МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ИНФОРМАЦИОННЫХ РЕСУРСОВ, ОТРАЖАЮЩИХ ПОЛИТИЧЕСКУЮ АКТИВНОСТЬ МОЛОДЕЖИ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Гущина Никиты Александровича

Научный руководитель
профессор, доктор социологических наук _____ Н.И.Мельникова
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2024

ВВЕДЕНИЕ

Актуальность проблемы. Тематическое моделирование является важным инструментом в анализе текстовых данных, играющим значительную роль в различных областях, таких как информационный поиск, обработка естественного языка, социальные науки, маркетинг и бизнес-аналитика. В условиях роста объема текстовой информации в цифровом мире необходимость в автоматизированных методах анализа текстов становится все более актуальной. Тематическое моделирование позволяет эффективно выявлять скрытые структуры в текстовых данных, что делает его незаменимым инструментом для современных исследователей и аналитиков.

Современный мир характеризуется огромным объемом текстовой информации, поступающей из различных источников: социальных сетей, новостных сайтов, научных публикаций, блогов и т.д. Традиционные методы анализа не справляются с таким количеством данных, что делает тематическое моделирование актуальным инструментом для обработки и анализа больших массивов текстов. Оно позволяет автоматически классифицировать документы по темам, что упрощает процесс поиска и анализа информации.

Актуальность тематического моделирования в современном мире трудно переоценить. Этот метод предоставляет мощные инструменты для анализа и интерпретации текстовых данных, что делает его незаменимым для различных областей применения. С ростом объемов текстовой информации и необходимости в ее автоматизированной обработке и анализе, тематическое моделирование становится все более востребованным и важным инструментом в арсенале исследователей и аналитиков.

Степень разработанности темы. В исследовании «Эффективные реализации алгоритмов тематического моделирования» можем увидеть обзор эффективных алгоритмов вероятностного тематического моделирования больших текстовых коллекций. Рассматриваются алгоритмы обучения моделей латентного размещения Дирихле (Latent Dirichlet allocation,

LDA), и аддитивно регуляризованных тематических моделей для многопроцессорных систем.

В исследовании «Применение методов тематического моделирования для идентификации групп интернет-ресурсов с целью снижения риска киберугроз» рассматриваются существующие методы определения сетевых угроз с помощью анализа журналов прокси-сервера и предлагается метод кластеризации интернет-ресурсов, направленный на снижение объема входных данных путем исключения групп безопасных интернет-ресурсов или выбором только подозрительных интернет-ресурсов.

В исследовании «Применение тематического моделирования в глубоком обучении» рассматривается внедрение тематического моделирования в область глубокого обучения. Основное внимание уделяется пониманию концепций, таких как LDA) и различным применениям этих методов в контексте глубокого обучения. Особый акцент делается на анализе текстов, рекомендательных системах и классификации документов.

Цель исследования:

Провести сравнительный анализ тематических моделей научных публикаций, отражающих политическую активность молодежи.

Задачи исследования:

1. Осуществить анализ и подбор научных публикаций российских исследователей, исследующих политическую активность молодежи.

2. Выполнить тематическое моделирование выбранных научных публикаций.

3. Сделать сравнительный анализ результатов тематического моделирования.

Объект исследования. Метод тематического моделирования в текстовой аналитике для социальных наук.

Предмет исследования. Научные публикации российских исследователей на тему политической активности молодежи за период с 2020гг по 2024гг.

Теоретическая значимость ВКР. Исследование политической активности молодежи и ее отражения в научных публикациях является актуальной задачей, поскольку позволяет лучше понять механизмы формирования общественного мнения и влияния на политические процессы.

Практическая значимость ВКР. Полученные результаты могут быть полезны как для научного сообщества, так и для практических целей, в том числе для разработки соответствующих стратегий и программ в области образования и молодежной политики.

Структура ВКР. Структурно работа состоит из введения, трёх разделов, заключения, списка использованных источников и двух приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Теоретические основания тематического моделирования» рассматриваются основные понятия тематического моделирования, латентного размещение Дирихле и вероятностного латентно-семантического анализа.

Тематическое моделирование представляет собой методику анализа текстовых данных, направленную на выявление скрытых тематических структур в большом объеме документов. Оно позволяет автоматизировать процесс извлечения смысловых тем из текстов, что особенно полезно при анализе больших корпусов данных, таких как новостные статьи, научные публикации, сообщения в социальных сетях и другие текстовые источники. Тематическое моделирование помогает выявлять скрытые темы и структуры в текстовых данных, которые не очевидны при поверхностном чтении. Это особенно важно в контексте социальных исследований и анализа мнений, где выявление скрытых тем может привести к глубокому пониманию общественных настроений и трендов.

