Статья для всероссийской студенческой научно-практической конференции Нижневартовского государственного университета

Гаврилов А.С.
Томский государственный университет
систем управления и радиоэлектроники
г. Томск, Россия

Использование вариационных автоэнкодеров как способ обнаружения политической пропаганды в сообщениях Telegram-каналов

Введение

В последние годы социальные сети, включая Telegram, стали основными каналами распространения информации. Их популярность обусловлена возможностью обмена новостными сведениями, мнениями, и идеями в режиме реального времени. Однако широкая доступность и отсутствие строгого контроля за публикуемым контентом способствуют распространению политической пропаганды и манипулятивных сообщений. В таких условиях выявление отклонений от привычного авторского стиля в текстах становится важной задачей, позволяющей идентифицировать потенциально вредоносный контент.

Одним из ключевых инструментов в этой области является анализ текстов на предмет аномалий. Нарушения авторского стиля могут указывать на вмешательство третьих лиц, использование автоматических генераторов текста или намеренное искажение фактов. Исследование данной тематики требует разработки методов, которые учитывают как общие особенности новостных постов в социальных сетях, так и уникальные особенности, связанные с тематикой и индивидуальными особенностями автора канала.

Методы и подходы анализа текста

Общий подход к анализу текста можно разделить на три вида [1]:

1) Статистические методы – методы, основанные на анализе свойств текста (анализ знаков препинания, частота использования частей речи, ключевые слова);
2) Алгоритмы машинного обучения – современные подходы к работе с большим набором данных;
3) Гибридные – методы, в основе которых используются алгоритмы машинного обучения, подкрепленные статистическим анализом текста.

Наиболее эффективным считается гибридный метод, поскольку он объединяет преимущества двух других подходов: детальный анализ, основанный на выделении уникальных характеристик конкретного текста, и способность обрабатывать большие объемы данных. Для выявления аномалий в тексте используются различные признаки, которые можно классифицировать на несколько групп [3]:

1. **Структурные признаки**: длина текста, наличие заголовков, абзацев, форматирование (жирный шрифт, курсив), использование списков и таблиц. Эти характеристики отражают общий формат текста.
2. **Синтаксические признаки**: частота использования частей речи, средняя длина предложений, наличие вопросов и восклицаний, частота использования местоимений. Они описывают грамматическую структуру текста.
3. **Лексические признаки**: частота использования ключевых слов, биграмм и триграмм, наличие специфических символов (@, #, $, и т.д.), доля уникальных слов, количество длинных слов.
4. **Семантические признаки**: темы или категории текста, ключевые слова, частота использования терминов, связанных с определённой тематикой.

5. **Признаки взаимодействия**: наличие упоминаний пользователей, ссылок, медиаконтента (изображения, видео), использование эмодзи. Данные признаки являются уникальными, так как они присущи только для социальных сетей.

С точки зрения машинного обучения, задачу выделения авторской атрибутики в тексте позволяют решать не только автоэнкодеры. Одним из таких методов являются сверточные нейронные сети (CNN) [2]. Данный подход эффективен для выявления важных особенностей в тексте. Однако, в ситуациях, когда для анализа текстов не доступен размеченный набор данных, содержащий примеры нормальных и аномальных сообщений, а также если нарушения локальных признаков, таких как специфические грамматические ошибки или редкие лингвистические конструкции, не имеют четко выраженный характер, то сверточные нейронные сети не смогут продемонстрировать достаточную точность для корректного выявления аномалий.

Другой метод, который может стать хорошей опорой для анализа текстов на предмет атрибутики автора – это LSTM (сети с долгой краткосрочной памятью) [2, 4]. Эта разновидность рекуррентной нейронной сети хороша тем, что за счет запоминающей ячейки она может хорошо справляться с анализом последовательных данных, таких как текст. Однако такие недостатки как высокая чувствительность к шумам (грамматические ошибки, сокращения, эмодзи) и потеря значимой информации при обработке больших текстов, делают эту модель непригодной для работы с социальными сетями.

Самым эффективным способом на данный момент является анализ текста в социальных сетях с использованием автоэнкодеров. Автоэнкодер [5] – это специализированная искусственная нейронная сеть, предназначенная для обучения эффективному восстановлению значимых признаков текста. За счет особенностей архитектуры сети, данный метод позволяет выявлять аномалии, опираясь не только на локальные признаки, но и учитывая все ключевые характеристики текста Также существенным преимуществом автоэнкодера является его независимость от длины сообщения, что делает данный подход особенно актуальным для анализа текстовых аномалий.

Подготовка модели к обучению

Перед обучением модели требовалось выполнить статистический анализ новостных каналов, с которыми в дальнейшем предстоит работать.

Так, для выявления аномалий в 20 отобранных каналах на тему политики, наиболее важными оказались синтаксические и лексические признаки, а также специфические характеристики взаимодействия, характерные для Telegram-каналов. Эти группы обеспечивают детальный анализ текста, выявляя не только отклонения в манере изложения, но и возможное наличие манипулятивного контента.

