Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

УДК 004.3.049.77

ПОТЕХИН
Александр Сергеевич

РАЗРАБОТКА ПРОГРАММЫ СБОРА ДАННЫХ О СТРУКТУРЕ ВЕБ-САЙТОВ

АВТОРЕФЕРАТ
диссертации на соискание степени
магистра информатики и вычислительной техники

по специальности 1-40 81 01 -- Информатика и технологии разработки программного обеспечения

Научный руководитель
Стержанов М.В.
к.т.н., доцент

Минск 2020

Работа выполнена на кафедре информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **СТЕРЖАНОВ Максим Валерьевич,**
кандидат технических наук, доцент кафедры информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Рецензент: **ПЛЮЩ Олег Борисович,**
кандидат физико-математических наук, доцент кафедры управления информационными ресурсами учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Защита диссертации состоится «26» июня 2020 г. года в $15^{\underline{00}}$ часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Гикало, 9, копр. 4, ауд. 114, тел. 293-85-91, e-mail: inform@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

Сентимент-анализ (далее СА; sentiment analysis) — это раздел текстомайнинга (text mining), система автоматического извлечения субъективных мнений из текста, дисциплина на стыке поиска информации и вычислительной лингвистики, которая исследует не столько содержание текста, сколько его тональность. Понятие сентимент-анализа имеет ряд синонимов и близких терминов: сентиментанализ (sentiment analysis), сентиментометрия (sentiment metrics), брендмониторинг (brand monitoring), социомедиаанализ (social media analysis), разведка мнений (opinion mining), «подслушивание» мнений (oppinion listening), анализ тональности текста и т.п. Говоря о тональности текста, следует выделять три параметра: субъект тональности (автора текста), тональную оценку (позитив, нейтрально или негатив либо более детальное деление) и объект тональности (предмет, о котором высказывается мнение, тональная оценка).

Из определения можно сделать несколько выводов о том, где теоретически (и, если уж на то пошло, практически) концепция анализа тональности текста могла бы найти применение и прояснить некоторые ее детали.

Во-первых, анализ тональности текстов способен помочь разобраться в законах, по которым живет естественный язык и научить компьютер воспринимать его на уровне, приближенном к человеческому. До недавнего времени машина понимала тексты на абстрактном уровне – в основном, через лексемы (слова), которые для нее обладали формой (набор букв) и содержанием (значение). Данная концепция предлагает ввести еще одну функцию – так называемую лексическую тональность текста (в простейшем случае она будет определяться как сумма лексических тональностей каждой отдельной лексемы).

Во-вторых, анализ тональности способен значительно повысить качество машинного перевода. Известно, что эталоном машинного перевода служит результат перевода текста человеком – профессиональным переводчиком. За 50 с лишним лет разработок в этой области исследователи убедились в том, что научить машину «думать, как переводчик» можно лишь приняв во внимание все те соображения, которыми пользуется профессионал, переводя тот или иной текст. Естественно, при переводе не обойтись без первичного анализа текста и отдельных слов – в том числе, анализа тональности как таковой.

В-третьих, целью анализа тональности текста может быть некое мнение автора или сам автор. Это – наиболее интересная сфера применения, поскольку здесь видится не только способ делегирования машине некоторых полномочий ученого (например, филолога, который исследует произведение того или иного автора), но и снова попытка приблизить образ мышления компьютера к человеческому. С этой точки зрения анализ тональности, возможно, является одним из самых важных и перспективных шагов к развитию искусственного интеллекта.

Автоматический анализ тональности текста базируется на технологиях лингвистической интерпретации эмоций, машинного обучения, извлечения

эмоционального смысла из информации и т.д. Технология может использоваться для автоматической оценки новостных событий, продуктов, персоналий, организаций, стран и т.д. К задачам СА относятся распознавание и интерпретация мнения, кластеризация текстов, исходя из полярных (позитивных или негативных) мнений; сегментация текстов по разным мнениям; прогнозирование мнений, исходя из анализируемых текстов.

Существующая на протяжении уже нескольких десятилетий, технология СА стала особенно актуальной с развитием Web, особенно Web 2.0, как инструмент мониторинга мнений миллионов пользователей Сети, которые постоянно высказывают свои мысли в разного рода социальных сетях, блогах, твиттах и т.п.

Интерес к данной технологии растет по мере того, как повышается доверие к информации из социомедиаресурсов.