Метод LDA является одним из наиболее распространенных методов тематического моделирования. Он предполагает, что каждый документ представляет собой смесь нескольких тем, и каждая тема представляет собой

распределение слов. LDA использует методы байесовского вывода для оптимизации параметров модели. Процесс включает несколько ключевых этапов:

1. Инициализация параметров модели: На этом этапе случайным образом задаются начальные значения параметров, таких как распределения тем по документам и распределения слов по темам. Это начальное случайное распределение служит отправной точкой для дальнейшего обучения модели.

2. Обновление параметров: Используя методы байесовского вывода, такие как вариационные методы или метод Гиббсовой выборки, параметры модели обновляются на основе текущих распределений тем и слов. Вариационные методы включают обновление параметров апостериорных распределений, минимизируя расхождение между истинным апостериорным распределением и аппроксимирующим распределением. Метод Гиббсовой выборки, с другой стороны, выполняет пошаговую случайную выборку из условных распределений для каждого слова, обновляя распределения тем по документам и слов по темам.

3. Оценка правдоподобия: На этом этапе вычисляется правдоподобие наблюдаемых данных при текущих параметрах модели. Это позволяет оценить, насколько хорошо текущая модель объясняет данные и служит критерием для оценки сходимости модели.

4. Итерация до сходимости: Процесс обновления параметров и оценки правдоподобия повторяется до тех пор, пока модель не стабилизируется и изменения правдоподобия становятся незначительными.

LDA позволяет эффективно выявлять темы в больших объемах текстовых данных, предоставляя мощный инструмент для анализа и интерпретации текстов. Тематическое моделирование и тематическая классификация часто используются вместе, но решают разные задачи. Тематическое моделирование выявляет скрытые темы в текстовых данных без заранее заданных меток, тогда как тематическая классификация распределяет тексты по заранее определенным категориям.

Вероятностный латентно-семантический анализ (pLSA) является еще одним популярным методом тематического моделирования, который подобно LDA используется для выявления скрытых тем в текстовых данных. pLSA основывается на вероятностной модели, где каждое слово в документе генерируется одной из скрытых тем. Модель pLSA строит матрицу вероятностей, описывающую вероятность появления слов в темах и тем в документах.

Процесс pLSA включает несколько этапов:

1. Построение матрицы терминов-документов: На этом этапе создается матрица, представляющая частоту появления слов в документах. Каждый ряд этой матрицы соответствует отдельному документу, а каждый столбец — отдельному слову из словаря.

2. Обучение модели: Оптимизация параметров, описывающих вероятностное распределение слов в темах и тем в документах, осуществляется с помощью алгоритма EM (Expectation-Maximization). Этот процесс включает:

E-шаг (Expectation), вычисление вероятности того, что каждое слово в документе связано с каждой темой на основе текущих оценок параметров.

M-шаг (Maximization), обновление оценок параметров для максимизации правдоподобия наблюдаемых данных, используя вероятности, вычисленные на E-шаге.

3. Интерпретация результатов: На этапе интерпретации анализируются выявленные темы и их распределение по документам. Это включает анализ матрицы P(z|d), которая показывает, какие темы наиболее вероятны для каждого документа, и матрицы P(w|z), которая показывает, какие слова наиболее вероятны для каждой темы.

pLSA позволяет анализировать и классифицировать тексты на основе выявленных тем, предоставляя гибкий инструмент для работы с большими объемами текстовых данных.

Векторные модели текста преобразу.т текстовую информацию в числовые векторы, что позволяет применять математические и статистические методы для анализа текстов. Основные векторные модели включают:

1. Модель мешка слов (BoW): Представляет текст как вектор частот слов, не учитывая порядок слов и синтаксические связи. Для создания BoW-вектора составляется словарь всех уникальных слов в корпусе документов, а затем для каждого документа строится вектор на основе частоты появления слов из этого словаря.

2. TF-IDF (Частота Терминов – Обратная Частота Документов): Улучшает модель BoW, учитывая важность слов в документе и в корпусе. Вектор TF-IDF для слова получается умножением частоты термина (TF) на обратную частоту документа (IDF), что снижает вес часто встречающихся, но малоинформативных слов и увеличивает вес редких, но значимых слов.

