Санкт-Петербургский государственный университет

Филатов Илья Сергеевич

Выпускная квалификационная работа

Методы агрегации и представления пользовательских данных в дискуссиях социальных сетей

Уровень образования: бакалавриат
Направление 02.03.02 «Фундаментальная информатика и информационные технологии»
Основная образовательная программа: «Программирование и информационные технологии»

Научный руководитель:
кафедра технологии программирования
к.ф. - м.н., доцент, заведующий кафедрой
Блеканов Иван Станиславович

Рецензент:
д.ф. - м.н., профессор
Крылатов Александр Юрьевич

Санкт-Петербург
2022 г.

Содержание

Введение

Актуальность

В настоящее время основным способом коммуникации между людьми являются социальные сети и мессенджеры. Практически все люди на планете пользуются данными онлайн платформами для повседневного общения, донесения важной информации, знакомств, распространения различной информации. Так же люди используют социальные сети и мессенджеры в качестве бизнес платформ, то есть в качестве средства, позволяющего людям зарабатывать деньги. Такое явление наблюдается из-за того, что для того, чтобы сгенирировать контент на онлайн платформах не нужно иметь специальных навыков. Из-за такого повсеместного использования социальные сети и мессенджеры содержат гигантское количество информации о конкретных событиях, причем мнение о событии со временем может меняться, также на онлайн платформах содержится информация о конкретных людях, группах людей, объединенных по интересам, различных отдельных организациях или группах организаций, объединенных по какому-либо признаку.

В современном мире возникает потребность проанализировать всю накопленную социальными сетями информацию для различных целей, например, для выявления общественных мнений из дискуссий о каком-либо событии. Для изучения подобной информации был создан Анализ социальных сетей (SNA), являющийся одновременно совокупностью методов и областью исследования социальных структур с использованием сетей, теории графов и методов машинного обучения. Анализ социальных сетей стал ключевым методом в современной социологии. Он также приобрел значительную популярность в следующих областях: биология, экономика, политология и в информационных технологиях.

В рамках области анализа социальных сетей существует много задач: контентный анализ, суммаризация текстов, детекция ботов и т.д. Но так как к анализу соцаильных сетей приходят все больше и больше экспертов из разных гуманитарных и социальных сфер, то одной из самых актуальных

задач в области SNA становится понятная визуализация пользовательских дискуссий в социальных сетях. Визуализация позволяет распознать скрытые зависимости и тренды, которые являются неочевидными при использовании других видов анализа. Используя полученную визуализацию, можно сразу определить простые паттерны, а также провести более серьезный анализ с помощью других методов, которые основываются на графиках визуализации. Для решения этой задачи возникает потребность в удобном сервисе, который будет предоставлять возможность получить визуализацию дискуссии в конкретном сценарии, согласно выбранным фильтрам, а также будет предоставлять возможность проанализировать дискуссию с помощью полученной визуализации. Так как выявлением паттернов в дискуссиях занимаются эксперты из разных областей, которые пользуются визуализацией и на ее основе строят те или иные гипотезы, поэтому ошибка в визуализации увеличивает риск того, что эксперт примет неверное решение относительно каких-либо событий. Из этого следует, что к визуализации предъявляются высокие требования, то есть очень важно визуализировать дискуссию четким и понятным образом, а также предоставить экспертам несколько видов визуализации для различных сценариев анализа.

Цель работы

Целью данной работы является разработка сервиса визуализации методов искусственного интелекта по анализу дискуссий пользователей в социальных сетях, который поможет экспертам из разных областей получить наглядное представление результатов алгоритмов машинного обучения, а также поможет по полученной визуализации провести дальнейший анализ.

Задачи работы

Основными задачами данной работы являются:

1. Изучение проблематики современного анализа социальных сетей;
2. Проведение сравнительного анализа инструментов, позволяющих реализовать слой пользовательского представления данных в сервисе;
3. Проведение сравнительного анализа инструментов визуализации;
4. Проведение сравнительного анализа существующих сервисов по анализу социальных сетей;
5. Реализация модулей системы, позволяющих провести визуализацию и последующий анализ дискуссий в социальных сетях.

