.**Министерство науки и высшего образования Российской Федерации**

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Владимирский государственный университет**

**имени Александра Григорьевича и Николая Григорьевича Столетовых»**

**(ВлГУ)**

СМК 30/05-22

Срок хранения 5 лет

ВЛГУ 10.05.04.06.05.00 ПЗ

**П О Я С Н И ТЕ Л Ь Н А Я З А П И С К А К   
 К У Р С О В О Й Р А Б О Т Е**

«Разработка системы определения достоверности текстовой информации из новостных источников по тематике «Информационная безопасность»»

Дисциплина: Методы анализа данных

Специальность: 10.05.04 – Информационно-аналитические системы безопасности

#### СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ к.т.н., доц. каф. ИЗИ Ю.М. Монахов

Исполнитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_студент гр. ИСБ-118 А.Н. Катков

СОДЕРЖАНИЕ

[АННОТАЦИЯ 4](#_Toc73017564)

[1 СЛОВАРЬ 5](#_Toc73017565)

[2 ДОСТОВЕРНЫЕ НОВОСТИ 8](#_Toc73017565)

[2.1 Источники достоверных новостей 8](#_Toc73017567)

[2.2 Парсер новостей 8](#_Toc73017567)

[3 НЕДОСТОВЕРНЫЕ НОВОСТИ 9](#_Toc73017574)

[3.1 ruGPT-3 XL 9](#_Toc73017566)

[3.2 Цепи Маркова 9](#_Toc73017566)

[4 ПРОВЕРКА ДОСТОВЕРНОСТИ 10](#_Toc73017574)

[ЗАКЛЮЧЕНИЕ 13](#_Toc73017574)

[СПИСОК ЛИТЕРАТУРЫ 14](#_Toc73017575)

[ПРИЛОЖЕНИЕ А 15](#_Toc73017576)

#### АННОТАЦИЯ

Тема курсовой работы - Разработка системы определения достоверности текстовой информации из новостных источников по тематике «Информационная безопасность».

Этапы разработки:

1. Создание словаря;
2. Определение списка новостных источников;
3. Парсер новостных статей;
4. Создание недостоверных новостей;
5. Проверка достоверности собранного датасета.

Словарь составляется на основании тематических документов.

Из определенных новостных источников, статьи при достижении определенного процентного порога совпадения содержащихся в них слов со словарем, попадают в датасет, в качестве достоверных новостей.

Недостоверные новости генерируются различными способами.

Проверка достоверности осуществляется посредством предобученной модели rubert-tiny2.

#### СЛОВАРЬ

Для отбора новостных статей, с сайтов был сгенерирован словарь. Словарь включает в себя тысячу самых употребляемых в текстах документов слов. Для создания словаря были взяты тематические документы с сайта Московского государственного университета имени М.В.Ломоносова, раздела информационная безопасность – документы.

В качестве документов были взяты тематические документы:

1. Федеральный закон "Об электронной подписи" от 06.04.2011 N 63-ФЗ;
2. Федеральный закон "О коммерческой тайне" от 29.07.2004 N 98-ФЗ;
3. Федеральный закон "О связи" от 07.07.2003 N 126-ФЗ;
4. Методические рекомендации по обеспечению с помощью криптосредств безопасности персональных данных при их обработке в информационных системах персональных данных с использованием средств автоматизации;
5. Федеральный закон "Об информации, информационных технологиях и о защите информации" от 27.07.2006 N 149-ФЗ;
6. Федеральный закон "О персональных данных" от 27.07.2006 N 152-ФЗ;
7. Методические рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 г. N 996 "Об утверждении требований и методов по обезличиванию персональных данных";
8. Федеральный закон "О безопасности" от 28.12.2010 N 390-ФЗ;
9. ГОСТ Р 50.1.056–2005 «Рекомендации по стандартизации. Техническая защита информации. Основные термины и определения»;
10. ГОСТ Р 50.1.053–2005 «Рекомендации по стандартизации. Информационные технологии. Основные термины и определения в области технической защиты информации»;
11. Информационное сообщение ФСТЭК об утверждении Требований к средствам антивирусной защиты от 30 июля 2012 г. No 240/24/3095;
12. ГОСТ Р 50739-95 Средства вычислительной техники. Защита от несанкционированного доступа к информации. Общие технические требования;
13. ГОСТ Р 50922-2006" Защита информации. Основные термины и определения";
14. Методический документ "Меры защиты информации в государственных информационных системах" (утв. Федеральной службой по техническому и экспортному контролю 11 февраля 2014 г.);
15. Постановление Правительства РФ от 01.11.2012 N 1119 "Об утверждении требований к защите персональных данных при их обработке в информационных системах персональных данных";
16. Приказ ФСТЭК России от 11.02.2013 N 17 (ред. от 28.05.2019) "Об утверждении Требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах";
17. Приказ ФСТЭК России от 18.02.2013 N 21 (ред. от 14.05.2020) "Об утверждении Состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных";
18. Приказ ФСТЭК России от 14.03.2014 N 31 (ред. от 15.03.2021) "Об утверждении Требований к обеспечению защиты информации в автоматизированных системах управления производственными и технологическими процессами на критически важных объектах, потенциально опасных объектах, а также объектах, представляющих повышенную опасность для жизни и здоровья людей и для окружающей природной среды";
19. Приказ ФСБ РФ N 416, ФСТЭК РФ N 489 от 31.08.2010 "Об утверждении Требований о защите информации, содержащейся в информационных системах общего пользования";
20. Сборник руководящих документов по защите информации от несанкционированного доступа;
21. Приказ Минкомсвязи РФ от 25.08.2009 N 104 "Об утверждении Требований по обеспечению целостности, устойчивости функционирования и безопасности информационных систем общего пользования";
22. Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 5 сентября 2013 г. N 996 "Об утверждении требований и методов по обезличиванию персональных данных";
23. <Письмо> Минкомсвязи РФ от 13.05.2009 N ДС-П11-2502 "Об осуществлении трансграничной передачи персональных данных"
24. Указ Президента РФ от 05.12.2016 N 646 "Об утверждении Доктрины информационной безопасности Российской Федерации";
25. Указ Президента РФ от 17.03.2008 N 351 (ред. от 22.05.2015) "О мерах по обеспечению информационной безопасности Российской Федерации при использовании информационно-телекоммуникационных сетей международного информационного обмена";

