МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

Интеллектуальный анализ текста

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Ал-джарвани Мустафа Али

Научный руководитель
доцент, к. ф.-м. н., доцент _____ Р.Н. Фадеев

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

Саратов 2022

Технология интеллектуального анализа данных помогает извлекать полезную информацию из различных баз данных. Хранилища данных оказались хорошими для числовой информации, но неудачными, когда дело дошло до текстовой информации. 21 век вывел нас за пределы ограниченного объема информации в сети. Это хорошо в том смысле, что больше информации обеспечит большую осведомленность и лучшее знание. Интеллектуальный анализ текстовых данных относится к процессу извлечения интересных и нетривиальных шаблонов или знаний из текстовых документов. Поскольку интеллектуальный анализ текста — это извлечение полезной информации из текстовых данных, он также известен как интеллектуальный анализ текстовых данных или обнаружение знаний из текстовых баз данных. Трудно найти точные знания в текстовых документах, чтобы помочь пользователям найти то, что они хотят [2,3].

В настоящее время большая часть информации в бизнесе, промышленности, правительстве и других учреждениях хранится в текстовой форме в базе данных, и эта текстовая база данных содержит полуструктурированные данные. Документ может содержать в основном неструктурированные текстовые компоненты, такие как аннотация, а также несколько структурированных полей, таких как заголовок, имя автора, дата публикации, категория и т. д. Интеллектуальный анализ текста — это разновидность области, называемой интеллектуальным анализом данных, которая пытается найти интересные закономерности в больших базах данных. Большое количество исследований, проведенных по моделированию и реализации полуструктурированных данных в недавних исследованиях баз данных. На основе этих исследований были разработаны методы поиска информации, такие как методы индексации текста, для обработки неструктурированных документов. При традиционном поиске пользователь обычно ищет уже известные термины, написанные кем-то другим. Проблема в том, что это не соответствует потребностям пользователей. Это цель интеллектуального анализа текста, чтобы обнаружить неизвестную

информацию, которая не известна и еще не записана [1].

Процесс интеллектуального анализа текста начинается со сбора документов из различных ресурсов. Инструмент интеллектуального анализа текста извлекает конкретный документ и предварительно обрабатывает его, проверяя формат и наборы символов. Затем документ будет проходить этап анализа текста. Анализ текста — это семантический анализ для получения высококачественной информации из текста. Доступно множество методов анализа текста; в зависимости от цели организации могут использоваться комбинации методов. Иногда методы анализа текста повторяются до тех пор, пока информация не будет извлечена. Полученная информация может быть помещена в информационную систему управления, предоставляя пользователю этой системы обширный объем знаний [4].

Актуальность темы исследования – обусловлена необходимостью применить методы интеллектуального анализа данных для обнаружения фейковых новостей, что позволит с большей вероятностью прогнозировать возможные фейковых новостей.

Предмет исследования – Модели и алгоритмы обработки данных, для обнаружения фейковых новостей.

Цель и задачи исследования. повышение эффективности процесса Алгоритмы прогнозирования для обнаружения фейковых новостей с помощью анализа текста. Для достижения сформулированной цели были поставлены и решены следующие задачи:

1. Исследование методов анализа текста и обработки естественного языка (NLP)

2. Изучение общего процесса интеллектуального анализа текста.

3. Идентификация проблем при проведении интеллектуального анализа

текста

4. Исследования приложений интеллектуального анализа текстов

5. Изучение инструментальных средств и методов новостного анализа

6. Изучение этапов обнаружения фейковых новостей

7. Сравнение алгоритмов прогнозирования для обнаружения поддельных новостей

Объект исследования – Интеллектуальный анализ текста для обнаружения фейковых новостей.

Научная новизна исследования – состоит в использовании методов и алгоритмов интеллектуального анализа данных для обнаружения фейковых новостей.

Практическая значимость – заключается в анализе новостей с использованием методов и алгоритмов интеллектуального анализа данных, который позволит с большей вероятностью прогнозировать возможные фейковых новостей.

Методы исследования – При решении указанных задач использовались: методы дискретной математики, интеллектуальной обработки данных, оценки эффективности алгоритмов.

Структура работы определена задачами исследования, логикой раскрытия темы. Работа состоит из введения, трех глав, заключения и списка использованных источников.

Во введении обосновывается выбор темы, актуальность исследования, определяются объект и предмет, цели и задачи, методы исследования, а также практическая значимость работы.

