УДК 006 https://doi.org/10.26160/2541-8637-2025-15-78-83

КЛАССИФИКАЦИЯ НОРМАТИВНОЙ ДОКУМЕНТАЦИИ С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРИЗАЦИИ ЭМБЕДДИНГОВ: МЕТОДИКА И ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

Теремов И.А.
МИРЭА – Российский технологический университет, Москва

Ключевые слова: классификация документов, кластеризация, эмбеддинги, большие языковые модели, агломеративная кластеризация, машинное обучение.
Аннотация. В данной работе представлена методика автоматизированной классификации нормативной документации на основе кластеризации эмбеддингов. Актуальность исследования обусловлена значительным объемом нормативных документов в промышленности и информационных технологиях, что затрудняет их систематизацию и анализ. Предложенный подход включает этапы предобработки текстов, извлечения эмбеддингов, кластеризации с использованием агломеративного метода и интерпретации результатов с помощью генеративных языковых моделей. В ходе экспериментов проведена оценка различных моделей для векторизации текстов, а также выполнена кластеризация с применением косинусной метрики. Результаты показали, что метод обеспечивает высокую точность выделения тематических групп, превосходя традиционные алгоритмы на основе частотного анализа (BoW, TF-IDF).

CLASSIFICATION OF REGULATORY DOCUMENTATION USING EMBEDDING CLUSTERING: METHODOLOGY AND PRELIMINARY RESULTS

Teremov I.A.
MIREA – Russian Technological University, Moscow

Keywords: document classification, clustering, embeddings, large language models, agglomerative clustering, machine learning.
Abstract. This paper presents a methodology for automated classification of regulatory documents based on embedding clustering. The relevance of the study is due to the significant volume of regulatory documents in industry and information technology, which complicates their systematization and analysis. The proposed approach includes the stages of text preprocessing, embedding extraction, clustering using the agglomerative method and interpreting the results using generative language models. During the experiments, various models for text vectorization were evaluated, and clustering was performed using the cosine metric. The results showed that the method provides high accuracy in identifying thematic groups, surpassing traditional algorithms based on frequency analysis (BoW, TF-IDF).

Актуальность проблемы

Современные отрасли промышленности и информационных технологий характеризуются чрезвычайно высоким объёмом нормативной документации, включающей федеральные законы, государственные стандарты (ГОСТ), регуляционные правовые акты, технические условия и иные регламентирующие документы. Такая многообразность источников приводит к существенным трудностям при их систематизации, анализе и применении на практике [1]. Наличие многочисленных противоречий, неоднозначностей и локальных

вариантов классификации усложняет процесс определения требований к промышленным и интеллектуальным продуктам, что особенно критично в условиях стремительного технологического развития и глобализации рынков.

Проблематика и цели исследования

Традиционные методы классификации нормативной документации, основанные на ручном анализе и авторских подходах, не способны справиться с возрастающими объёмами данных и сложностью их семантического содержания. В то же время современные технологии машинного обучения и обработки естественного языка (NLP) открывают новые возможности для автоматизации анализа текстовых данных. В данной работе предлагается разработка методики построения классификатора, основанного на извлечении эмбеддингов из текстов стандартов с последующей их кластеризацией. Целью исследования является создание адаптивного и универсального инструмента, который позволит: 1) автоматически группировать нормативные документы по тематическому признаку; 2) выявлять скрытые семантические связи между документами; 3) обеспечить высокую точность и интерпретируемость полученных кластеров.

Обзор существующих подходов

Существующие методы классификации нормативной документации зачастую ограничены авторскими подходами, применяемыми в узких областях знаний. Традиционные алгоритмы, такие как методы наивного Байеса или опорных векторов, не учитывают глубинную семантику текстов и зачастую требуют ручного уточнения. Современные модели трансформеров, такие как BERT, Sentence-BERT (SBERT) и Universal Sentence Encoder (USE), позволяют получать векторные представления (эмбеддинги), которые более полно отражают смысловую нагрузку документов. Использование эмбеддингов как основы для кластеризации открывает возможность построения иерархических структур, способных объединить нормативные документы по тематикам и адаптироваться к изменениям в нормативно-правовой базе.

