МЕТОДЫ ПОИСКА АНОМАЛИЙ В СОЦИАЛЬНЫХ СЕТЯХ С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Василенко В.И.[1], Сурвилов А.О.[2]
Научный руководитель: Савельев А.О.[3]
[1] НИ ТПУ, студент гр. 8К12 ОИТ, ИШИТР, e-mail: viv8@tpu.ru
[2] НИ ТПУ, студент гр. 8К12 ОИТ, ИШИТР, e-mail: aos45@tpu.ru
[3] НИ ТПУ, к.т.н., доцент ОИТ, ИШИТР, e-mail: sava@tpu.ru

Аннотация

Данная работа посвящена методам поиска аномальных публикаций социальных сетей среди набора семантически схожих. Статья рассматривает применение методов машинного обучения для решения данной задачи, а именно анализ эмоционального окраса и выделение именованных сущностей.

Ключевые слова: Обработка естественного языка, машинное обучение, социальные сети, hugging face, эмоциональный окрас, Reddit, NER, кластеризация.

Введение

На данный момент в связи со стремительным развитием web-технологий существенная часть социальных коммуникаций производится при помощи социальных сетей. Это делает социальные сети важным источником данных для проведения социальных исследований. Однако большой объем данных ведет к тому, что исследователи вынуждены решить множество подготовительных задач во время проведения исследований, таких как сбор и поиск целевой информации. В связи с этим возникает потребность автоматизации этих процессов.

Одной из важных задач в рамках социальных исследований является поиск аномалий среди публикаций социальных сетей, которые представляют собой публикации, вызывающие нестандартные реакции среди участников сообщества или которые сами выделяются по каким либо критериям среди остальных постов. Анализ данных публикаций может помочь в изучении аудитории, которая участвует в дискуссии, а также причины, по которым возникают данные аномалии. Кроме того, поиск и устранение аномалий выступает важным этапом подготовки данных для исследования. В связи с этим в рамках данной работы было решено рассмотреть методы автоматизации обнаружения аномалий среди публикаций сообществ в социальной сети, основанные на моделях машинного обучения. Аномальные публикации– это такие посты, реакция на которые сильно отличается (метод трех сигм) от реакции на семантически схожие, например, внутренними метриками числом ответов, лайков и другими. Однако эти метрики не всегда могут быть информативны, кроме того, они не позволяют выявить причины аномальности. В связи с этим предлагается использовать методы машинного обучения для решения этой задачи. Как итог цель данной работы оценить смогут ли методы машинного обучения предоставить дополнительные критерии для выявления аномальных публикаций. Для достижения этой цели планируется выполнить следующий список работ.

1. Выбор сообщества для анализа и сбор данных.
2. Разбиение данных на семантически схожие подгруппы.
3. Выявление аномальных постов на основе метрик социальной сети.
4. Выявление дополнительных маркеров для поиска аномалий, на основе эмоциональной оценки и выявления именованных сущностей.

Сбор и подготовка данных

Сбор необходимых для работы данных осуществлялся в социальной сети Reddit, при помощи официального API данной платформы. API Reddit позволил собрать данные о 58179 записях из сообщества minipainting, каждая из которых является либо публикацией либо

комментарием к публикации. После сбора данных был отдельно выделен датасет [1] содержащий только информацию о постах общим количеством 998. Изначальные данные, полученные при помощи API содержат поля, представленные в таблице 1.

