Интеллектуальный алгоритм поиска текстов экстремистской направленности

Д.О. Фадеев[1], В.С. Мошкин[1], И.А. Андреев[1]

[1]*Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027*

Аннотация

В работе приводится описание алгоритма нахождения текстов экстремистской направленности в социальных сетях путем их классификации. Алгоритм включает 3 этапа: предобработка, фильтрация с использованием словарей и классификация текстов на базе байесовского алгоритма. Описана архитектура информационной системы и краткие результаты проведенных экспериментов.

Ключевые слова

Анализ текста, социальная сеть, байесовский классификатор, классификация

1. Введение

В настоящее время участились случаи привлечения пользователей социальных сетей к уголовной и административной ответственности за публикуемую (post) и распространяемую (repost) информацию, носящую характер экстремизма или терроризма. Актуальность данной работы обусловлена ужесточением российского законодательства в области распространения информации в сети Internet и социальных сетях. При этом пользователь социальной сети зачастую не имеет верного представления о том, что его текстовое сообщение может являться нелегальным с точки зрения законодательства о борьбе с экстремизмом, терроризмом и прочими правонарушениями.

2. Алгоритм интеллектуальной фильтрации текстовых сообщений

В рамках данного исследования был разработан алгоритм автоматической проверки текстовых сообщений пользователя в социальной сети на наличие материалов, запрещенных к опубликованию (в частности, экстремистской направленности). Данный алгоритм включает следующие этапы:

1. Предобработка текстовых сообщений, извлекаемых из социальных сетей.

Предобработка включает проведение графематического и морфологического анализа, а также лемматизацию слов с использованием грамматического словаря системы Lucene [1].

2. Поверхностная классификация на основе тезауруса опасных фраз и высказываний.

Результатом данного этапа является набор предложений и коротких текстов, которые предварительно помечены опасными.

3. Классификация текстов методом наивного байесовского классификатора.

Была выбрана модель «мешок слов». Согласно данной модели, любое текстовое сообщение представляется в виде множества слов и словосочетаний [2]. По наличию в тексте определенных слов, неявно соответствующих классу, происходит классификация. Вероятность отнесения текста d_i к классу k определяется следующей моделью:

$$P(k|d_i) = \frac{P(d_i|k)*P(k)}{P(d_i)},$$

где $P(d_i|k)$ представляет собой вероятность нахождения документ d_i во множестве документов k; $P(k)$ предполагает безусловную вероятность класса k в обучающей выборке; $P(d_i)$ – безусловная вероятность текста d_i в корпусе текстов обучающей выборки [3].

Наиболее вероятный класс для текста определяется, используя оценку апостериорного максимума [4]:

$$k_{map} = \arg\max_{k \in K} \left[\log \mathrm{P}(\mathrm{k}) + \sum_{s=1}^{n} \log \mathrm{P}(w_s \mid \mathrm{k}) \right]$$

Для реализации предложенного алгоритма был разработан модуль классификации текста на языке программирования python, представляющий из себя отдельный REST сервис.

Общая архитектура системы представлена на рисунке 1.

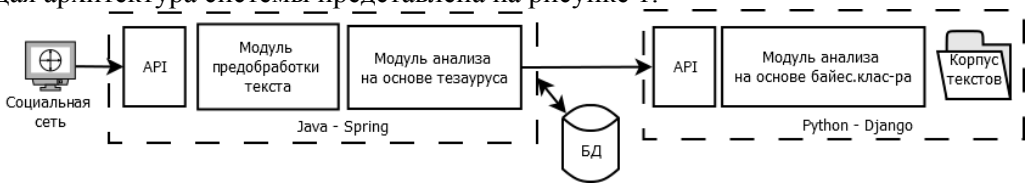

Рисунок 1: Архитектура системы классификации текста

Для обучения наивного байесовского классификатора были использованы размеченные тексты, полученные в предыдущих исследованиях. Из них было сформировано два корпуса текстов, 190 документов опасных постов и 820 документов подозрительных постов. Разделение на обучающую и тестовую выборку составило 9 к 1 соответственно.

В результате экспериментов были получены следующие результаты:

- Точность классификатора на тестовой выборке составила чуть более 84%.
- В ходе нагрузочного тестирования время ожидания при 100 запросах изменилось с 17 секунд до 28, Увеличение времени обработки одного запроса увеличилось на 0,11 секунд, что является приемлемым.

3. Заключение

Разработанная в рамках исследования система является набором микросервисов. Она позволяет осуществлять трехэтапную классификацию текстов. Была повышена точность классификации без значительного увеличения времени обработки, что позволит реализовать дополнительные этапы проверки для уточнения результатов определения класса текстового ресурса.

4. Благодарности

Работа выполнена при финансовой поддержке РФФИ, гранты № 18-47-730035 и 18-47-732007.

5. Литература

[1] Павлыгин, Э.Д. Разработка программного комплекса для интеллектуального анализа социальных медиа / Э.Д. Павлыгин, А.Г. Подлобошников, Р.А. Савинов, Н.Г. Ярушкина, А.М. Наместников, А.А. Филиппов, А.А. Романов, В.С. Мошкин, Г.Ю. Гуськов, М.С. Григоричева. – Автоматизация процессов управления. – 2019. – № 2(56). – С. 23-36. DOI: 10.35752/1991-2927-2019-2-56-23-36.

[2] Афанасьева, Т.В. Онтологический и нечеткий анализ слабоструктурированных информационных ресурсов / Т.В. Афанасьева, В.С. Мошкин, А.М. Наместников, И.А. Тимина, Н.Г. Ярушкина. – Ульяновск: УлГТУ, 2016. – 130 с.

[3] Ермаков, П.Д. Исследование методов машинного обучения в задаче автоматического определения тональности текстов на естественном языке / П.Д. Ермаков, Р.В. Федянин. – Москва: МГТУ им. Н.Э. Баумана, 2015. – С. 600-614.

[4] Галимов, Р.Г. Основы алгоритмов машинного обучения – обучение с учителем. – Уральский государственный экономический университет. – 2017. – С. 807-809.