Широта охвата аудитории в миллионы человек и оперативность извлечения информации (она доступна практически в режиме реального времени) позволили получать недостижимые ранее результаты исследований. Если раньше, чтобы выявить мнение по какому-либо вопросу, нужно было проводить опросы, то сегодня высказывания по огромному количеству популярных тем уже есть в Сети, надо только выявить их, распознать и оценить. На базе СА-технологий разработан богатый набор программных приложений.

Наиболее простой метод автоматического определения мнения автора состоит в выделении и подсчете в тексте количества слов, имеющих позитивную или негативную окраску. Если в нем преобладают слова типа «удобный», «практичный», «стильный», то, скорее всего, тональность текста, описывающего предмет, положительная, и, наоборот, слова «скучный», «плохой», «проблемный» свидетельствуют об отрицательном отношении автора к нему.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Современные технологии по выявлению тональности текста позволяют классифицировать полярности документа. Кроме того, технология сентимент-анализа позволяет находить мнения в тексте и выявлять их свойства. Данная технология может использоваться для автоматической оценки новостных событий, продуктов, персоналий, организаций, стран и т.д. В сфере новостных статей технология особенно актуальна, так как позволяет распознавать и интерпретировать мнения, кластеризировать тексты исходя из полярных мнений, сегментировать тексты по разным мнениям, прогнозировать мнения исходя из анализируемых текстов.

В связи с вышесказанным, актуальной является разработка анализа тональности текста на базе нейронной сети.

Степень разработанности проблемы

Исследование анализа тональности текста осуществлялось на основе построения теоретических моделей с использованием работ российских и белорусских ученых: М.И. Горлова, В.А. Емельянова, Л.П. Ануфриева, В.Ф. Алексеева, Г.А. Пискуна, Л.Н. Кечиева, Е.Д. Пожидаева, В.А. Каверзнева, Г.Д. Грошева, а так же зарубежных авторов: Ч. Джоввета, Кая Есмарка, А. Шваба, Э. Хабигера, Steven H. Voldman и др.

Одним из недостатков исследований, представленных в современной технической литературе, является неполное рассмотрение особенностей и условий для моделирования анализа.

Предложенное исследование направлено на устранение этого недостатка на основе модификации алгоритма анализа текста с использованием нейронных сетей.

Цель и задачи исследования

Целью диссертационной работы является создание приложения для сбора данных о структуре веб-сайтов для последующего исследования научного и практического применения анализа тональности текста для выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов, последующей обработки и использования, в частности, при исследовании медиасферы.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Исследовать выбранные технологии на предмет применимости в отношении анализа тональности текста.
2. Обосновать необходимость применения сентимент-анализа текста в медиасфере.
3. Разработать архитектуру приложения для сбора данных из новостных статей.
4. Разработать архитектуру приложения для анализа собранных данных.

Область исследования

Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры) ОСВО 1-39 81 01-2012 специальности 1-40 81 01 «Информатика и технологии разработки программного обеспечения».

Теоретическая и методологическая основа исследования

В основу диссертации легли работы белорусских и зарубежных ученых в области определения анализа текстов с помощью нейронных сетей, а также анализ приложений для скачивания информации из интернет-ресурсов.

Информационная база исследования сформирована на основе литературы, открытой информации, технических нормативно-правовых актов, сведений из электронных ресурсов, а также материалов научных конференций и семинаров.

Научная новизна

Научная новизна и значимость полученных результатов работы заключается в разработке методики сбора данных с веб-сайтов и анализа тональности на основе изучения мнений на уровне коллекций что позволяет более точно исследовать мнения и их аспекты.

Теоретическая значимость работы заключается в детальном анализе протекающих процессов при сборе данных с веб-сайтов и построения модели нейронной сети для выявления тональности текста.

Практическая значимость диссертации состоит в разработанной схемотехнической модели структур, которая позволит оптимизировать процесс сбора данных с веб-сайтов и анализировать собранные данные.

Основные положения, выносимые на защиту

1. Систематизация механизмов работы со сбором данных с веб-сайтов, основанная на анализе особенностей веб-сайтов и процедуры сбора данных, позволившая более детально описать специфику работы и струкруты веб-сайтов.

2. Модель построения анализа тональности текстов, построенная на базе моделей SLDA и ASUM, позволяющая производить автоматизированный анализ текстов по моделируемым парам аспект-сентимент, которые реализуют мультиномиальное распределение над словами, которые описывают похожие слова в тексте.

3. Экспериментально установленный диапазон оптимальных параметров на задачах анализа тональности на различной размерности, а так же зависимость полученного результата от инициализации нейронной сети.

