АННОТАЦИЯ

диссертации на соискание ученой степени «доктор философии» (Ph.D) по специальности 6D070400 - Вычислительная техника и программное обеспечение

Якунин Кирилл Олегович

“Разработка моделей и методов сбора, анализа и классификации медиа-публикаций на базе методов обработки естественных языков”

В диссертационной работе были предложены методы классификации текстов на естественном языке для решения задач медиа-мониторинга. В основе предложенных методов лежит идея применения больших объемов разнородных данных (Big Data) для извлечения из них знаний о скрытых структурах текстового корпуса, позволяющих построить эффективные векторные представления текстов, а также решать задачи классификации. При этом предложенные методы имеют низкую вычислительную сложность и позволяют создавать классификаторы текстов по произвольному критерию оценки с минимальным объемом ручной разметки, либо в ряде случаев автоматически (без разметки).

Цель работы. Целью работы является разработка моделей и методов для автоматической многокритериальной оценки текстовой информации из медиа-источников и социальных сетей в рамках распределенной информационной системы.

Задачи исследования

Для достижения цели в работе поставлены следующие задачи:

– выявить основные факторы, ограничивающие возможность построение эффективных (с точки зрения производительности и требуемой экспертной разметки) классификационных моделей для текстов;

– разработать метод эффективной векторизации текстов на базе неразмеченных данных (Big Data);

– разработать метод самообучения (self-learning) классификационной модели на основе объективных явных показателей, для распознавания скрытых параметров документов;

– разработать методику многокритериальной оценки документов и медиа-источников;

– разработать распределенную информационную систему для сбора, хранения, обработки и классификации текстовой информации из масс-медиа и социальных сетей и верифицировать точность результатов работы информационной системы.

Актуальность исследования. Задачи медиа-мониторинга пользуются большой популярностью на мировом рынке, в первую очередь в виде продуктов по менеджменту репутации. Однако более высокоуровневые задачи, связанные с принятием решений на основе данных из медиа-пространства на государственном уровне, на данный момент практически не решаются в автоматическом режиме.

Несмотря на значительную проработанность темы классификации текстов, современные SoTA (state of the art – лучшие на текущий момент) модели требуют больших объемов разметки для классификации по заданному критерию, а также высокой производительности (сотни и тысячи процессоров для обучения). Следовательно, существует потребность в более эффективных с точки зрение требуемой экспертной разметки и вычислительной производительности моделях классификации и анализа текстовой информации.

Объектом исследования являются методы классификации текстовых документов.

Предмет исследования - методы многокритериальной оценки текстовых документов на базе тематического моделирования

Цель диссертационного исследования – разработка моделей и методов для автоматической многокритериальной оценки текстовой информации из медиа-источников и социальных сетей в рамках распределенной информационной системы.

Методы исследования. В диссертационной работе применяются следующие методы исследования: методы классификации, методы векторизации текстовых документов, методы многокритериального анализа для принятия решения, технологии проектирования и разработки информационных систем.

Научная новизна

1. Предложен метод векторизации текстовых документов с помощью тематической модели BigARTM
2. Предложен метод оценки тематического межкорпусного дисбаланса для самообучения классификационной модели
3. Предложена методика многофакторной оценки социальной значимости публикации
4. Предложена методика многокритериальной оценки масс-медиа ММА на базе байесовской системы агрегации, метода анализа иерархий (AHP) и тематического моделирования

Основные результаты исследования заключаются в: разработке методики многокритериальной оценки текстовых документов и медиа-источников и сопутствующих методов для самообучения и обучения

соответствующих моделей; разработке распределённой информационной системы для сбора, хранения, обработки и классификации данных масс-медиа и социальных сетей, являющейся программной реализацией предложенных методов; многофакторной методики оценки социальной значимости публикаций в СМИ и социальных сетях. Можно выделить следующие практические выводы и рекомендации, полученные при выполнении диссертационной работы:

1 Проведен анализ рынка систем, предоставляющих услуги медиа-мониторинга, а также анализ нормативно-правовой основы и технических особенностей. Выявлены слабые стороны существующих решений, сформированы рекомендации;

2 Исследован вопрос влияния открытых информационных источников на обществе, выявлены основные направления влияния, сформирован перечень информативных признаков, на основе которых можно оценить это влияние;

