

УДК 004.056.5(075.8)

ББК 32.97я73

ЦЗ49

Печатается по решению кафедры информационно-аналитических систем безопасности Института компьютерных технологий и информационной безопасности Южного федерального университета (протокол № 9 от 25 мая 2023 г.)

Рецензенты:

доктор технических наук, профессор,
заведующий кафедрой информатики Таганрогского института
имени А. П. Чехова (филиал) РГЭУ (РИНХ) *Я. Е. Ромм*
доктор технических наук, профессор, профессор кафедры
информационно-аналитических систем безопасности
Южного федерального университета *А. В. Боженюк*

Целых, А. Н.

ЦЗ49

Выявление инцидентов информационной безопасности и мошеннических транзакций методами машинного обучения : учебное пособие / А. Н. Целых, Э. М. Котов ; Южный федеральный университет. – Ростов-на-Дону ; Таганрог : Издательство Южного федерального университета, 2023. – 116 с.

ISBN 978-5-9275-4515-5

Пособие посвящено рассмотрению подходов по применению методов обработки естественного языка sentiment analysis для обнаружения угроз информационной безопасности в сети интернет, а также выявлению мошеннических транзакций с помощью методов машинного обучения.

Пособие предназначено для студентов высших учебных заведений, обучающихся по специальности 10.05.04 – Информационно-аналитические системы безопасности (специализация: «Автоматизация информационно-аналитической деятельности») по курсу «Математические методы анализа больших данных» и направлению 10.03.01 – Информационная безопасность (направленность: «Информационно-аналитические системы безопасности») по курсу «Модели и методы инженерии знаний».

УДК 004.056.5(075.8)

ББК 32.97я73

ISBN 978-5-9275-4515-5

© Южный федеральный университет, 2023

© Целых А. Н., Котов Э. М., 2023

© Оформление. Макет. Издательство

Южного федерального университета, 2023



СОДЕРЖАНИЕ

ВВЕДЕНИЕ	6
1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА	10
1.1. Обработка естественного языка NLP	10
1.2. Понятие Sentiment Analysis и подходы к классификации тональности текстов	11
1.3. Этапы процесса анализа текстов	14
1.4. Подходы к классификации тональности текста	14
1.5. Виды анализа настроений	17
1.6. Существующие инструменты для анализа настроений	19
1.7. Предварительная обработка текста	20
1.8. Общие методы обработки естественного языка, исполь- зуемые для предварительной обработки текстов	22
1.9. Применение Brand24	30
2. АЛГОРИТМЫ И ИНСТРУМЕНТЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА	34
2.1. Сервис Brand24	34
2.2. Kaggle notebook	35
2.3. Библиотека NLTK	36
2.4. Датасет Sentiment140	37
2.5. Инструмент векторизации TfidfVectorizer	39
2.6. Определение тональности текста с помощью алгоритмов машинного обучения	40
2.7. Метрики для оценки методов машинного обучения	44
3. ТЕСТИРОВАНИЕ МЕТОДОВ И АНАЛИЗ РЕЗУЛЬТАТОВ	47
4. МЕТОДЫ ОБНАРУЖЕНИЯ МОШЕННИЧЕСКИХ ТРАНЗАКЦИЙ	56

4.1. Проблемы обнаружения мошенничества в финансовой сфере	56
4.2. Этапы обнаружения мошеннических транзакций методами машинного обучения	57
4.3. Методы обнаружения мошеннических транзакций	59
4.4. Инструменты обнаружения мошеннических транзакций	60
4.5. Алгоритмы машинного обучения, используемые для обнаружения мошеннических транзакций	62
4.5.1. Деревья решений	62
4.5.2. <i>Random Forest</i>	63
4.5.3. Метод опорных векторов	64
4.5.4. Метод <i>k</i> -ближайших соседей	65
4.5.5. Наивный Байес	66
4.5.6. <i>Adaptive Boosting</i>	66
4.5.7. <i>CatBoost</i>	67
4.5.8. <i>XGBoost</i>	68
4.5.9. <i>LightGBM</i>	68
4.6. Метрики, используемые для оценки методов машинного обучения	69
5. ПРОЕКТИРОВАНИЕ МОДЕЛИ ОБНАРУЖЕНИЯ МОШЕННИЧЕСКИХ ТРАНЗАКЦИЙ	71
5.1. Описание используемого набора данных	71
5.2. Описание используемых библиотек Python	73
5.2.1. Библиотека <i>Pandas</i>	73
5.2.2. Библиотека <i>NumPy</i>	73
5.2.3. Библиотека <i>Matplotlib</i>	74
5.2.4. Библиотека <i>Seaborn</i>	74
5.2.5. Графическая библиотека <i>Plotly</i>	75
5.2.6. Библиотека <i>Scikit-learn</i>	75
5.3. Описание выбранных методов машинного обучения	75
5.3.1. Метод <i>Random Forest</i>	75
5.3.2. Метод <i>AdaBoost</i>	77
5.3.3. Методы <i>XGBoost</i> , <i>CatBoost</i> и <i>LightGBM</i>	80
5.4. Обобщение на другие функции потерь	82

Содержание

6. ПРИМЕР РЕАЛИЗАЦИИ ВЫЯВЛЕНИЯ МОШЕННИЧЕСКИХ ТРАНЗАКЦИЙ	94
6.1. Предварительный анализ данных	94
6.2. Определение параметров прогнозирования и целевых зна- чений	97
ЗАКЛЮЧЕНИЕ	110
СПИСОК ЛИТЕРАТУРЫ	112