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

УДК 004.934.2

Федоров
Павел Андреевич

Система анализа эмоциональной окраски текста

АВТОРЕФЕРАТ

на соискание степени магистра
по специальности 1-40 80 04 – Информатика и технологии программирования

Научный руководитель
Анисимов В.Я.
к.ф.-м.н., доцент

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

Отнесение документов к определённой категории на основание его содержимого называется классификацией документа. Данная задача является одной из задач информационного поиска. Процесс классификации может осуществляться как полностью вручную, так и с помощью применения методов машинного обучения, в частности, сверточных нейронных сетей. Также следует отличать классификацию от кластеризации, где в последнем случае тексты тоже группируется по категориям, которые заранее определены. Задача классификации текстов становится все более востребованной в связи с постоянным ростом информации в интернете и необходимостью в ней ориентироваться. Например, задача классификации текстов применима к решению следующих задач:

– Персонификация рекламы.

– Разделение сайтов по тематическим каталогам.

– Борьба со массовой рассылкой корреспонденции рекламного характера.

– Распознавание тональности текстов.

В тоже время задача интеллектуального анализа текстовой информации, которая способна определять автора и пол автора текста, возраст, уровень образования, эмоциональное состояние автора в момент написания текста также является актуальной задачей. Под тональностью будем понимать эмоционально окрашенную лексику и эмоциональную оценку, выраженную автором относительно чего-либо. Анализ тональности имеет важное практическое применение:

– Оценка качества товаров и услуг на основании отзывов пользователей интернет-ресурсов.

– Противодействие экстремизму и терроризму.

– Анализ ситуации на фондовых рынках и прогнозирование волатильности финансовых активов.

– Составление текстов с заранее заданными эмоциональными характеристиками.

Социальные сети (Twitter, Facebook, LinkedIn) — пожалуй, самая популярная бесплатная доступная широкой общественности площадка для высказывания мыслей по разным поводам. Миллионы твитов (постов) публикуется ежедневно — там кроется огромное количество информации. В частности, Twitter широко используется компаниями и обычными людьми для описания состояния дел, продвижения продуктов или услуг. Twitter также является прекрасным источником данных для проведения интеллектуального анализа текстов: начиная с логики поведения, событий, тональности высказываний и заканчивая предсказанием трендов на рынке ценных бумаг. Там кроется огромный массив информации для интеллектуального и контекстуального анализа текстов.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы состоит в исследовании и разработке методов анализа эмоциональной окраски текста с применения сверточных нейронных сетей.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить классы задач интеллектуальной обработки естественного языка человека.
2. Рассмотреть основные понятия и произвести анализ существующих алгоритмов анализа эмоциональной окраски теста.
3. Спроектировать и реализовать систему для анализа эмоциональной окраски текста с применение сверхточной нейронной сети.
4. Провести демонстрацию и анализ разработанной системы.

Объектом исследования являются высказывания людей по заданной тематики, представленные в виде текстов на естественном языке и доступные через интернет.

Предметом исследования выступают задачи извлечения информации из высказываний людей для определения их эмоциональной окраски.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования методов машинного обучения для обработки естественного языка. В частности, применение сверточных нейронных сетей для анализа эмоциональной окраски текста.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Проделанная работа основана на росте популярности социальных сетей в интернете, а именного платформ микроблогинга. Данные платформы ежедневно представляют миллионы уникальных высказываний реальных людей по заданной тематике в виде текста, которые могут подлежать дальнейшему анализу.

Личный вклад соискателя

Результаты и методы, описанные в диссертации, получены соискателем лично. Вклад научного руководителя Анисимова В. Я. заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения были представлены в докладе на международной научной конференции «Информационные технологии и системы 2019» (Минск, Беларусь, Белорусский государственный. университет информатики и радиоэлектроники)

Публикации результатов диссертации

По теме диссертации опубликована 1 печатная работа.

Структура и объем диссертации

Диссертация состоит из введения, пяти глав, заключения, списка используемых источников и приложений. В первой главе проведен анализ предметной области и выявлены основные классы задачи обработки естественного языка. Вторая глава посвящена теоретическим сведениям. В данной главе рассмотрены основе понятия и термины. В третьей главе рассмотрены существующие алгоритмы анализа эмоциональной классификации, в частности анализ с помощью сверточных нейронных сетей. Четвертая глава посвящена проектированию и реализации системы. В пятой главе проведена демонстрация и анализ результатов.

