

Самая распространенная из тем, согласно результатам кластеризации, также относится к новостным сводкам, что свидетельствует о снижении популярности социальной сети Twitter среди пользователей на территории Республики Беларусь, так как большая часть сообщений в настоящее время производится СМИ и аккаунтами тех или иный организаций, предоставляющих информацию о COVID-19. Несмотря на то, что оба алгоритма хорошо справились с выделением этих тем, в перспективе можно отказаться от исследования тем, произведенных подобными аккаунтами и сделать уклон в сторону исследования сообщений пользовательских аккаунтов, включающего также анализ тональности.

Для улучшения результатов в будущих исследованиях необходимо предпринять следующие шаги:

1 Совершенствование и расширение словаря стоп-слов. Сортировка и отбор наиболее частотных прилагательных/глаголов, слов разговорного стиля, опечаток, сдвоенных слов, лемм. Выделение биграмм/триграмм. Несмотря на

фильтрацию с помощью расширенного списка стоп-слов и TF-IDF, в числе самых частотных слов полученных тем присутствует шум в виде вышеперечисленных токенов.

2 Фильтрация сообщений на языках, не используемых в качестве объекта исследования. Библиотека Tweepy учитывает выбор языка при поиске и загрузке постов пользователей, однако, с учётом того, что все сообщения на монгольском языке написаны некоторыми пользователями на кириллице, они попадают в категорию русского языка.

3 Сбор корпуса большего размера с захватом более широкого временного промежутка. Это может повлиять на количество обсуждаемых тем/событий и дополнить уже существующие кластеры тем, таким образом, результат получится более репрезентативным. В настоящее время корпус собран в соответствии с наличием вычислительных мощностей.

4 Обучение модели на пользовательском корпусе. В данном эксперименте были использованы векторы из предварительно обученной модели.

5 Настройка оптимальных гиперпараметров/количества кластеров.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

А-1 Новиков, А.А. Использование методов тематического моделирования при разработке проектных решений / А.А. Новиков, М.В. Стержанов // Веснік сувязі. – №5. –2023. – С. 60 – 62.

А-2 Новиков, А. А. Использование методов тематического моделирования при разработке проектных решений / А. А. Новиков // Информационная безопасность : сборник материалов 59-й научной конференции аспирантов, магистрантов и студентов БГУИР, Минск, 17–21 апреля 2023 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2023. – С. 132–137.

А-3 Новиков, А. А. Тематическое моделирование на службе обеспечения безопасного веб-пространства / А. А. Новиков, К. А. Радкевич // Технические средства защиты информации : тезисы докладов XXI Белорусско-российской научно-технической конференции, Минск, 6 июня 2023 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Т. В. Борботько [и др.]. – Минск, 2023. – С. 66–67.