3. Word2Vec: Обучает нейронную сеть для создания плотных векторов слов, отражающих семантические отношения между словами. Две основные архитектуры Word2Vec — Continuous Bag of Words (CBOW) и Skip-gram. CBOW предсказывает текущее слово на основе контекста, тогда как Skip-gram предсказывает контекстные слова на основе текущего.

4. GloVe (Глобальные Векторы для Представления Слов): Использует статистическую информацию о частотах совмещения слов для создания векторных представлений. GloVe моделирует отношения между словами через матрицу совмещения, где каждый элемент представляет частоту совместного появления пары слов.

5. Doc2Vec: Расширяет идею Word2Vec для создания векторных представлений для документов. Doc2Vec обучает вектора документов так, чтобы они могли предсказывать слова, содержащиеся в этих документах.

Во втором разделе «Подготовка исходных данных для тематического моделирования» описывается процесс подготовки данных, включающий несколько этапов: сбор публикаций, предварительная обработка данных, токенизация, удаление стоп-слов, приведение слов к начальной форме, создание словаря и корпуса документов.

Для сбора научных публикаций был использован сервис Google Академия. Основным критерием поиска была тема «Политическая активность молодежи в цифровой среде». Google Академия предоставляет доступ к различным источникам, что позволяет получить максимально полное представление о текущих исследованиях в выбранной области.

Отбор публикаций проводился по следующим критериям: тематика, период, авторы.

Тематика определялась вопросами политической активности молодежи в цифровой среде в публикациях, выпущенных в период с 2020 по 2024 годы российскими исследователями, чтобы обеспечило релевантность и актуальность данных в контексте российского общества.

Для тематического моделирования использовались только аннотации и ключевые слова, так как они содержат сжатую информацию о содержании публикаций и основных исследовательских вопросах.

На этапе очистки текстов были удалены все ненужные символы, такие как пунктуация, цифры и специальные символы, которые не несут смысловой нагрузки для тематического моделирования. Также были удалены дублирующиеся пробелы и знаки препинания. Очистка текстов является важным шагом, так как наличие лишних символов может негативно повлиять на результаты анализа.

Токенизация — это процесс разбиения текста на отдельные слова или токены. Этот процесс важен для дальнейшего анализа, так как тематическое моделирование работает с отдельными словами. Токены являются основными единицами анализа в большинстве алгоритмов обработки естественного языка.

Стоп-слова — это часто встречающиеся слова, которые обычно не несут значимой информации для анализа. Примеры стоп-слов включают: "и", "в", "на", "это" и т.д. Удаление стоп-слов позволяет сосредоточиться на более значимых терминах и улучшить качество тематического моделирования. Все стоп-слова были удалены из текстов, чтобы повысить качество тематического моделирования.

Лемматизация — это процесс приведения слов к их базовой или начальной форме (лемме). Например, слова "бегу", "бежал", "бежит" приводятся к лемме "бегать". Лемматизация помогает объединить различные формы одного и того же слова и улучшить качество анализа.

Стемминг — это процесс отсечения окончаний слов для получения основы слова. Например, слова "running", "runner" будут преобразованы в "run". В русском языке примером может быть преобразование слов "бежал", "бегающий" в основу "бег". Стемминг помогает уменьшить разнообразие словоформ и сосредоточиться на базовых корнях слов. Стемминг использовался в качестве дополнения к лемматизации для улучшения качества анализа.

Словарь был построен на основе всех уникальных слов, обнаруженных в текстах аннотаций и ключевых слов. Это позволило создать полный список терминов, используемых в публикациях, что является основой для дальнейшего тематического моделирования. Каждый термин в словаре был связан с уникальным идентификатором, что облегчает последующую обработку текстов.

Формирование корпуса документов включало преобразование каждого текста аннотации и ключевых слов в вектор токенов. Этот процесс был выполнен с использованием словаря, созданного на предыдущем этапе. Для векторизации текстов использовались методы, предоставляемые библиотекой Gensim, которые позволяют эффективно работать с большими объемами текстовых данных.

В третьем разделе «Проведение тематического моделирования» описывается ход проведения тематического моделирования для собранных данных и показан рабочий код.