Общий объем работы составляет 59 страниц, из которых основного текста – 50 страниц, 31 рисунок на 31 странице, список использованных источников на 15 наименований на 1 странице и 2 приложения на 9 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и основные направления исследования диссертационной работы, показаны актуальность и востребованность выбранной темы, кратко описаны исследуемые вопросы и обозначена практическая ценность работы.

В **первой главе** проведен обзор подходов и решаемых задач, которые применяются в рамках обработки естественного языка человека.

Интеллектуальная обработка текстов включает в себя множество задач и все из них так или иначе подвластны человеку и связаны с пониманием текста. В данной главе назовем и кратко прокомментируем основные задачи обработки текстов. Будем идти от простого к сложному и условно разделим их на три класса:

– Синтаксические.

– Требующие понимая текста.

– Требующие понимания и порождения нового текста.

Задачи первого класса можно условно назвать синтаксическими. Здесь задачи, как правило, очень хорошо определены и представляют собой задачи

классификации или задачи порождения дискретных объектов, и решаются многие из них сейчас уже довольно неплохо, например:

– Выделение границ предложения.
– Морфологическая сегментация.
– Частеречная разметка.
– Разрешение смысла.
– Другой вариант задачи о морфологии отдельных слов — стемминг.
– Распознавание именованных сущностей.
– Синтаксический.
– Разрешение кореференций.

Второй класс — это задачи, которые в общем случае требуют понимания текста, но по форме все еще представляют собой хорошо определенные задачи с правильными ответами (например, задачи классификации), для которых легко придумать не вызывающие сомнений метрики качества. К таким задачам относятся, в частности:

– Анализ тональности.
– Информационный.
– Ответы на вопросы.
– Выделение отношений или фактов.
– Языковые модели.

И наконец, к третьему классу отнесем задачи, в которых требуется не только понять уже написанный текст, но и породить новый. Здесь метрики качества уже не всегда очевидны, и мы обсудим этот вопрос ниже. К таким задачам относятся, например:

– Автоматическое.
– Диалоговые модели.
– Машинный перевод.
– Задачи порождения нового текста.

Вторая глава посвящена теоретическим сведениям. В данной главе рассмотрены определение эмоциональной окраски и основные классы текстовой информации, и виды классификации.

Термин «анализ эмоциональной окраски текста», который используется в данной работе используется, переведен из оригинального термина «sentiment analysis», который также можно перевести как, как «анализ тональности текста».

Согласно определению, анализ эмоциональной окраски текста – это задача автоматического анализа мнений и эмоционально окрашенной лексики, которые выражаются людьми в тексте.

В анализе эмоциональной окраски текста считается, что текстовая информация делится на два класса:

– Факты
– Мнения.

Главным понятием является определение мнения. Так как основной задачей анализа эмоциональной окраски является выявление мнений в тексте

и определить их свойств. Множество определяемых свойств, которые будут подвержены исследованию, будет зависеть уже от поставленной задачи. Например, анализу может подлежать определённый автор, которому принадлежит определенное мнение.

Мнения делятся на два типа:

– Простое мнение.

– Сравнение.

В современных системах автоматического определения эмоциональной оценки текста чаще всего используется одномерное эмотивное пространство: позитив или негатив (хорошо или плохо). Однако известны успешные случаи использования и многомерных пространств.

Основной задачей в анализе тональности является классификация полярности данного документа, то есть определение, является ли выраженное мнение в документе или предложении позитивным, негативным или нейтральным. Более развёрнуто, «вне полярности» классификация тональности выражается, например, такими эмоциональными состояниями, как «злой», «грустный» и «счастливый». Существуют несколько основных видов классификации эмоциональной окраски:

– По бинарной шкале.

– По многомерной шкале.

– Системы шкалирования.

– Субъективность/объективность.

В **третьей главе** рассмотрены существующие алгоритмы анализа эмоциональной окраски текста. Данную задачу обычно определяют как одну из задач компьютерной лингвистики, то есть подразумевается, что мы можем найти и классифицировать тональность, используя инструменты обработки естественного языка.

Анализ тональностей может быть разделен на две отдельные категории:

– Ручной.

– Автоматизированный анализ тональности.