Первая глава диссертационной работы посвящена аналитическому обзору Интеллектуальный анализ текста, и Области интеллектуального анализа текста, Общий процесс интеллектуального анализа текста, Обработка естественного языка и текстовая аналитика ,Методы и модели, используемые в интеллектуальном анализе текста ,Виды извлекаемой информации, Приложения интеллектуального анализа текстов.

Во второй главе работы рассмотрены описание инструментальных средств и методов новостной анализ, Фейковые новости, Типы фейковых новостей, Фейковые новости в социальных сетях, существующие методы автоматического обнаружения фейковых новостей, Этапы обнаружения фейковых новостей, и также алгоритмы классификации.

В третьей главе рассмотрены Фейковые новости, Векторизатор TF-IDF, Обнаружение поддельных новостей с помощью python, Шаги для обнаружения поддельных новостей с помощью python.

В заключении сформулированы основные результаты диссертационной работы, отмечены ее научная значимость и практическая ценность, определены перспективы дальнейшей работы.

В первой главе «Интеллектуальный анализ текста» рассматривается понятие целом интеллектуальный анализ текста состоит из анализа текстовых документов путем извлечения ключевых фраз, понятий и т. д. и подготовки обработанного текста для дальнейшего анализа с использованием методов интеллектуального анализа данных. В этой главе обсуждались концепция, процесс, методы, инструменты, проблемы и модели, используемые в интеллектуальном анализе текста, и приложения интеллектуального анализа текста, которые могут применяться во множестве областей, таких как Веб-добыча, медицина, фильтрация резюме и т. д. Это также исследование в области

интеллектуального анализа текста и мотивация для дальнейшего изучения. интеллектуальный анализ текста включает в себя ряд действий, которые необходимо выполнить для эффективного извлечения информации. как показано на рисунке 1.