Апробация диссертации и информация об использовании ее результатов

Результаты исследований, вошедшие в диссертацию, докладывались и обсуждались на Международной научно-практической конференции «Информационные технологии и системы 2019» (г. Минск, 2019 г.), XII международной молодежной научно–практической конференции аспирантов, магистрантов и студентов БГУИР (г. Минск, Беларусь, 2019 г.), 12-й Международной научно-технической конференции «Научный потенциал молодежи – будущему Беларуси» (г. Пинск, 2018 г.).

Отдельные положения диссертации могут быть использованы при преподавании дисциплины «Нейронные сети».

Публикации

Изложенные в диссертации основные положения и выводы опубликованы в 2 печатных работах. В их числе 2 статьи в сборниках материалов научных конференций и 2 тезиса докладов на научных конференциях.

Общий объем публикаций по теме диссертации составляет 5 страниц.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, трех глав с краткими выводами по каждой главе, заключения, библиографического списка и приложений.

В первой главе приведено теоретическое исследование используемых технологий веб-скрепинга и сентимент-анализа, а также рассмотрено возможность применения данных технологий в медиасфере.

Во второй главе представлен обзор медиасферы, ее недостатки и способы решений выявленных недостатков.

В третьей главе представлено теоретическая реализация моделей скрепера новостных сайтов, сентимент-анализа и пользовательского приложения.

В четвертой главе представлено практическое исследование результатов работы, в том числе тестирование полученного решения, прикладной анализ применения решения в реальной жизни в медиасфере, а так же представлено сравнение полученного решения с существующими аналогами.

В приложении представлены примеры кода по реализации сервиса аналитики, сервиса по скачиванию данных с веб-сайтов, сервиса по обработке данных с веб-сайтов и сервиса клиентского приложения.

Общий объем работы составляет 102 страницы, из которых основного текста – 52 страницы, 27 рисунков на 14 страниц, список использованных источников из 94 наименований на 7 страниц и 4 приложения на 29 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы анализа тональности текста как основополагающей проблеме развития искусственного интеллекта, рассмотрены основные направления в области сентимент-анализа и техлоголии на которых он базируется, указаны основные направления исследований, проводимых по данной тематике, а также описано обоснование актуальности темы.

В **общей характеристике работы** показана актуальность проводимых исследований, степень разработанности проблемы, сформулированы цель и задачи диссертации, обозначена область исследований, научная (теоретическая и практическая) значимость исследований, а также апробация работы.

В первой главе приведено теоретическое исследование используемых технологий веб-скрепинга и сентимент-анализа, а также рассмотрено возможность применения данных технологий в медиасфере.

Из анализа следует, что проблема реализации веб-скрепинга в отсутствии точных алгоритмов и методик моделирования воздействия на веб-сайты,

так как многие из них реализованы в разных средах программирования с использованием разных фреймворков. Решение этой проблемы позволит оптимизировать процесс скачивания и обработки данных с веб-сайтов, а также сократить затраты времени и ресурсов на работу скрепера, что обуславливает актуальность проводимых исследований.

Проанализированы особенности реализации сентимент-анализа. Выявлено, что сентимент-анализ имеет множество проблем для реализации, таких как зависимость тональности текста от предметной области, использование отрицаний и саркастичной лексики в текстах, зависимость тональности от автора текста. При проведении анализа проблем было выявлено, что существуют подходы, которые позволят обойти данные проблемы.

Во второй главе представлен обзор медиасферы, ее недостатки и способы решений выявленных недостатков.

Проводя анализ проблем в медиасфере можно обозначить основные, которые может решить сентимент-анализ:

1) Аналитика. Анализ разнообразных ресурсов позволяет получать информацию о том, как меняется во времени мнение потребителей, клиентов и конкурентом о разнообразных продуктов компании, что позволит компаниям получать полную и достоверную информацию, по которой можно строить различные бизнес-планы.

2) Быстрое реагирование на мнения. Реализация мониторинга социальных разговоров позволяет компаниям получать достоверное мнение об их продуктах, позволит понимать что больше всего необходимо клиентам.

3) Проведение исследований. Анализ большого количества ресурсов и выделение ключевых слов и мнений авторов значительно упрощает проведение исследований, основанных на медиасфере.

4) Прогнозирование. Выявление обсуждаемых темы, юридических лиц, проводидение их классификацию, выполнение сентимент-анализа как по отдельным темам, так и по всему тексту, позволяет анализировать ряд параметров, например выяснить, что автор планирует, насколько он уверен в своих намерениях, оценить степень эмоциональной выраженности автора и ответить на его вопросы.