Отобранные признаки включали:

- **Синтаксические**: количество предложений, использование восклицательных и вопросительных знаков;
- **Лексические**: частота использования ключевых слов и биграмм, наличие длинных слов;
- **Признаки взаимодействия**: количество ссылок, хэштегов, эмодзи, наличие медиафайлов.

Описание метода реализации

В данной работе для анализа текстов используются два ключевых инструмента: TF-IDF и вариационные автоэнкодеры (VAE).

1. **TF-IDF (Term Frequency-Inverse Document Frequency) [2]** — это статистический метод, оценивающий важность термина в документе относительно всего корпуса текстов. Формула расчёта выглядит следующим образом:

$$\mathrm{TF-IDF(t, d)} = TF(t, d) * IDF(t) \quad (3)$$

где TF(t, d) – частота термина (TF) для слова “t” в документе “d”, IDF(t) – обратная частота документа (IDF) для слова “t”.

Данный метод позволяет выделить ключевые слова, характерные для определённого Telegram-канала, и использовать их как признаки для дальнейшего анализа. Это позволило учитывать специфику контента и повысить точность анализа.

2. **Вариационные автоэнкодеры (VAE)** [5] — это разновидность автоэнкодера, которая кодирует входные данные в вероятностное пространство, а затем восстанавливает их (Рисунок 1).

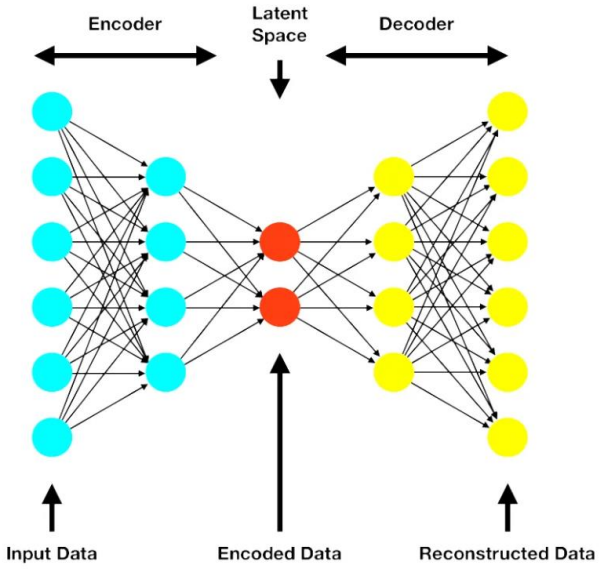

Рис. 1 – Схема работы VAE

Результаты

После обучения модель была протестирована с использование и без использования индивидуализированных признаков на сообщениях канала, где заранее были размечены сообщения на аномальные и нормальные (Рисунки 2 – 3).