Таблица 1. Свойства набора данных, извлечённого из Reddit

Название	Описание	Наличие пропусков
full_name	Идентификатор текстовой записи из Reddit API.	Нет
text_body	Текстовое содержание записи, кодированное в UTF-8, если оно не было удалено автором или модератором.	Да
author_name	Идентификатор автора текстовой записи, если учётная запись не была удалена.	Да
votes	Рейтинг записи.	Нет
responds_to	Идентификатор записи, на которую отвечает данная. Корневая запись не отвечает на другие.	Да
parent_submission_name	Идентификатор публикации, в дереве которой находится данная запись. В корневой записи это свойство не указывается.	Да
submission_flair	Тэг текстовой записи, который един для всего дерева публикации.	Да
created_timestamp	UNIX метка времени создания записи по данным Reddit.	Нет
parsed_timestamp	UNIX метка времени извлечения и сохранения записи.	Нет
controversiality	Отметка о спорности записи по данным Reddit.	Нет

Публикации, полученные в результате прошлого шага, принадлежат сабредиту (сообществу в сети Reddit) с весьма узкой специализацией в связи с чем публикации в сообществе должны быть семантически схожи. Для работы с менее узконаправленными сообществами, где посты не являются гарантированно семантически схожими, необходимо выявить среди общего набора публикаций подгруппы схожих. Для решения данной задачи было решено использовать векторизацию TF-IDF и метод K-means для разбиения на кластеры. Итоговое количество кластеров выбирается исходя из необходимости максимизировать оценку компактности кластера [2] (для гарантирования достаточной семантической схожести), а также из необходимости иметь внутри кластера количество постов, достаточное для проведения анализа.

Следующим этапом производится подсчет комментариев для каждой публикации, после чего опираясь на внутренние метрики Reddit (votes и число комментариев), были выбраны посты имеющие высокий шанс быть аномалией. Под такими постами подразумевается посты имеющие votes выше (или ниже) 90 (10) перцентиля от votes внутри своего кластера, аналогично для комментариев. Таким методом было обнаружено 332 аномалии.

Выявление дополнительных маркеров

Следующим этапом была произведена оценка эмоционального контекста публикаций, для этого было решено использовать модель Emotion English DistilRoBERTa-base [3, 4], данная модель возвращает данные в формате предсказаний 6 эмоций-классов по теории эмоций Пола Экмана и дополнительный нейтральный класс на основе текста. Каждое предсказание оценивает от 0 до 1 (от минимального к максимальному) насколько вероятна данная эмоция в предложенном тексте. После чего для каждой записи были получены эмоциональные оценки поля text_body. Кроме того, для каждой публикации были также определенны средние эмоциональные оценки комментариев к ней. Подсчет производился следующим образом:

1. Расчет эмоциональных предсказаний для текстов в постах и их комментариях.
2. Агрегация эмоциональных предсказаний комментариев с общей публикацией – родителем.
3. Расчёт среднего значения.

Также было решено выделить именованные сущности для каждой публикации. Для этого использовалась модель WikiNEuRal [5]. Данная модель возвращает именованную сущность, ее тип и внутреннюю оценку уверенности модели в правильности. Для того чтобы отсеять возможные ошибки в работе модели было решено отсеять те сущности, уверенность в выявлении которых меньше, чем медиана. Что позволило выявить 418 уникальных сущностей.

После применения данных моделей было решено провести сравнение результатов их работы для публикаций выделенных ранее аномальных публикаций и для остальных публикаций. Для начала было произведено сравнение медиан оценок по каждой эмоции между основным набором и набором аномалий. Были получены, которые показаны в таблицах 2 и 3.

Таблица 2. Сравнение медиан эмоций агрегаций комментариев

Эмоция	Медиана аномалий	Медиана основных данных
Anger	0.042130	0.040181
Disgust	0.021382	0.022143
Fear	0.021628	0.019839
Joy	0.299936	0.304723
Neutral	0.284702	0.284867
Sadness	0.070268	0.071219
Surprise	0.212150	0.213253

Таблица 3. Сравнение медиан эмоций публикаций

Эмоция	Медиана аномалий	Медиана основных данных
Anger	0.012777	0.011814
Disgust	0.004582	0.003996
Fear	0.005812	0.005554
Joy	0.036018	0.050114
Neutral	0.366309	0.337490
Sadness	0.023063	0.022284
Surprise	0.071670	0.075168

Результаты данного анализа приводят к тому, что аномальные публикации и их комментарии практически не отличаются от остальных содержащихся в сабредите. Такие результаты могут быть обусловленным малым объемом датасета, который не позволяет получить более конкретные отличия между эмоциями аномалий и основных данных.