Практическая значимость работы

Данная дипломная работа несет следующие аспекты практической значимости:

- Разработанным сервисом смогут пользоваться социологи для исследования влияния какого-либо события на общество в целом или конкретные группы людей, а также проследить изменилось ли отношения общества к этому событию с течением или нет, а если изменилось то каким образом. Также социологи смогут находить лидеров мнений и что конкретно привело к их лидерству;
- Также разработанным сервисом смогут пользоваться психологи для исследования того, какие именно фразы в дискуссии могут положительно или отрицательно влиять на человека;
- Еще одной областью применения данной платформы является экономика, эксперты из данной области смогут по отношению пользователей к конкретной организации спрогнозировать рост или падение акций;

Обзор литературы

В работе [5] рассмотрены история и теория анализа социальных сетей, представлен сборник методов, моделей и приложений анализа данных социальных сетей. В [6] предложен сравнительный анализ социальных сетей в различных аспектах, так как анализ социальных сетей обеспечивает как математический, так и визуальный анализ человеческих взаимоотношений. В этой статье дается современный обзор работ, выполненных в области анализа социальных сетей, начиная от строго математического анализа (формальные методы, графы, матрицы и статистические модели) и заканчивая современным анализом социальных сетей в semantic web (Social network analysis (SNA)). Современному анализу социальных сетей (SNA), моделированию и визуализации сетей посвящен ряд работ [8], [9], [7].

В работах [10], [11], [12] описан анализ тональности (Sentiment Analysis). Основной задачей анализа тональности является классификация полярности входного текста с точки зрения положительной, отрицательной или нейтральной полярности. Методы в этой области используются для вывода общественного мнения. До сих пор большая часть усилий в области анализа настроений была посвящена анализу текста, в то время как ограниченные усилия были предприняты для определения настроений по изображениям и видео, которые становятся все более распространенными в социальных сетях.

В работах [13], [14], [15] описаны тематические модели, обеспечивающие удобный способ анализа больших объемов неклассифицированного текста и выявления скрытых тем. Тема содержит группу слов, которые часто встречаются вместе. Представленные методы тематического моделирования – скрытый семантический анализ (LSA), вероятностный латентный семантический анализ (PLSA), коррелированная тематическая модель (CTM).

Работы [16], [17], [18] посвящены инфографике, проектированию интерфейсов, которые могут визуализировать данные социальных сетей. А именно, в работах описано проектирование статистических графиков данных, диаграмм, таблиц, временных рядов, карты данных, многомерных схем с анализом того, как отображать данные для точного, эффективного и быстрого анализа, а также как обнаружить графический обман.

С помощью докуметаций библиотек [1], [2], [3], [4] проведен сравнительный анализ инструментов, позволяющих создать слой пользовательского представления данных, которые видит пользователь в браузере, и позволяющих взаимодействовать с приложением в целом.

Глава 1. Проблематика современного анализа социальных сетй

1.1 Общая проблематика анализа дискуссий в социальных сетях

Под социальной сетью в контексте изучения пользовательских дискуссий понимается совокупность социальных сущностей из реального мира, связанных совокупностью социальных отношений, например, таких как дружба, сотрудничество или обмен информацией. Социальными сущностями могут быть, например, люди, организации или государства. Например, люди, общающиеся друг с другом в мессенджере Telegram образуют социальную сеть, так как они связаны отношением общения. Преимуществом такой структуры, как социальная сеть, является то, что она показывает, как связаны сущности этой сети через различные заранее определенные связи. Так же особенностью социальных сетей является то, что их можно разделять на подгруппы, поведение которых может многое сказать о поведении сети в целом.