5) Продвижение брендов. Сантимент-анализ предоставляет возможность не только оценить тональность высказываний о бренде, но и получить целый ряд дополнительных инструментов, упрощающих управление социальной аудиторией, интересующейся брендом, установление контактов, обмен информацией, влияние на взращивание социального контента, поиск лидеров мнений социального сообщества, снабжение их информацией и привлечение к продвижению бренда.

6) Перспективность. Технология сентимент-анализа выглядит молодой и имеющей разнообразные варианты своей эволюции.

В третьей главе представлено теоретическая реализация моделей скрепера новостных сайтов, сентимент-анализа и пользовательского приложения.

На основании проведенного анализа были выбраны следующие технологии для реализации приложения. Для сервера будет использоваться Heroku,

для реализации скрепера новостных сайтов будет использоваться язык Ruby, для анализа тональности текстов будут использоваться сервисы машинного обучения и аналитики на основе языка Python и фреймворков, для хранения промежуточных данных и данных анализа будет использоваться база данных PostgreSQL.

Для реализации скрепера был выбран язык программирования Ruby из-за того, что на нем быстро и удобно реализовывать оболочку, которая будет управлять работой библиотек, написанных на C++. Скрапер поддерживает граф связей узлов, различные фильтры и нормализаторы URL. Он позволяет использовать различные хранилища данных, такие как Cassandra, Hbase и др. Скрапер также является масштабируемым (до 100 узлов в кластере) и легко настраивается и расширяется, в полной мере является “вежливым”.

Основным ядром скрепера будет библиотека для получения данных ‘curb’ (CUrl-RuBy), она предоставляет собой связывание языка Ruby и libcurl - полнофункциональной клиентской библиотеки для передачи данных с URL-адресов.

Порядок работы скрапера:

1. На вход поступает URL адрес сайта с новостями.

2. В папке с проектом или в переданных файлах происходит поиск инструкций разбора страниц данного сайта и настроек для потоков скачки.

3. Скрипт в процессе работы анализирует страницы, строит дерево обхода сайта, нормализует URL на следующие страницы и если страница подходит под страницу с необходимым контентом, загружает ее в базу.

Для реализации анализа тональности текстов будет использована нейросеть реализованная на языке программирования Python. Так же для работы нейронной сети будет использоваться тональный словарь, полученный благодаря краудсорсинговому веб-ресурсу Linis Crowd (http : //linis − crowd.org/) Лаборатории интернетисследований НИУ ВШЭ. Результатом проекта является общедоступная коллекция размеченных пользовательских интернет-текстов общественно-политического содержания и общедоступный тональный словарь, созданный на основе коллекции с помощью технологии краудсорсинга и, таким образом, учитывающий восприятие слов широким кругом самих интернет-пользователей.

Последней доступной версией является словарь за 2016 год. Он представляет собой список слов с тональностями от -2 до 2, указанными пользователями. Каждое слово встречается столько раз, сколько раз его оценивали. После обработки всех оценок слов и удаления нейтральных итоговый размер составил 2454 слов.

Для расширения словаря использовался подход с добавлением синонимов. В качестве аналога WordNet для русского языка использовался RuWordNet. Это вариация тезауруса РуТез, которая была получена его автоматизированным преобразованием в стандартную структуру Wordet. Он содержит синсеты для существительных, прилага21 тельных и глаголов и связи между ними. В результате такой обработки размер словаря учеличился до 3419 слов.

Дерево зависимостей позволяет определить явные синтаксические зависимости между элементами. Поэтому в работе используется этот подход синтаксического разбора. С этой целью использовался фреймворк SyntaxNet (рисунок 1). Эта система была разработана с целью придания возможности компьютерным системам читать и понимать человеческий язык. Она поддерживает более 40 языков, в том числе и русский.