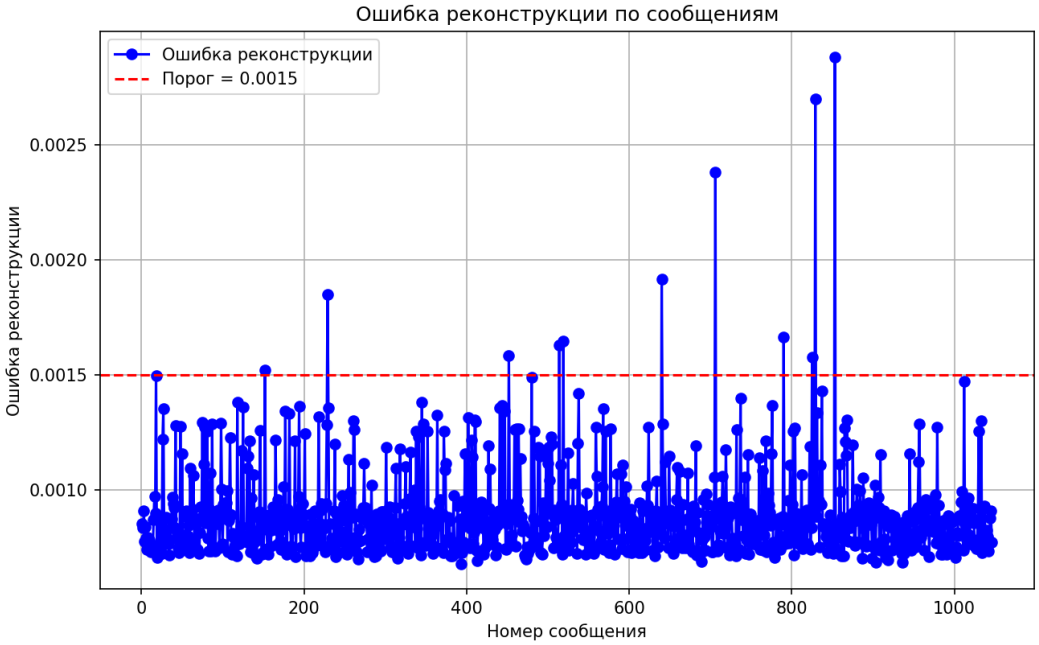

Рис. 2 – Поиск аномалий VAE с использованием специализированных признаков

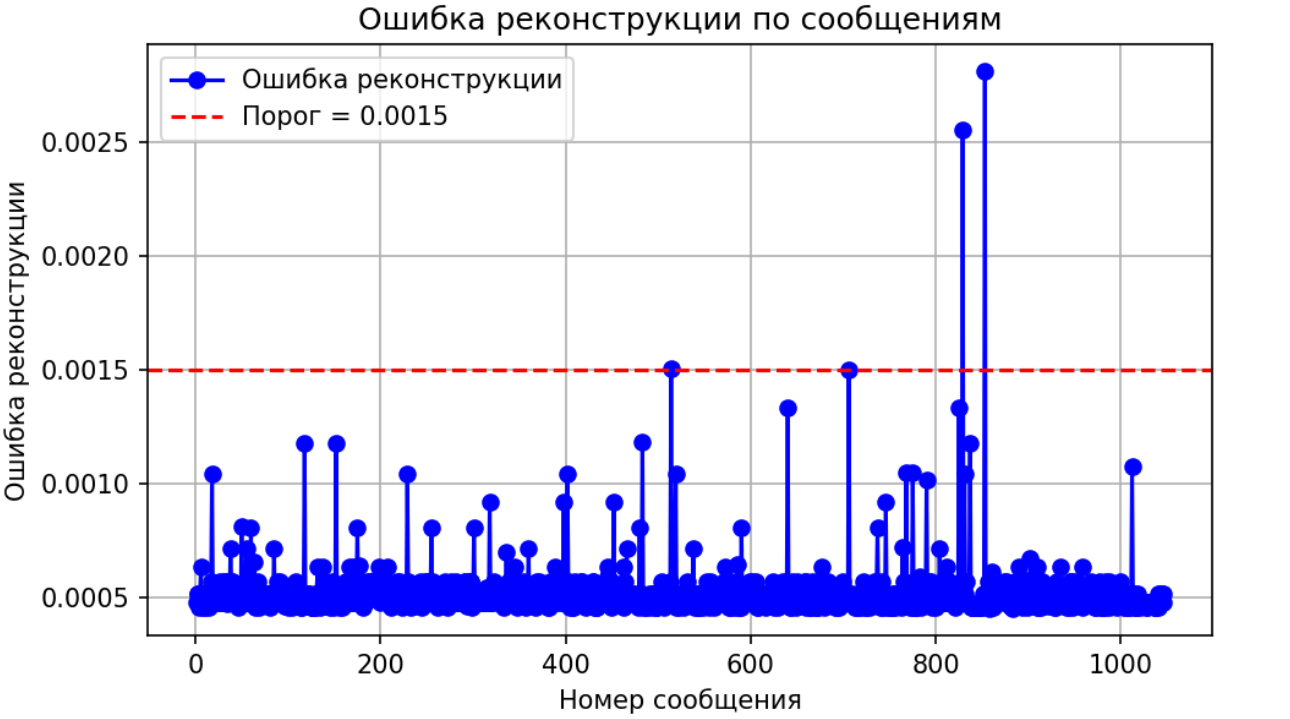

Рис. 3 – Поиск аномалий VAE без использования специализированных признаков

Визуализация результатов продемонстрировала, что VAE с выбранным набором индивидуализированных признаков успешно справился с задачей, корректно выявив все аномалии. В то же время использование модели вариационного автоэнкодера без учета признаков взаимодействия и TF-IDF, хотя и позволило обнаружить часть аномалий, не достигло сопоставимого уровня точности.

Заключение

В ходе исследования был разработан и протестирован подход к выявлению нарушений авторского стиля в текстах Telegram-каналов. Использование TF-IDF и вариационных автоэнкодеров в сочетании с индивидуализированным набором признаков позволило эффективно обнаруживать аномалии. Предложенный метод может быть применён для идентификации политической пропаганды и манипулятивного контента, что делает его важным инструментом в современных условиях информационной безопасности.

Литература

1. Шкодырев В.П., Ягафаров К.И., Баштовенко В.А., Ильина Е.Э. Обзор методов обнаружения аномалий в потоках данных // International Journal of Open Information Technologies: 2022. № 2.
2. Tang X. Author identification of literary works based on text analysis and deep learning // Heliyon. – 2024. – Vol. 10, no. 3.
3. Таюпова О.И. Категории и признаки текста // Вестн. Башкир. гос. ун-та. 2011. № 4.
4. Д.Е. Савицкий, М.Е. Дунаев, К.С. Зайцев Выявление аномалий при обработке потоковых данных в реальном времени // International Journal of Open Information Technologies 2022. № 1.
5. Акинина Н.В. Акинин М.В. Соколова А.В. Никифоров М.Б. Таганов А.И. Автоэнкодер: подход к понижению размерности векторного пространства с контролируемой потерей информации // Известия Тульского государственного университета. Технические науки – 2016