Чтобы проанализировать социальную сеть необходимо сначала произвести ряд подготовительных действий:

1. Важно выбрать коллекцию сущностей с полной информацией обо всех сущностях коллекции, например, все люди, которые комментируют видео, все друзья отдельного человека или все члены отдельного сообщества в мессенджере Telegram;

2. Важны ограничения выборки, обычно это максимум 50 сущностей, чтобы можно было сделать понятную визуализацию;

3. Нужно определить тип связи между сущностями, который будет использоваться в социальной сети. Обычно это простые отношения, такие как дружба. Так же возможно определение более косвенных связей, например, два человека могут быть определены, как связанные в сети, если они прокомментировали одно видео. Если нужно определить несколько связей между сущностями, то в таком случае создаются несколько социальных сетей, каждая с теми же сущностями, но с различными определениями связей;

После того, как проведен подготовительный этап можно анализировать социальную сеть. Выделяют два основных способа анализа социальных:

1. Структурный анализ социальных сетей;
2. Анализ с помощью визуального представления методов машинного обучения;

1.2 Структурный анализ социальных сетей

Структурный анализировать социальных сетй предполагет использование определенных метрик, которые представляют собой формулы или методы для измерения некоторых аспектов сети. Метрики так же позволяют проверить выявленные с помощью визуализации паттерны, но при этом с помощью метрик можно выявить и закономерности, которые сложно распознать с помощью визуализации. Метрики могут применяться как к отдельным узлам сети, так и к сети в целом. Рассмотрим основные метрики.

1. Наиболее частой метрикой для узлов сети является степень центральности. Степенью центральности узла является число других узлов, которые связаны с этим узлом.
2. Более сложной метрикой является степень посредничества. Для любой пары узлов в сети существует по крайней мере один кратчайший путь между этими узлами, и степень посредничества для узла равна числу кратчайших путей через этот узел. Узлы с самой большой степенью посредничества важны, так как если удалить их сети, то сеть будет не связной или информация от узлов будет доходить более долго, так как кратчайший путь пропадет. Обычно узлы с высокой степенью посредничества имеют так же и высокую степень центральности. Узлы с высокой степенью посредничества могут связывать различные группы узлов между собой.

Также в рамки структруного анализа входит задача определения наиболее важных связанных групп узлов. Существуют две основные стратегии для декомпозиции на связанные группы:

1. Clique calculations.В основе данной стратегии лежит понятие p-clique. В сети p-clique – это набор p узлов, которые соединены друг с другом. Стратегия анализа p-clique заключается в том, чтобы определить самое большое число p, чтобы существовали такие наборы p-clique, а затем перечислить и прокомментировать все такие наборы p-clique, как важные группы в сети с точки зрения связности. Если существует только один максимальный размер p-clique, то p может быть уменьшено для нахождения больших групп. К недостаткам такой стратегии можно отнести то, что наборы p-clique могут накладываться друг на друга, что может затруднить интерпретацию результатов. Так же из-за условия, что все узлы в наборе p- clique соединяются со всеми другими узлами из такого набора, какие-либо важные группы могут быть не обнаружены из-за того, что один узел соединяется со всеми узлами, кроме одного;

2. Альтернативным подходом является кластеризация. Сеть может быть разделена на кластеры узлов так, что узлы в каждом кластере имеют тенденцию соединяться друг с другом больше, чем с узлами вне кластера. В таком подходе исправляется недостаток стратегии p-clique, что важные группы могут быть не обнаружены из-за условия, что все узлы в наборе должны соединяться со всеми другими узлами из набора.Но в таком подходе появляется новая проблема, а именно что существует множество способов поделить узлы на кластеры, из-за чего становится сложно выбрать условие деления на кластеры. Чтобы легче принять решение, можно заранее решить, сколько кластеров должно быть найдено с помощью некоторых алгоритмов. Так же можно попробовать разные разделения на кластеры и выбрать такое разделение, которое дает наиболее четкое разделение;

1.3 Анализ с помощью визуального представления методов машинного обучения

С каждым днем использование таких социальных платформ, как Facebook и Twitter, как общественностью, так и организациями, быстро растет. Лица, принимающие решения в организациях, используют социальные сети для взаимодействия со своими клиентами, поскольку они склонны выражать свое мнение об определенных продуктах и услугах с помощью этого популярного механизма. Поэтому в социальных сетях находится очень большое количество информации, которую хочется использовать для анализа, однако основным препятствием для этого является получение именно значимой информации с этих платформ из-за неструктурированных данных, которые они представляют. Для того чтобы проанализировать такое огромное количество неструктурированных данных в социальных сетях, их нужно четко визуализировать подходящим способом для конкретной цели. Наиболее частыми задачами являются проведение анализа тональности(Sentiment Analysis) и тематического моделлирования(Topic Modeling), а также визуализация, связанная с этими задачами.