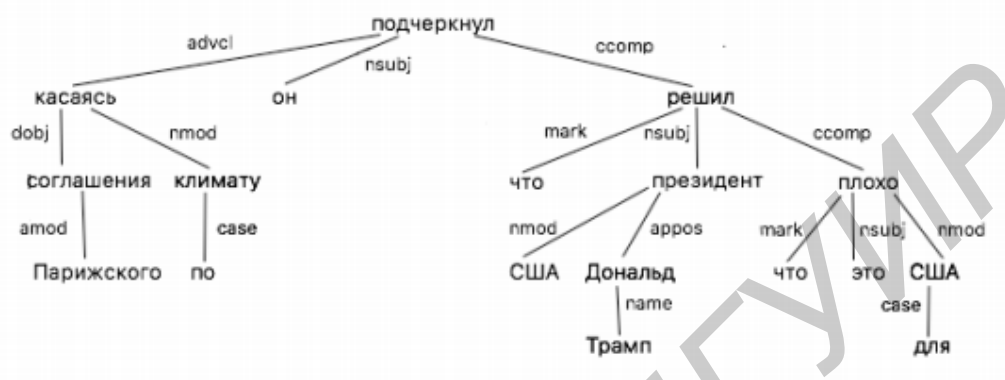

Рисунок 1 - Дерево зависимостей предложения, построенное с помощью SyntaxNet

На вход SyntaxNet подается файл, каждая строка которого соответствует одному предложению. При разборе происходит автоматическое определние частей речи слов, зависимостей между ними, тип связи и некоторая другая информация. Индексы родительских слов и тип связи помогают определить струтуру дерева.

В результате по построенному дереву можно более точно определить как и на какие слова влияют тоанльные. В модальности тематической модели добавляются тональные слова и «окрашенные» ими, если они являются терминами, и пара тональное+термин. Влияние тональных слов определялось следующим образом:

1. Если тональное слово прилагательное, существительное или наречие, то влиянию подвергается родительское слово (рисунок 2 а).

Рисунок 2 – Правила по влиянию тональных слов: a) связи для глагола и прилогательного; b) влияние частицы «не»

2. Если это глагол, то «окрашивались» слова объекты и субъекты (связь типа obj, subj), для которых это слово было родительским (рисунок 2 а).

3. Если одним из «детей» тонального слова является частица «не» (связь типа neg), то знак меняется на противоположный (рисунок 2 б).

Примеры влияния тональных слов в документах изображены на рисунке 3. Тональные слова, подкрашены красным или зеленым цветом в зависимости от тональности, оранжевым — окрашенные ими термины.

Мнение 1: Бывший кандидат в президенты Соединенных Штатов Америки Хиллари Клинтон считает исторической ошибкой(-1) решение президента Дональда Трампа выйти из Парижского соглашения по климату. В Организации Объединенных Наций разочарованы(-1) решением президента США Дональда Трампа выйти из Парижского соглашения по климату.

Мнение 2: 1 июня Трамп заявил, что принял решение о выходе из Парижского соглашения ради защиты(+1) Америки и ее граждан. По его словам, Вашингтон начнет переговоры о заключении соглашения по климату на условиях, более выгодных(+1) для США. Трамп не раз публично называл глобальное потепление мистификацией(-1). 28 марта он подписал указ об энергетической независимости , которым отменил ряд решений администрации Обамы по борьбе с глобальным потеплением.

Рисунок 3 - Окраска тематических слов тональными

BigARTM. BigARTM — это библиотека тематического моделирования с открытым кодом (http://bigartm.org). В ней реализуется идея APTM и возможности, описанные в предыдущих главах, имеется набор регуляризаторов и метрик качества для оценки тематических моделей.

В данной работе будет испробован метод, который использует тематическое моделирование для поиска мнений в текстах. Эта возможность будет реализована в виде дополнительного регуляризатора.

В четвертой главе представлено практическое исследование результатов работы, в том числе тестирование полученного решения, прикладной анализ применения решения в реальной жизни в медиасфере, а так же представлено сравнение полученного решения с существующими аналогами.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Выполнен анализ существующих методов сбора данных о структуре веб-сайтов и выявлены основные правила написания программ по сбору данных с интернет-ресурсов [1, 2].

Рекомендации по практическому использованию результатов

Полученные результаты могут быть использованы для исследования больших объемов текстовых статей, в частности для исследований медиасферы.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Тезисы конференций

1. Потехин А.С., ПРОГРАММА СБОРА ДАННЫХ О СТРУКТУРЕ ВЕБ-САЙТОВ / Потехин А.С. / Информационные технологии и системы 2019 (ИТС 2019): Материалы международной научной конференции, БГУИР, г. Минск, 30 октября 2019 / Белорусский государственный университет информатики и радиоэлектроники – Минск, 2019 – С 326-327

2. Потехин, А. С. ПРОГРАММА СБОРА ДАННЫХ О СТРУКТУРЕ ВЕБ-САЙТОВ / Потехин А.С. / Научный потенциал молодежи – будущему Беларуси: материалы XII международной молодежной научно–практической конференции, УО "Полесский государственный университет", г. Пинск, 6 апреля 2018 г. Ч.1 / Министерство образования Республики Беларусь [и др.]; редкол.: К.К. Шебеко [и др.]. – Пинск: ПолесГУ, 2018. – С. 262-263