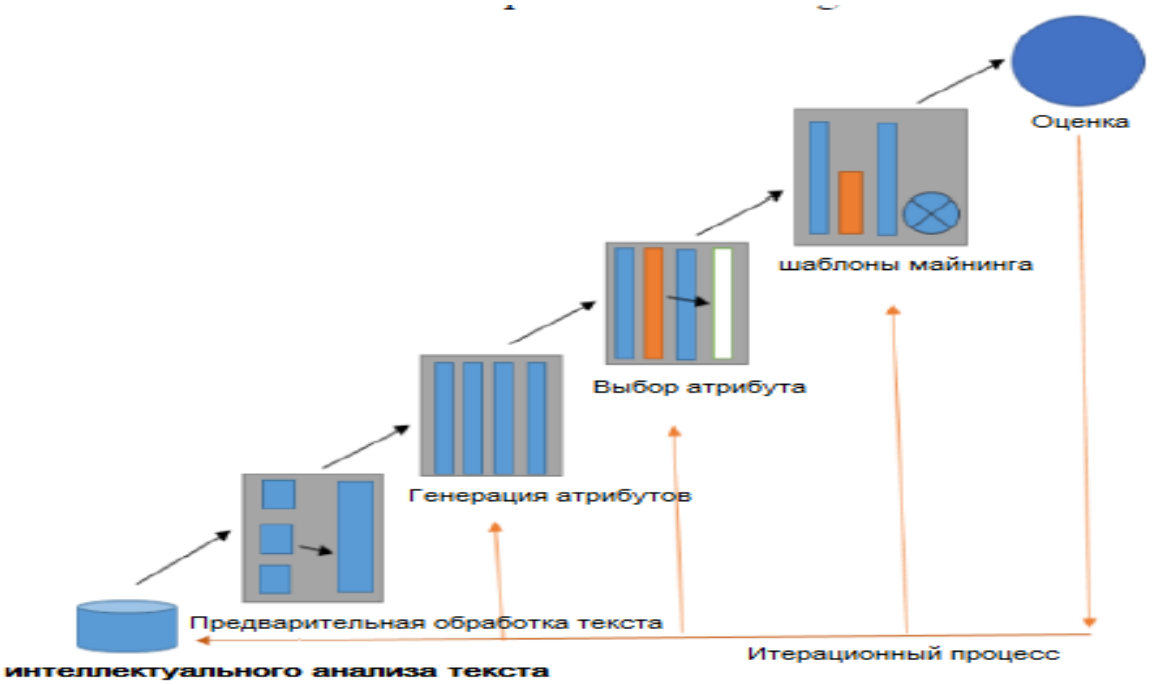

Рисунок 1. Действия/процесс анализа текста

Во второй главе «Инструментальных средств и методов новостно анализ» рассмотрим в проблеме фейковых новостей уделяется огромное внимание. По мере появления новых статистических данных и исследований, те, кто распространяет такие новости в интернете, составляют в мире около 62% взрослых. В настоящем исследовании я предложил модель обнаружения фейковых новостей с использованием метода извлечения признаков TF-IDF. Кроме того, мы используем четыре различных метода алгоритмов. Реализованная модель достигла максимальной точности в случае использования классификатора случайного леса. Максимальная оценка точности составила 93%. Достигнутый результат лучше, чем перечисленные родственные работы, поэтому использование этого алгоритма повышает точность классификации.

Общее обсуждение этих шагов проиллюстрировано в этом разделе шагом

является выбор подходящего набора данных фейковых новостей и предварительная обработка набора данных. После этого применяется TF-IDF для извлечения признаков слов после разделения набора данных с использованием перекрестной проверки (10-кратной). Следующим шагом является классификация набора данных с использованием классификаторов , как показано на рисунке 2.

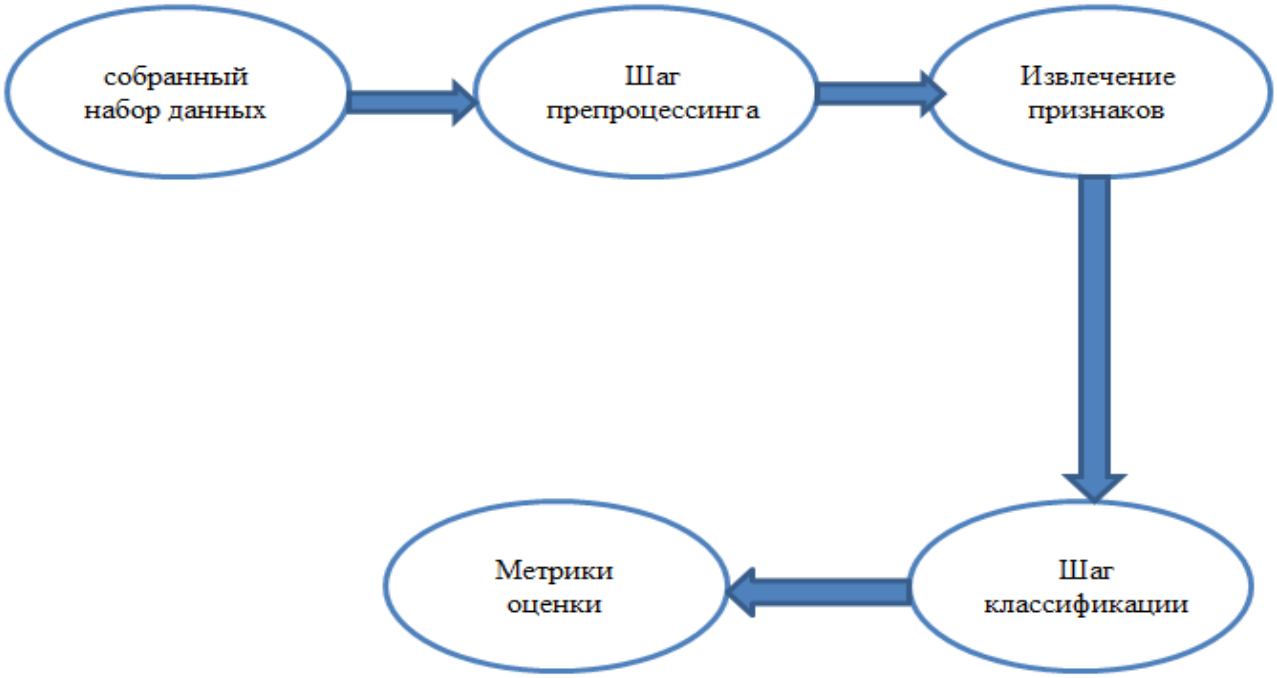

Рисунок 2. Рабочий этап обнаружения фейковых новостей.

В третьей главе «Обнаружения поддельных новостей с помощью python» обсуждаются результаты, полученные с помощью алгоритмов логистическая регрессия (LR), Дерева решений (DT), Повышения градиента (GBC), Случайного леса (RFC). И исследований шаги для обнаружения поддельных новостей с помощью python.

