МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТВИТОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Головко Михаила Михайловича

Научный руководитель
доцент, к. э. н. _____ А. Р. Файзлиев

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. В мире количество информации только растет, и человек не способен усвоить всю эту информацию. С данной проблемой может справиться машинный анализ текста, который сможет ее классифицировать. В свою очередь классификация поможет из огромного количества информации выбрать именно то, что необходимо человеку.

Целью магистерской работы является исследование, модификация, программная реализация метода машинного обучения для анализа тональности твитов.

Объект исследования - машинное обучения при помощи мешка слов

Предмет исследования - социальная сеть Twitter и програмное обеспечение R.

Для достижения поставленных целей в работе необходимо решить следущие **задачи**:

— определить основные понятия, необходимые для машинного обучения;
— анализ методов машинного обучения;
— рассмотреть основные алгоритмы машинного обучения;
— определить инструменты для исследования;
— провести анализ выборки твитов;

Практическая значимость производимого исследования заключается в разработке программного обеспечения машинного обучения для анализа информации.

Структура и содержание магистерской работы. Работа состоит из введения, 8 разделов, заключения, списка использованных источников. Общий объем работы составляет 55 страниц.

Основное содержание работы

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи,отмечается практическая значимость полученных результатов.

В **первом** разделе рассматривается задача предстоящей работы и области применения. Автоматическое распознавание тональности текстов находит широкое применение в различных сферах деятельности человека. Примеры :

1. Маркетинговые исследования. Проводятся для разнообразных целей, включая изучение потребительских предпочтений, измерение степени удовлетворения потребностей потребителей, определение эффективности распространения продуктов или услуг.

2. Финансовые рынки. В каждом акционерном обществе существуют многочисленные публикации новостей, статьи, блоги и сообщения в Твиттере. Система анализа тональности может использовать эти источники для нахождения статей, в которых обсуждаются такие общества, и извлекать отзывы, что позволит создать автоматическую торговую систему. Одной из таких систем является «The Stock Sonar» . Система показывает графически ежедневные позитивные и негативные настроения о каждой акции рядом с графиком цены акции. По настроениям предсказывается дальнейший рост или падение цены акции.

3. Рекомендательные системы. Анализируются отзывы и обзоры различных продуктов с целью помощи покупателям при выборе товара. Например, система не будет рекомендовать продукт, если он получил много отрицательных отзывов.

4. Анализ новостных сообщений. Анализируются новостные ресурсы на предмет тональности сообщений относительно различных персон и событий.

5. Политологические исследования. Собираются данные о политических взглядах населения. Это может иметь существенное значение для кандидатов, выступающих от разных партий. Такой подход применяется организаторами предвыборной кампании для выявления того, что думают избиратели в отношении различных проблем, и как они связывают эти проблемы со словами и действиями кандидатов.

6. Социологические исследования. Анализируются данные из социаль-

ных сетей, например для выявления религиозных взглядов или различия между мужчинами и женщинами в употреблении эмоционально-окрашенных слов в сообщениях.

7. Поддержка поисковых систем и систем извлечения информации. В таких системах анализ может служить для отделения фактов от мнений.

8. Анализ обратной связи от пользователей. При диалоге с пользователем система распознает его эмоции, и при помощи обратной связи может реагировать в соответствии с ними.

9. Анализ экстремистских ресурсов. Анализируются Интернет-ресурсы экстремистского содержания на предмет подозрительной активности.

10. Психологические исследования. Определение депрессии у пользователей социальных сетей.

Во **втором** разделе производится анализ видов машинного обучения

Machine Learning (ML, с английского – машинное обучение) — это методики анализа данных, которые позволяют аналитической системе обучаться в ходе решения множества сходных задач. Машинное обучение базируется на идее о том, что аналитические системы могут учиться выявлять закономерности и принимать решения с минимальным участием человека.

Давайте представим, что существует программа, которая может проанализировать погоду за прошедшую неделю, а также показания термометра, барометра и анемометра (ветрометра), чтобы составить прогноз. 10 лет назад для этого написали бы алгоритм с большим количеством условных конструкций If (если):

1 Если дует сильный ветер, возможно, он нагонит облака.

2 Если среди облаков есть тучи, будет пасмурно.

3 Если температура воздуха упала, но выше нуля, то пойдет дождь, если ниже – снег.

От программиста требовалось описать невероятное количество условий, чтобы код мог предсказывать изменение погоды. В лучшем случае использовался многомерный анализ данных, но и в нем все закономерности указывались вручную. Но даже если такую программу называли искусственным интеллектом, это была лишь имитация.