Для построения классификатора нормативной документации разработан конвейер, включающий этапы сбора данных, предобработки текстов, извлечения эмбеддингов, кластеризации и интерпретации полученных кластеров. Ниже описаны ключевые этапы предлагаемой методики.

Сбор и предобработка данных

Первичным этапом является формирование датасета нормативных документов. В исследовании в качестве исходного материала использовались стандарты системы ГОСТ, доступные как в машиночитаемом формате, так и посредством сканированных изображений, обработанных с помощью технологий оптического распознавания символов (OCR). Для обеспечения качества дальнейшей обработки применялись следующие шаги:

– *Нормализация текста:* приведение текстовых данных к единому формату, устранение опечаток и лишних символов.

– *Токенизация и лемматизация:* разбиение текста на отдельные элементы (токены) с последующим приведением к базовой форме, что способствует унификации терминологии.

– *Удаление стоп-слов:* исключение высокочастотных, но семантически нейтральных слов для уменьшения шума и повышения качества последующего извлечения эмбеддингов.

Извлечение эмбеддингов

Для получения числовых векторных представлений нормативных документов используются современные модели на базе архитектуры трансформеров. В рамках исследования проводился сравнительный анализ следующих подходов:

– *BERT и производные модели:* позволяют учитывать контекстуальные зависимости между словами, однако требуют значительных вычислительных ресурсов.

– *Sentence-BERT (SBERT):* оптимизирован для получения эмбеддингов предложений, демонстрируя высокое качество представления коротких текстовых фрагментов.

– *Universal Sentence Encoder (USE):* обеспечивает эффективное кодирование текстов на разных языках, что актуально при работе с нормативами, содержащими интернациональную терминологию.

Анализ показал, что применение SBERT позволяет достичь наилучшей семантической дифференциации между документами при приемлемых вычислительных затратах [2][3]. Таким образом, для каждого документа вычисляется вектор, отражающий его смысловую нагрузку, что становится фундаментом для дальнейшей кластеризации.

Кластеризация эмбеддингов с использованием агломеративного подхода

Ключевой частью методики является группировка эмбеддингов для выявления тематических кластеров. Для этой цели выбрана агломеративная кластеризация, обладающая рядом преимуществ:

– *Иерархическая структура:* Алгоритм последовательно объединяет наиболее близкие по смыслу документы, формируя дендрограмму, что позволяет визуально интерпретировать взаимосвязи между кластерами.

– *Отсутствие жесткого требования к количеству кластеров:* в отличие от методов, таких как K-средних, агломеративная кластеризация не требует предварительного задания числа групп, что важно при анализе неоднородных данных.

– ***Гибкость и интерпретируемость:*** Полученная иерархия кластеров может быть сопоставлена с существующими классификационными онтологиями (например, ОКС), что обеспечивает практическую применимость решения.

Этапы кластеризации:

1. Расчёт матрицы расстояний: с использованием косинусного расстояния определяется степень сходства между эмбеддингами.

2. Построение дендрограммы: на основе выбранной метрики и метода связи (например, полного соединения) строится иерархическая структура, отображающая слияние документов в кластеры.

3. Определение оптимального разреза: с использованием метрик, таких как силуэт-коэффициент и внутрикластерная дисперсия, определяется оптимальное число кластеров, позволяющее выделить тематически однородные группы.

Интерпретация кластеров и интеграция с существующими классификационными системами

Полученные в результате кластеризации группы документов представляют собой совокупности эмбеддингов, не имеющих изначально семантических ярлыков. Для автоматической интерпретации кластеров применяется генеративная модель на базе больших языковых моделей (LLM), таких как Llama или Qwen. Эти модели анализируют тексты, входящие в каждый кластер, и генерируют краткое тематическое описание, которое затем используется для сопоставления с существующими классификационными схемами, например, с Общероссийским классификатором стандартов (ОКС).

Интеграция полученных результатов с онтологической базой позволяет не только унифицировать классификацию нормативной документации, но и обеспечить адаптивность системы при изменении нормативно-правовой базы.

Предварительные результаты

В рамках предварительных экспериментов был реализован полный конвейер от сбора исходных данных до формирования тематических кластеров. Результаты демонстрируют высокую потенциал применения предлагаемого подхода для автоматизированной классификации нормативной документации.