В этом раздел используется набор данных (фальшивые и настоящие новостные статьи в формате .CSV), собранный из интернета новостные сайты (RT, BBC, ABC) и сайты социальных сетей. Этот набор данных содержит около 44921 записей из различных статей, найденных в Интернете, и их атрибуты (текст, данные, заголовок и метка). После применения шага предварительной обработки размер набора данных стал 43251 запись. Эти данные разделены на

два класса: 19748 настоящих новостей и 23503 фейковых новостей. В этом работе для выявления классификаторов фейковых новостей используются только два признака (текст, метка). Нулевая метка присваивается ложным новостям (или фейкам), а единица присваивается реальным новостям. После применения кода полученные результаты, как показано на следующем рисунке:

```
news = str(input())
manual_testing(news)
```

```
The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Trump, @realDonaldTru
mp and @POTUS.  The opinions expressed are his own.Â Reuters has not edited the statements or confirmed their accurac
y.  @realDonaldTrump : - Together, we are MAKING AMERICA GREAT AGAIN! bit.ly/2lnpKaq [1814 EST] - In the East, it cou
ld be the COLDEST New Yearâ€™s Eve on record. Perhaps we could use a little bit of that good old Global Warming that
our Country, but not other countries, was going to pay TRILLIONS OF DOLLARS to protect against. Bundle up! [1901 EST]
-- Source link: (bit.ly/2jBh4LU) (bit.ly/2jpEXYR)


LR Prediction: True News
DT Prediction: True News
GBC Prediction: True News
RFC Prediction: True News
```

Из результатов выше таким образом отметим, что сила алгоритма классификатор случайного леса (RFC) наиболее точна в предсказании, а точнее в новостях на 93% и алгоритм логистическая регрессия (LR) на 65%. Наименее мощным в прогнозировании является алгоритм Классификатор дерева решений (DT) и классификатор повышения градиента (GBC) на 53%.

Заключение.

Проведенное исследование позволяет сделать следующие выводы.

В настоящее время технология интеллектуального анализа текста широко используется для решения различных задач в сфере бизнеса, научных исследований, управления, разведки и безопасности, прогнозирования фейковых новостей, которые связаны с текстовым анализом текстовой информации. Дан краткий обзор четырех алгоритмов, реализующих методы вышеназванной технологии, которые могут помочь исследователю при работе с некорректной текстовой информацией, в том числе при решении бизнес-задач. И сравнение функционала упомянутых алгоритмов. Поэтому Использование информации и знаний, извлеченных из большого количества данных, приносит пользу многим приложениям, таким как анализ рынка, управление бизнесом и анализ новостей. Во многих приложениях базы данных хранят информацию в текстовом виде, поэтому интеллектуальный анализ текста является одной из наиболее актуальных областей исследований. Извлечение требуемой пользователем информации является сложной задачей. Интеллектуальный анализ текста — важный этап процесса обнаружения знаний. Поэтому для этой цели применяются различные методы, такие как классификация, кластеризация и извлечение информации. Разработан ряд методов категоризации текста.

Из вышеприведенных результатов делаем вывод, что:

1. Классификатор случайного леса лучше других классификаторов.

2. логистическая регрессия лучше, чем дерева решений, в точности классификации набора данных фейковых новостей.

3. Случайный лес больше подходит для больших наборов данных, потому что деревья решений генерируются случайным образом и зависят от голосования между результатами, чтобы выбрать лучший результат.

4. Шаги предварительной обработки с использованием нашего набора данных дают лучшие результаты. Эти шаги оказали существенное влияние на повышение точности классификации.

5. Тип собранного набора данных (поддельные и настоящие новостные

статьи набора данных) также оказывает значительное влияние на точность классификации этой работы.

	Истинно положительные	Истинно отрицательные	Ложно положительные	Ложно отрицательные
Логистическая регрессия	65%	60%	35%	40%
Деревья решений	53%	51%	47%	49%
Классификатор повышения градиента	55%	53%	45%	47%
Классификатор случайного леса	93%	90%	7%	10%

СПИСОК ЛИТЕРАТУРЫ

1. Divya NASA, “Text Mining Techniques- A Survey” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.

2. Rashmi Agrawal and Mridula Batra, “A Detailed Study on Text Mining Techniques” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

3. Пескова О. В. Алгоритмы классификации полнотекстовых документов // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. — М.: МИЭМ (Московский государственный институт электроники и математики), 2011. — С. 170—212.

4. Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M. W. Berry. — 2004. — Springer, 2003. — 261 p.