Машинное обучение же позволяет дать программе возможность само-

стоятельно строить причинно-следственные связи. ИИ получает задачу и сам учится ее решать. То есть компьютер может проанализировать показатели за несколько месяцев или даже лет, чтобы определить, какие факторы оказывали влияние на изменение погоды.

Пример от Google DeepMind. Программа получала информацию от виртуальных рецепторов, а ее целью было перевести модель из точки А в точку Б. Никаких инструкций по этому поводу не было – разработчики лишь создали алгоритм, по которому программа обучалась. В результате она смогла самостоятельно выполнить задачу.

ИИ, словно ребенок, пробовал разные методы, чтобы найти тот, который лучше всего поможет добиться результата. Также он учитывал особенности моделей, заставляя четвероногую прыгать, человекообразную – бежать. Также ИИ смог балансировать на двигающихся плитах, обходить препятствия и перемещаться по бездорожью.

В **третьем** разделе рассматриваются виды машинного обучения

Всего есть 3 вида машинного обучения:

1 С учителем (Supervised machine learning). Исходя из введенных данных, программа может построить причинно-следственные связи и помочь учащимся с профориентацией. Например, она может предположить, что ученик может поступить на филологический факультет потому, что получил высший балл по литературе и имеет гуманитарный склад ума. Так же ученик со склонностью к техническим наукам и хорошими результатами по геометрии может смотреть в сторону профессии инженера-проектировщика.

2 Без учителя (Unsupervised machine learning). Эта программа получила задание от разработчика – добраться до точки Б. Но она не знала, как это сделать – ей даже не показали, как выглядит ходьба, но это не помешало ИИ выполнить задачу.

3 Глубокое обучение (Deep learning). Глубокое обучение может быть как с учителем, так и без, но оно подразумевает под собой анализ Big Data – настолько большой информации, что одного компьютера будет недостаточно. Поэтому Deep Learning использует для работы нейронные сети.

В **четвертом** разделе был рассмотрены более подробно предлагаемые подходы, методов классификации текстов :

Наивный метод Байеса

Пусть P(c|d) - Вероятность того, что документ d принадлежит классу c. В наивной байесовской классификации задача состоит в нахождении наиболее вероятного класса cm для данного документа d

$$c_m = argmax \;\; P(C|D) \tag{1}$$

используя формулу Байеса, можно переписать выражение для P(c|d) ,Где

$$c_m = argmax \frac{P(d|c)P(C)}{P(d)} = argmax P(d|c)P(c) \tag{2}$$

знаменатель P(d) не зависит от c и, следовательно, не влияет на нахождение максимума, поэтому его можно опустить;

Деревья

Идея данного метода состоит в построении разрешающего дерева на «обучающем» наборе документов. Дерево строится по следующему правилу: выбираем терм, документы, содержащие этот терм кладем направо, остальные налево. Таким образом, документы разделились на две непересекающиеся коллекции. Для каждой коллекции выбирается новый терм и повторяется описанная выше процедура. Так продолжается до тех пор, пока не получится однородная коллекция, то есть коллекция, в которой либо все документы соответствуют категории, либо все документы соответствуют ее дополнению.

Random Forest

Алгоритм Random Forest – ансамблевый метод машинного обучения, который использует ансамбль деревьев решений. Он основывается на основных подходах бэггинга и выбора случайных подмножеств признаков. Этот алгоритм позволяет достичь высокой точности классификации. Деревья в ансамбле строятся друг от друга независимо. [9]

Метод опорных векторов (SVM)

Метод опорных векторов (Support Vector Machine, SVM) – метод, в котором основой является построение (оптимальной) разделяющей гиперплоскости. Некоторая выборка линейно разделима, если в ней возможно получить

(построить) линейный пороговый классификатор:

$$sign(\sum_{i=1}^{m} w_i * x^i - w_0) = sign(< w, x > -w_0) \quad (3)$$

Модель мешок слов

Для реализации методов машинного обучения существует классическая модель “Мешок слов” (Bag of Words). Формальная постановка задачи выглядит следующим образом:

Пусть f1,...,fm множество, состоящее из m признаков (атрибутов), которые могут присутствовать в документе; пусть ni(d) – это количество вхождений признака fi в документ d. Далее каждый документ d представляется в виде вектора следующим образом:

$$d = (n1(d), n2(d), ..., nm(d)) \quad (4)$$

Выделяют два основных типа атрибутов:

a. Частотные - каждое значение в векторе d соответствуют количеству вхождений признаков в документ d;

b. Бинарные (наличия/отсутствия), каждое значение в векторе d бинарное (true/false или 0/1) и отражает факт присутствия признака.