После этапа предобработки текстов для каждого нормативного документа были получены векторные представления с использованием модели SBERT. Проведённый анализ показал, что полученные эмбеддинги обладают высокой дисперсией, что указывает на их способность различать тонкие семантические различия между документами. Это является важным условием для дальнейшей успешной кластеризации.

На основе вычисленных эмбеддингов выполнена агломеративная кластеризация с использованием косинусной метрики и метода полного соединения. Результатом данного этапа стало формирование дендрограммы, демонстрирующей иерархическую структуру групп документов. Основные результаты включают:

– *Выделение тематических ветвей:* Документы с близкой тематикой сгруппированы в отдельные кластеры, что подтверждает корректность выбора метода кластеризации.

– *Определение оптимального числа кластеров:* С использованием таких метрик, как силуэт-коэффициент и внутрикластерная дисперсия, был определён разрез дендрограммы, позволяющий достичь максимальной однородности внутри кластеров при чётком разделении между ними.

Оценка качества кластеризации

Для количественной оценки качества полученных кластеров использовались следующие показатели:

– *Силуэт-коэффициент:* Высокие значения данного показателя указывают на хорошо разделённые кластеры с минимальным перекрытием между группами.

– *Внутрикластерная дисперсия:* Анализ показал компактность кластеров, что свидетельствует о высокой семантической однородности документов внутри каждой группы.

– *Сравнение с традиционными методами:* Первоначальный сравнительный анализ с алгоритмами, основанными на частотном анализе (BoW, TF-IDF), подтвердил, что подход с использованием эмбеддингов обеспечивает более точное выделение тематических групп.

Интерпретация и перспективы развития

Для автоматического присвоения тематических меток полученным кластерам была использована генеративная модель на базе LLM. Полученные описания позволяют не только идентифицировать основные тематические направления, но и сопоставить их с существующими классификационными схемами (например, ОКС). Несмотря на предварительный характер экспериментов, результаты свидетельствуют о следующих преимуществах предлагаемого подхода:

– Повышенная точность классификации за счёт использования глубоких семантических представлений.

– Гибкость метода, позволяющая адаптироваться к новым данным и изменениям нормативной базы.

– Возможность автоматизации процесса анализа нормативной документации, что существенно сокращает затраты времени и ресурсов.

В дальнейшем планируется расширение датасета, оптимизация параметров моделей эмбеддингов и кластеризации, а также интеграция разработанного классификатора в информационные системы сопровождения производственных процессов и управления жизненным циклом продукции.

Заключение

Разработанная методика классификации нормативной документации на основе кластеризации эмбеддингов демонстрирует высокую потенциал для практического применения в условиях постоянно растущего объёма и сложности нормативных текстов. Использование современных моделей глубокого обучения (таких как SBERT) в сочетании с агломеративной кластеризацией позволяет формировать интерпретируемые и гибкие иерархические структуры, которые могут быть непосредственно интегрированы в существующие системы классификации, например, ОКС [4].

Предварительные результаты исследования свидетельствуют о том, что предложенный подход обеспечивает значительное улучшение качества классификации по сравнению с традиционными методами, позволяя более точно выделять тематические группы и выявлять скрытые семантические взаимосвязи между документами. Это особенно важно в условиях динамично изменяющейся нормативно-правовой базы и необходимости быстрой адаптации к новым требованиям.

Таким образом, предложенный подход представляет собой перспективное направление в области автоматизированной классификации нормативной документации, способное существенно повысить эффективность систем контроля качества, снизить трудозатраты на анализ нормативов и обеспечить унификацию

информации в условиях многопланового регулирования промышленности и информационных технологий.

Список литературы

1. Денисова О.А., Дмитриева С.Ю. Стандарт на SMART-стандарт: документ в деталях // Стандарты и качество. – 2023. – № 10. – С. 44-48.
2. Краснов Ф.В. Использование языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов на электронных торговых площадках // International Journal of Open Information Technologies. – 2023. – Т. 11, №9. – С. 33-40.
3. Салып Б.Ю., Смирнов А.А. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка // StudNet. – 2022. – Т. 5, №5. – С. 33.
4. Краснов Ф.В. Использование языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов на электронных торговых площадках // International Journal of Open Information Technologies. – 2023. – Т. 11, №9. – С. 33-40.

Сведения об авторе:
Теремов Иван Алексеевич – аспирант.