#### ДОСТОВЕРНЫЕ НОВОСТИ

* 1. Источники достоверных новостей

В качестве источников достоверных новостей, были выбраны следующие новостные сайты:

1. www.securitymedia.org
2. www.anti-malware.ru
3. www.xakep.ru
4. www.securitylab.ru
   1. Парсер новостей

В качестве уровня, для попадания новости в пул достоверных, относящихся к области ИБ, был определен порог в 7 процентов. Всё что ниже это отметки, не будет относится к теме ИБ.

При поиске новостей, сначала определяется количество страниц на сайте, затем начинается поиск подходящих статей. Поиск будет окончен либо при достижении заданного количества статей, либо когда статьи на сайте иссякнут.

В процессе поиска, происходит считывание статьи с помощью библиотеки Newspaper3k. Из статьи удаляются все лишние символы, все слова приводятся в нижний регистр, она лемматизируется и сравнивается со словарем. Затем статья добавляется в датасет, с соответствующей меткой.

#### неДОСТОВЕРНЫЕ НОВОСТИ

Найти русскоязычные недостоверные новости и составить из них датасет задача крайне сложная. В силу данного обстоятельства, использовались следующие стратегии по добавлению недостоверных новостей в датасет:

1. Поиск датасета недостоверных новостей в интернете;
2. Генерация новостей.

При поиске в интернете был найден датасет news.csv c фейковыми и реальными новостями.

* 1. ruGPT-3 XL

Первый способ генерации – посредством предобученной нейросети ruGPT-3 XL. На вход подавался заранее составленный перечень начал новстей, вместе с ключевым словом новость. На выход поступали сгененрированные новости, которые затем чистились от лишнего наполнения как программно, так и руками. Генерация осуществлялась в Google Colab и заняла приличное количество времени, однако качество новостей как по связности, так и по наполнению оставляет желать лучшего.

* 1. Цепи Маркова

Второй способ генерации – посредством цепей Маркова. На вход генератору подаются статьи, которые были найдены парсером на новостных сайтах и добавлены в датасет реальных новостей, на выходе выдаются вектора размером от 8 до 15 слов, в количестве 20 штук. Они имитируют новостные статьи не по смысловой, но по словесной нагрузке.

#### ПРОВЕРКА ДОСТОВЕРНОСТИ

Проверка достоверности осуществляется посредством нейросети BERT. В качестве токенайзера и предобученной модели используется rubert-tiny2. Данная модель ориентирована только на русский язык, имеет словарный запас в 83828 токена, а также достаточно быстро работает. Длинна вектора для bert была взята равно 512, поскольку длинна большинства статей в выборках, как показано на рисунках 1, 2, 3 колеблется в большинстве своем до 512 слов.

