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

УДК 004.42:316.472.4-047.44

Разумов
Андрей Васильевич

Средства лингвистического анализа сообщений в социальных сетях

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности 1-40 80 03 «Вычислительные машины и системы»

_____

Научный руководитель
Насуро Екатерина Валериевна
Кандидат технических наук

_____

Минск 2018

КРАТКОЕ ВВЕДЕНИЕ

Общественное мнение сегодня является важным индикатором состояния социально-экономической системы, поскольку отражает уровень социальной напряженности. Учет и контроль этого уровня позволяет выстраивать стратегическое планирование для обеспечения устойчивого развития социально-экономической системы, будь то промышленное предприятие или государство в целом. В связи с этим, мониторинг общественного мнения является важным и актуальным инструментом управления, активно применяемым социально-политическими, финансово-экономическими и общественными структурами.

Активный рост аудитории социальных медиа в сети Интернет, таких как социальные сети, форумы, блоги и интернет-СМИ, привел к становлению этих ресурсов в качестве нового источника данных о мнении и настроении граждан. Специфика работы с такими данными несет в себе целый ряд преимуществ и недостатков. К преимуществам относится скорость доступа к информации, охват аудитории и спектр выражаемых мнений. Одним из главных достоинств, как и серьезным препятствием, является объем этих данных. Так, согласно статистике, на 2016-й год, ежемесячно в русскоязычных социальных сетях около 30 миллионов уникальных авторов публикуют почти 580 миллиардов сообщений.

Крупные компании используют социальные сети для исследования мнений о своих продуктах. Такой подход, в отличие от использования опросов на сайте производителя и работы фокус-групп, обеспечивает большую широту исследования мнений.

Для решения задач, связанных с выявлением и дальнейшим анализом эмоционально окрашенной лексики в тексте, используются методы, общее название которых – анализ тональности текста (Sentiment Analysis). Другое название данной области на русском языке – «анализ эмоциональной окраски текста». Анализ тональности текста входит в область задач компьютерной лингвистики и является подзадачей получения и обработки информации (Information Retrieval).

Миллиарды публикаций, оставляемых пользователями ежемесячно, невозможно обработать вручную при проведении исследования общественного мнения. Этот факт выдвигает на первый план потребность в методах автоматизированного интеллектуального анализа текстовой информации, позволяющих за короткое время обработать большие объемы данных и понять смысл пользовательских сообщений. Именно понимание смысла сообщений является наиболее важным и сложным элементом автоматизированной обработки.

Таким образом, актуальность данной работы обусловлена необходимостью развития методологического аппарата, который позволил бы использовать большие объемы публикаций пользователей социальных сетей для решения комплекса задач по автоматизации мониторинга общественного мнения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью диссертации является разработка классификатора текстовых сообщений социальной сети Твиттер на основе наивного Байесовского классификатора, учитывающий особенности выбранной социальной платформы с целью улучшения базового алгоритма.

Для выполнения поставленной цели в работе были сформулированы следующие задачи:

− проанализировать особенности задачи в отражении на микроблоги;

− исследовать особенности написания текстовых сообщений в социальной сети с целью выбора необходимого для анализа метода;

− сравнить базовые методы обучения с учителем и выбрать лучший по параметрам точности, полноты результатов и времени обучения;

− предложить и обосновать новый метод на основе выбранного;

− оценить результаты работы нового метода.

Объектом исследования является процесс автоматической классификации текстовых сообщений социальных сетей.

Предметом исследования является использование особенностей сети Твиттер для улучшения точности принятия решения относительно входного сообщения.

Положения, выносимые на защиту:

1. Классификация теоретических и практический подходов в области автоматического анализа текстовых сообщений по используемых в них методах и матиматических моделях, позволившие упростить, а главное ускорить процесс обработки большого объема текстовой информации.

2. Модель классификатора, которая в отличие от существующих учитывает большинство особенностей целевой области применения: хештеги, смайлы, пролонгирование, опечатки и т. д.

3. Экспериментальные результаты, полученные в результате работы разработанной системы, показавшие улучшение базового алгоритма и увеличение точности классификации. Зависимости воздействия различных комбинаций использования особенностей социальной сети Твиттер на итовый результат работы программного решения.

4. Рекомендации по дальнейшему развитию и улучшению представленного алгоритма на основе полученных результатов.