Далее для каждой именованной сущности в датасете с аномалиями было вычислено количество ее упоминаний. После чего было произведено разделение на два набора: набор именованных сущностей и числа их упоминаний для аномальных постов, и для всех постов. Было обнаружено, что большинство сущностей, упоминаемых среди аномальных постов, используется лишь 1. При этом 90,6 процентов всех сущностей, которые упоминаются лишь 1 раз среди всех записей, упоминается именно в аномальных статьях. Это может свидетельствовать, что редкие сущности могут быть маркером аномальной публикации

После чего были рассчитаны медианные значения параметров votes и числа комментариев для всех постов в датасете, которые содержат упоминания самых популярных сущностей среди аномалий. Популярные сущности в данном случае – это такие сущности, число упоминаний которых больше медианы. Результаты расчетов показаны в таблице 4.

Таблица 4. Сравнение медиан постов с аномальными сущностями со всеми

	С упоминанием популярных аномальных сущностей	По всем публикациям
Медиана votes	1801	1147
Медиана комментариев	37	42

Можно заметить разницу в числе votes, в постах с аномальными сущностями медиана данного критерия на 36 процентов выше, чем медиана по всем публикациям. Что может свидетельствовать, что самые популярные среди аномалий сущности тоже могут быть маркером для расширения списка аномалий.

Ограничения

Одним из заметных ограничений в данной работе является точность модели Emotion English DistilRoBERTa-base, используемой для определения эмоциональных предсказаний даже при учете того, что модель обучалась 6 различных датасетах вероятность неверно определяемых эмоциональных весов остаётся. Кроме того, было обнаружено что модель не всегда строит свои выводы исходя из контекста предложенного текста – из-за чего многим речевым оборотам присваиваются неверные эмоции. Также не стоит исключать ограниченность объема данных, собранных для анализа что может привести к тому что на определенных данных в силу их специфики результаты могут сильно варьироваться.

Таким же важным является ограничение работы модели WikiNEuRal которая также способна на неверное определение именованных сущностей.

Заключение

В ходе работы было проведено исследование на тему способности методов машинного обучения выявить новые маркеры для поиска аномальных публикаций в социальных сетях. Был реализован механизм выявления семантически схожих групп при помощи кластеризации с последующим выявлением аномалий в этих группах, которые позже были обработаны методами машинного обучения. По итогам чего было выявлено, что эмоциональный окрас не может являться точным маркером для выявления аномальных публикаций. Однако учитывая как ограничения датасета, так и самой модели полученный результат требует дополнительных исследований.

Кроме того, было обнаружено, что именованные сущности также могут быть маркером для дополнительного выявления аномальных постов. При помощи отбора постов, содержащих уникальные для всего датасета именованные сущности, так и при помощи выбора постов, содержащих сущности, которые часто встречаются среди аномалий.

Литература

1. Pandas documentation – pandas 2.2.3 documentation // Pydata.org [Электронный ресурс]. – URL: pandas.pydata.org/docs/ (дата обращения: 16.03.2025).

2. Оценка качества в задаче кластеризации // Ifmo.ru [Электронный ресурс]. – URL: neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче_кластеризации (дата обращения: 16.03.2025).

3. J-hartmann/emotion-english-distilroberta-base · Hugging face // Huggingface.co [Электронный ресурс]. URL: huggingface.co/j-hartmann/emotion-english-distilroberta-base (accessed: 16.03.2025).

4. Transformers // Huggingface.co [Электронный ресурс]. – URL: huggingface.co/docs/transformers/v4.49.0/en/index (accessed: 16.03.2025).

5. Babelscape/wikineural-multilingual-ner · Hugging Face // Huggingface.co [Электронный ресурс]. – URL: huggingface.co/Babelscape/wikineural-multilingual-ner (accessed: 16.03.2025).