Различия между этими двумя заключаются в точности и эффективности анализа. Эксперт, конечно же, гораздо корректнее обрабатывает входные данные, но при этом не может соревноваться с вычислительной машиной в объемах и скорости обрабатываемых массивов данных

Сделав большое обобщение, для автоматизированного анализа можно разделить существующие подходы на следующие категории:

– Подходы, основанные на правилах.

– Подходы, основанные на словарях.

– Машинное обучение с учителем.

– Машинное обучение без учителя.

На рисунке 1 представлена сравнительная характеристика наиболее распространенных методов анализа тональности.

	Точность	Автоматизация	Данные для обучения	Простота применения	Применяемость в коммерческих системах
Подход на правилах	наиболее точный	подлежит	не требует данных	–	+
Подход со словарем	не универсален	в рамках одной предметной области	требует данные	+	–
Машинное обучение	точный	автоматический	требует данные	+/–	+
Обучение без учителя	низкая точность	автоматический	не требует данных	+	+

Рисунок 1 – Сравнение методов анализа тональности

В **четвертой главе** описано проектирование и реализация системы. В данной главе подробно описано сбор и подготовка данных для анализа с помощью диаграмм классов и листинга кода.

Были рассмотрены основные способы получения данных твиттера с помощью основных API:

– Search API.

– Streaming API.

Оба данных API являются неделимой частью программирования при создании приложений по сбору коллекций твитов. Между двумя эти API существенная разница. Рассмотрим, когда и какой API следует использовать для максимальной эффективности.

Search API возвращается во времени, а Streaming API идет вперед. Если мы решили собирать твиты по какой-либо тематике и нам не с чего начать. Поисковый API нужно использовать тогда, когда мы хотим получить твиты за последние семь дней по нашей тематики. Это часто называют обратным заполнением. Если же мы хотим получать твиты начиная с данного момента и дальше, то для этого нам следует воспользоваться Streaming API. Благодаря этому API мы имеем возможность перехватывать все твиты в будущем.

Рассмотрен и использован для обучения корпус коротких текстов Юлии Рубцовой, сформированный на основе русскоязычных сообщений из Twitter. Он содержит 114 991 положительных, 111 923 отрицательных твитов, а также базу неразмеченных твитов объемом 17 639 674 сообщений.

Произведено проектирование сверхточной нейронной сети с использованием векторного отображения текстов. Наивысший показатель F_1=76.80% на валидационной выборке был достигнут на третьей эпохе обучения. Качество работы обученной модели на тестовых данных составило F_1=78.1% (рисунок 2).

Метка класса	Точность	Полнота	F_1	Количество объектов
Negative	0.78194	0.78243	0.78218	22457
Positive	0.78089	0.78040	0.78064	22313
avg / total	0.78142	0.78142	0.78142	44770

Рисунок 2 – Качество анализа тональности на тестовых данных.

В **пятой главе** проведена демонстрация работы разработанной системы и анализ полученных результатов. В данной главе приведен алгоритм работы с приложением и проведен анализ:

- Используемых платформ.
- Тональность высказываний.
- Количество лайков и ретвитов.
- График твитов на оси времени.
- Местоположение.
- Облако слов.
- Граф связей.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведен обзор и анализ задач обработки естественного языка, история развития и современные подходы к их решению. Рассмотрены основные понятия и термины в области эмоциональной окраски текста

2. Спроектирована и разработана система по сбору, хранения и анализу текстовых сообщений пользователей из социальных сетей по заданной тематике.

3. Продемонстрирована эффективность использования методов машинного обучения для интеллектуального анализа текста. Спроектирована и обучена сверточная нейронная сеть для классификации сообщений пользователей.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для решения задача анализа эмоциональной окраски текста. Данные знания могут быть использованы для ускорения процесса проектирования, разработки и улучшения качества классификации в новых системах.

2. Разработанная система готова к внедрению в существующие системы, занимающиеся обработкой текстов.

3. Результаты работы могут использоваться в качестве практического пособия для персонала, занимающегося интеллектуальным анализом текстовой информации в интернете для решения задач эмоциональной окраски текста.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Нестеренков, С. Н. Использование сверточных нейронных сетей для классификации и анализа тональности текстов / С. Н. Нестеренков, П. А. Федоров, В. А. Денисов // Информационные технологии и системы 2019 (ИТС 2019) : материалы междунар. науч. конф., Минск, 30 окт. 2019 г. / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. - Минск, 2019. - С. 248-249.