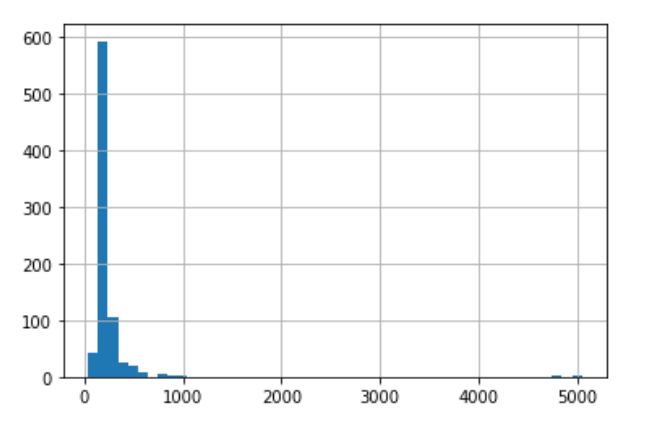


Рисунок 1 – Длины новостей с цепями Маркова

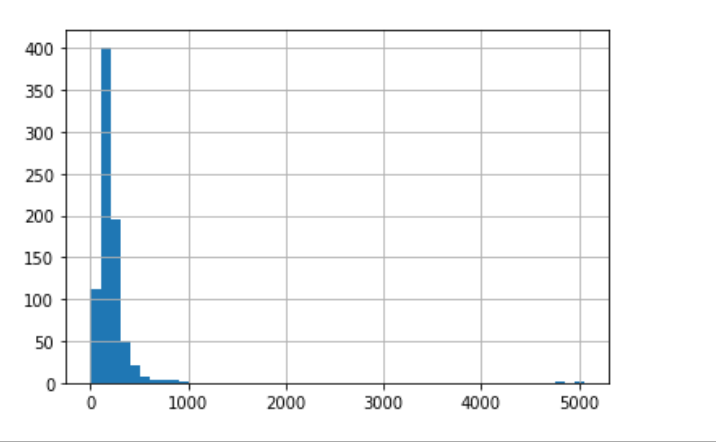


Рисунок 2 – Длины новостей с сгенерированными нейронной сетью

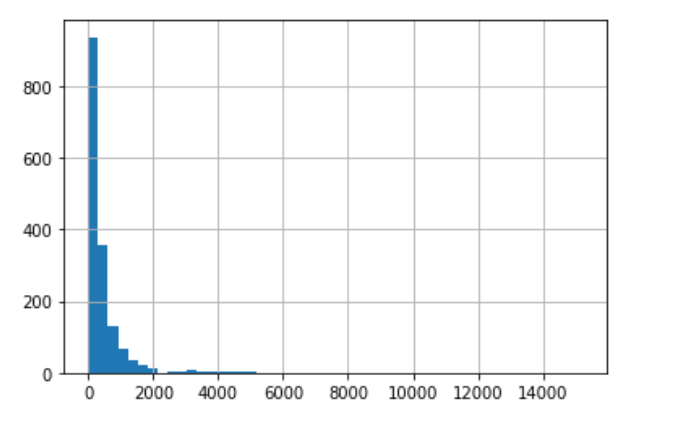


Рисунок 3 – Длины новостей с фейками из интернета

Результаты работы классификатора представлены ниже на рисунках 4, 5, 6. Как видно ниже, выборки, где данные которые были сгенерированы дают 100 процентную точность, что означает плохое качество датасетов. Примесь фейков из выборки найденной в интернете дала точность 70 процентов.

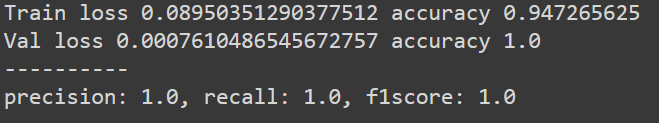


Рисунок 4 – Результат работы с цепями Маркова

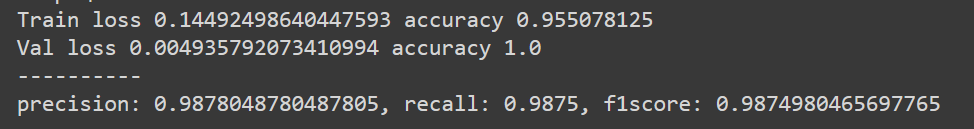


Рисунок 5 – Результат работы с сгенерированными новостями



Рисунок 6 – Результат работы с новостями из датасета, из интернета

ЗАКЛЮЧЕНИЕ

В результате выполнения работы была разработана система определения достоверности текстовой информации из новостных источников по тематике «Информационная безопасность».

СПИСОК ЛИТЕРАТУРЫ

1. How I Created a Fake News Detector with Python. Электронный ресурс. URL: https://towardsdatascience.com/how-i-created-a-fake-news-detector-with-python-65b1234123c4 (дата обращения: 15.03.2022);
2. Классификация текста с использованием моделей трансформеров. Электронный ресурс. URL: https://newtechaudit.ru/klassifikacziya-teksta-s-ispolzovaniem-modelej-transformerov/ (дата обращения: 15.04.2022);
3. Распознавание нейросетью фейковых новостей по их тексту. Электронный ресурс. URL: https://newtechaudit.ru/raspoznavanie-nejrosetyu-fejkovyh-novostej-po-ih-tekstu/ (дата обращения: 25.04.2022);
4. BERT для классификации русскоязычных текстов. Электронный ресурс. URL: https://habr.com/ru/post/567028/ (дата обращения: 10.05.2022);

ПРИЛОЖЕНИЕ А **–** QR-КОД ССЫЛКА НА КОД



Рисунок 7 – Ссылка на репозиторий с кодом