Результаты исследования были представлены на 53-й научной конференции аспирантов, магистрантов и студентов БГУИР в 2017 году.

Основные положение работы и результаты диссертации изложены в работе, опубликованной на студенческой конференции от БГУИР.

Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, четырех глав и заключения, библиографического списка и приложений. Общий объем диссертации − 66 страниц. Работа содержит 7 таблиц, 4 рисунка. Библиографический список включает 27 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении рассмотрено современное состояние проблемы автоматического сбора и анализа большого объема публикаций пользователей в социальных сетях, определены основные направления исследований, а также дается обоснование актуальности темы диссертационной работы.

В первой главе производится анализ предметной области, проблемы и методы анализа информации, а также существующие методы анализа эмоциональной окраски текстовых сообщений. Сформулирована цель и задача работы, показана связь с научными программами и проектами, даны сведения об объекте исследования и обоснован его выбор.

Во второй главе рассмотрены модели и методы, положенные в основу разрабатываемой системы. Проведен анализ основных особенностей выбранной социальной сети – Твиттер. Подробно рассмотрена математическая модель Байесовского классификатора, а также проблемы его реализации (проблема неизвестных слов) и способы их решения.

В третьей главе произведено планирование разработка средства автоматического анализа текстовых сообщений на основе Байесовского классификатора. Описаны основные этапы обработки сообщений, точки расширений и варианты улучшения алгоритма.

В четвертой главе произведено тестирование и анализ полученных результатов разработанной системы. Проведено сравнение результатов работы системы в различных режимах, сделаны выводы и предложены варианты дальнейшего развития системы.

В приложении приведены графические материалы (презентация), необходимые для защиты данной работы.

ЗАКЛЮЧЕНИЕ

В данной работе был проведен анализ существующих методов и средств, помогающих анализировать текстовые сообщения. Были выделены основные достоинства и недостатки каждого из существующих методов, а также их области применения.

На основе анализа был выбран метод классификации наивным Байесовским классификатором, время обучения которого на порядок ниже других, с целью исследования возможности его улучшения. В ходе работы были проанализированы особенности задачи анализа высказываний на примере социальность сети Твиттер и найдены варианты их использования.

На основе наивного Байесовского классификатора предложен, обоснован и реализован новый метод. Для улучшения использовались: переход к байесовскому подходу (апостериорные вероятности), использование смайлов и хештегов, как особенностей выбранной социальной сети, устранение проблемы неизвестных слов сглаживающим методом, а также обучение в процессе работы приложения. Структурная модель системы поддерживает возможность расширения и использования в различных режимах, тем самым являясь универсальным скелетом для дальнейшей модификации и применения новых подходов для улучшения точности классификации.

Полученные эмпирические результаты были проанализированы и дополнены пояснениями, на их основаниях были сделаны выводы о целесообразности использования разработанного метода анализа текстовых сообщений социальной сети Твиттер.

На основе проведенных исследований можно сделать следующие выводы:

1. Наивный Байесовский классификатор хоть и является простой моделью, но хорошо подходит для использования в анализе эмоциональной окраски текстовых сообщений. Имеет низкое время обучения, по сравнению с другими рассмотренными моделями.

2. Разработанная система обеспечивает более высокую точность классификации входных данных, за счет использования особенностей рассматриваемой соц. сети: языковых маркеров (смайлов), которые являются одними из самых точных индикаторов настроения автора сообщения.

3. На основании проведенных исследований установлено, что разработанную систему целесообразно применять в случаях, когда скорость обучения и классификации является ключевым параметром, так как она выше, чем у других методов. Но вместе с тем не уступает в точности результата.

4. Вариантами улучшения системы могут послужить использование других принципов к отношению между словами (n-граммы), также использование онтологий для замены неизвестных слов. Также имеется возможность использование взаимосвязи сообщений и ответов на него.

5. Система может быть использована аналитиками для определения отношения пользователей к какому-либо событию, продукту и т.д. Отдельной задачей стоит способы анализа собранной информации и возможности автоматически делать на ее основе выводы.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Разумов, А. В. Методы лингвистического анализа текстовых сообщений / А. В. Разумов // Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов (Минск, 2 – 6 мая 2017 г.). – Минск: БГУИР, 2017. – С. 43 – 44.