Метод автоматического порождения гипотез (ДСМ)

В **пятом** разделе был более подробно рассмотрен метод автоматического порождения гипотез (ДСМ). ДСМ-метод - это метод автоматического порождения гипотез. Был предложен В. К. Финном в конце 1970-х гг. Свое название метод получил от инициалов известного английского философа, логика и экономиста Джона Стюарта Милля. ДСМ-метод представляет собой формализацию правдоподобных рассуждений, которая позволяет на основе анализа имеющихся данных формировать гипотезы о том, какими свойствами могут обладать рассматриваемые объекты. ДСМ-метод - это синтез трех познавательных процедур - эмпирической индукции, структурной аналогии и абдукции. В данной работе мы рассмотрим только два этапа этого метода - этапы индукции и аналогии.

В **шестом** разделе были подробно рассмотрены метрики качества

Правильность и ошибочность

На практике не бывает систем, абсолютно точно определяющих правильные соотношения классов и принадлежащих им объектов. Классификатор будет работать с ошибками относительно тестовой выборки. Для оценки успешности сопоставления классов и объектов используется метрика правильности:

$$A = \frac{TP + TN}{FN + FP + TN + TP} \tag{5}$$

где в числителе - количество объектов, по которым классификатор принял правильное решение, в знаменателе - размер классифицируемой выборки. Для оценки процента ошибок используется метрика ошибочности:

$$A = \frac{FP + FN}{FN + FP + TN + TP} \tag{6}$$

Точность и полнота.

Точность P и полнота R являются метриками, которые используются при оценке большей части систем анализа информации. Иногда они используются сами по себе, иногда в качестве базиса для производных метрик, таких как F1-мера.

Метрика точности определяется формулой

$$P = \frac{TP}{TP + FP} \tag{7}$$

Метрика точности характеризует, сколько полученных от классификатора положительных ответов являются правильными. Чем больше точность, тем меньше число ложных попаданий. Но эта метрика не дает представление о том, все ли правильные ответы вернул классификатор. Для этого существует метрика полноты, определяемая формулой

$$R = \frac{TP}{TP + FN} \tag{8}$$

Седьмой раздел посвящен предметной области.

Twitter — пожалуй, самая популярная бесплатная доступная широкой общественности площадка для высказывания мыслей по разным поводам. Миллионы твитов (постов) ежедневно — там кроется огромное количество информации. В частности, Twitter широко используется компаниями и обычными людьми для описания состояния дел, продвижения продуктов или услуг.В Twitter можно узнать, что происходит в мире и о чем говорят люди прямо сейчас. Твиттер доступен с компьютера или мобильного устройства. 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.

API-интерфейсы Twitter — это способ общения компьютерных программ друг с другом для запроса и предоставления информации. Для этого программное приложение вызывает так называемую конечную точку — адрес, соответствующий определенному типу информации (как правило, конечные точки уникальны, подобно телефонным номерам). Twitter предоставляет доступ к некоторым своим службам с помощью API-интерфейсов, чтобы программисты могли разрабатывать программное обеспечение, тесно взаимодействующее с Twitter.

Язык R. R — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом в рамках проекта GNU.Широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ.

Восьмой раздел посвящен разработке на языке R.

Первый шаг — зарегистрироваться на портале разработчиков для Twitter и пройти авторизацию.

Нам потребуется

```
api_key = "Ваш ключ API"
api_secret = "Ваш api_secret пароль"
access_token = "Ваш токен доступа"
```

```
access_token_secret = "Ваш пароль токена доступа"
```

После получения этих данных авторизируемся для получения доступа к Twitter API:

Следующий шаг — загрузить массив позитивных и негативных тональных слов (словарь) в рабочую папку R. Слова мы можем достать из переменных, positive и negative, как показано ниже.

```
positive=scan('positive-words.txt',what='character',comment.char=';')
negative=scan('negative-words.txt',what='character',comment.char=';')
```

Следующий шаг — задать строку поиска по twitter-сообщениям и присвоить ее значение переменной, findfd. Количество твитов, которые будут использованы для анализа, присваивается другой переменной, number. Время на поиск по сообщениям и извлечение информации зависит от этого числа. Медленное соединение с Интернет или сложный поисковый запрос могут привести к задержкам.

```
findfd= "CyberSecurity"
number= 5000
```

Основные результаты

1. Рассмотрены некоторые подходы машинного обучения. Выбран машинное обучение с учителем.
2. Рассмотрены несколько стратегий анализа текстов.
3. Изучена необходимая информация о машинном обучении.;
4. Построена и проанализированна модель анализа тональности твитов. Определенны тональности 5000 твитов на основе поиска по «CyberSecurity».