Основные пакеты, использованные в работе, включают os, re, nltk, gensim и pyLDAvis. Библиотека os предоставляет функции для взаимодействия с операционной системой, re — для работы с регулярными выражениями, nltk — для обработки естественного языка, gensim — для тематического моделирования, а pyLDAvis — для визуализации результатов моделей LDA.

В данном коде используется библиотека stopwords из пакета nltk для работы со стоп-словами. После загрузки стоп-слов их можно импортировать и использовать для фильтрации текста, чтобы удалить ненужные слова и улучшить качество дальнейшего анализа.

Стемминг — это метод обработки естественного языка (NLP), который сводит слова к их базовой или корневой форме. В данном случае для работы с русским языком был использован SnowballStemmer из библиотеки nltk, который эффективен и подходит для этой задачи.

Gensim присваивает каждому слову в документе уникальный идентификатор. Полученный корпус представляет собой набор пар (word_id, word_frequency), где каждый элемент описывает частоту появления слова с конкретным идентификатором в документе. Этот корпус затем используется в качестве входных данных для построения модели LDA.

LDA модель строится на основе подготовленного корпуса документов. Процесс построения включает выбор количества тем, обучение модели и интерпретацию полученных тем. Основные этапы включают:

1. Выбор количества тем (k): Количество тем выбирается исходя из структуры и объема данных. Слишком большое количество тем может привести к дублированию и пересечению тем, а слишком малое — к недостаточной детализации.
2. Обучение модели: Применение алгоритма LDA для нахождения скрытых тем в тексте.

3. Интерпретация результатов: Анализ ключевых слов для каждой темы и их значимость.

После создания модели LDA следующим шагом является анализ полученных тем и связанных с ними ключевых слов. Для этого отлично подходит интерактивная визуализация с помощью библиотеки pyLDAvis.

Каждый пузырь на диаграмме pyLDAvis представляет отдельную тему. Чем больше пузырь, тем более распространена эта тема в наборе данных. При наведении курсора на пузырь справа отображаются ключевые слова, формирующие эту тему.

Для проверки тем на устойчивость проводилось изменение параметра num_topics на различные значения (например, 6, 12, 15). Анализ результатов для каждого значения num_topics позволил выявить оптимальное количество тем и оценить их устойчивость. Темы проверялись на наличие пересечений и уникальности ключевых слов.

Анализ результатов для различных значений. Определяемых переменной num_topics выполнялся для 6, 12 и 15 тем.

При выделении 6 тем они касались политической активности молодежи, влияния цифровых технологий, социальных аспектов и участия в политической деятельности.

При выделении 12 тем были выделены более детализированные аспекты политической активности молодежи, включая влияние цифровых технологий, социальные и культурные аспекты, а также исследовательские методы.

При выделении 15 тем оказалось, что темы становятся еще более узконаправленными, что позволяет глубже понять различные аспекты политической активности и участия молодежи.

Визуализация и интерпретация результатов тематического моделирования показали, что LDA модель позволяет эффективно выявлять и анализировать скрытые темы в текстовых данных, предоставляя ценные инсайты о структуре и направлениях исследований в области политической активности молодежи.

ЗАКЛЮЧЕНИЕ

Полученные результаты моделирования позволили выявить ключевые темы и направления исследований в области политической активности молодежи в цифровой среде. Анализ тем показал, что наиболее часто встречающиеся темы касаются наиболее часто встречающихся тем касаются влияния цифровых технологий на политическую активность молодежи, её участие в общественно-политических процессах через интернет и социальные сети. Также важными аспектами являются социальные взаимодействия и формирование гражданской позиции среди молодых людей в онлайн-пространстве. Анализ тематических моделей показал, что вопросы цифровой грамотности, информационной безопасности и влияния онлайн-платформ на формирование политических предпочтений также остаются актуальными и часто встречающимися темами исследований в этой области.

Результаты данного исследования могут быть использованы для разработки образовательных программ и стратегий по повышению политической активности молодежи в цифровой среде. Практическая значимость работы заключается в возможности использования полученных данных для анализа текущих тенденций и прогнозирования будущих изменений в политической активности молодежи.

В ходе работы были выявлены некоторые ограничения, связанные с качеством исходных данных и необходимостью более глубокой обработки текстов. В дальнейшем можно улучшить результаты моделирования за счет использования более сложных алгоритмов и методов глубокого обучения. Также перспективным направлением является расширение базы данных и включение в анализ большего количества текстов из различных источников.

Таким образом, проделанная работа позволила создать тематическую модель, выявить ключевые темы и направления исследований.