3 Исследованы существующие подходы классификации документов и векторизации текстов, выявлены проблемы и слабые стороны текущих решений, сформированы рекомендации;

4 Разработан подход векторизации текстов на основе тематической модели;

5 Разработан метод оценки межкорпусного тематического дисбаланса, позволяющий автоматически или полуавтоматически получать веса топиков по отношению к заданному признаку;

6 Разработан метод мультикритериальной оценки медиа-источников ММА на базе байесовской модели агрегации;

7 Разработана распределенная информационная система на базе Open Source решений, позволяющая производить сбор (скрапинг), хранение, обработку текстовой информации, а также построение тематических моделей и классификаторов с возможностью визуализации полученных результатов;

8 Собран корпус, состоящий из более чем 6 миллионов публикаций из казахстанских и российских источников, включая как тексты публикаций, так и метаданные;

9 Проведена валидация предложенных моделей и методов. Основная серия вычислительных экспериментов была связана с определением так называемых опасных новостей (социально значимые, резонансные негативные публикации). Было проведено сравнение с моделью глубокого обучения BERT, метрики качество предложенных моделей сопоставимы, при меньшей вычислительной сложности и меньших требованиях к объему ручной экспертной разметки.

Научные результаты позволили разработать модели, методы и программные инструменты, позволяющие классифицировать текстовые документы по ряду признаков (критериев) с минимальным объемом ручной разметки, с применением так называемой высокоуровневой разметки топиков, либо метаданных статей. Разработанные в рамках проводимого исследования информационная система является легко масштабируемой (как

функционально, так и с точки зрения производительности) и позволяет хранить, обрабатывать, агрегировать и визуализировать большие объемы текстов данных.

Рекомендации и исходные данные по конкретному использованию результатов

Результаты научного исследования, в частности разработанная информационная система может быть использована для ряда целей:

1. Использование исследователями и учёными. Как показано в разделе 5.3.1 работы, существует большой потенциал для использования разработанной информационной системы для самых разных гуманитарных исследований.
2. Использование крупными компаниями и государственными органами для поддержки принятия решений.
3. Использование крупными компаниями для решения задача менеджмента репутации.
4. Использование обычными пользователями для разведывательного поиска интересующих данных.

При этом нужно отметить универсальность системы – она может быть применена к самым разным корпусам текстовых данных, например к внутренним документам организаций, научным публикациям, личными перепискам и т.п.

Положения, выносимые на защиту:

1. Разработана методика мультифакторной оценки социальной значимости на базе многокритериальной методики оценки масс-медиа ММА с использованием тематической векторизации текстовых документов.
2. Разработана архитектура и программная реализация распределенной информационной системы для сбора, обработки, оценки и визуализации текстовых данных масс-медиа и социальных сетей Media Analytics, включающий предложенную методику оценки, а также ее оценку и верификацию.

Связь темы с планами научно–исследовательских программ. Представленные результаты получены при выполнении проекта ИИВТ КН МОН РК (источник финансирования Комитет науки МОН РК): программно-целевого финансирования (ПЦФ) КН МОН РК BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» в 2018-2020 годы;

Апробация результатов исследования. Основные положения и результаты исследования докладывались на: Procedia 9th International Young Scientist Conference (Scopus, 2020), International Conference on Digital Transformation and Global Society (Springer, 2019, 2020) и других конференциях. Основные результаты исследования опубликованы в журнале Symmetry, имеющем импакт-фактор 2.51 (Q1). Разработанная информационная система внедрена в МОН РК (Приложение А Диссертации).

Всего по теме диссертации опубликовано 16 работ, из которых 6 статьей опубликованы журналах, входящих в базы Scopus и Thomson Reuters (2 статьи - Q1, 2 статьи - Q2, 2 статьи - Q3), одна из них опубликована в изданиях, рекомендованных Комитетом по контролю в сфере образования и науки МОН РК, 3 статьи в журналах имеющих CiteScore (2 статьи - CiteScore 17%, 1 статья – CiteScore 69%) в Scopus без присвоенного квартиля, 2 статьи в журналах, входящих в РИНЦ (импакт-фактор 0.482 и 1.385), 6 статей опубликованы в сборниках международных научно-практических конференций. Оформлено 4 авторских свидетельства на результаты работы.

Структура и объем диссертации. Работа состоит из введения, пяти разделов, заключения и списка использованной литературы. Общий объем диссертации 130 страниц.