1.3.1 Анализ тональности

В условиях неструктурированных данных анализ настроений рассматривается как один из лучших инструментов для исследования идей или мнений в огромном объеме данных. Основной задачей анализа тональности является классификация полярности входного текста с точки зрения положительной, отрицательной или нейтральной полярности. Методы в этой области используются для вывода общественного мнения о каком-либо событии, либо личности или организации. Анализ тональности состоит из четырех этапов:

1. Извлеечние данных. Этот этап предполагает сбор данных по ключевым словосочетаниям и выбранным фильтрам в одной или нескольких социальных сетях. Сбор можно осуществлять с помощью API платвоформы, но в таком случае возникает препятствие в виде большого количества не полных данных, а также ограниченности доступных настроек. Наиболее лучшим вариантом является использование роботов-краулеров,

которые решают недостатки описанные выше недостатки;

2. Предварительная обработка данных. Этот этап подразумевает нормализацию данных, то есть удаление повторяющихся постов и нежелательных словосочетаний, исправление орфографии и приведение слов к одной лексической форме;

3. Предварительный анализ, который подразумевает использование методов машинного обучения для классифиыкации текста на положительный, отрицательный или нейтральный окрас. Говоря о тональности текста, следует выделять три параметра: субъект тональности (автора текста), тональную оценку (позитив, нейтральность или негатив) и объект тональности, то есть предмет, о котором высказывается мнение. Под «нейтральностью» подразумевается, что текст не содержит эмоциональной окраски. Процесс анализа тональности очень похож на процессы с применением машинного обучения: необходимо собрать коллекцию документов для обучения классификатора, затем каждый документ из обучающей коллекции нужно представить в виде вектора признаков, затем для каждого документа нужно указать «правильный ответ», то есть тип тональности, по этим ответам и будет обучаться классификатор, затем нужно выбрать алгоритм классификации и обучить классификатор;

4. Графическое представление полученных результатов с целью дальнейшего анализа;

1.3.2 Тематическое моделирование

Еще одним способом, обеспечивающим удобный анализ больших объемов неклассифицированного текста является тематическое моделирование. Данное направление позволляет понять, о чем говорят люди в целом, а также найти шаблонные скрытые темы. Тема содержит группу слов, которые часто встречаются вместе.

Идея тематического моделирования заключается в том, тексты представляют собой смесь тем, где тема - это распределение вероятностей по сло-

вам. Также важным аспектом тематического моделирования является идентификация темы во времени, то есть возможность проследить эволюцию темы с течением времени. Например, тематическое моделирование может помочь в ситуации, когда исследователь хочет выбрать тему исследования в определенной области и хотел бы знать, как эта тема развивалась с течением времени, и попытаться определить те тексты, которые объясняют эту тему. В рамках тематического моделирования существует несколько моделей:

1. Скрытая семантическая модель. Основная цель этой модели заключается в том, чтобы создать векторное представление текстов для создания семантического содержания. С помощью векторного представления вычисляется сходство между текстами, чтобы выбрать наиболее связанные слова. Латентный семантический анализ использует сингулярное разложение;

2. Вероятностная скрытая семантическая модель, основная задача которой состоит в том, чтобы идентифицировать и различать контексты употребления слов, не прибегая при этом к словарю. Такая модель позволяет устранить неоднозначность слов, то есть ситацию, когда слово может иметь много значений. Также данная модель раскрывает тематические сходства, группируя вместе слова, которые имеют общий контекст;

3. Немарковский метод непрерывного времени. Данная модель в отличии от рассмотренных выше идентифицируют тему во времени. Это очень важно, так как с течением времени темы текстов эволюционируют. Поэтому важно моделировать эволюцию темы, чтобы люди могли идентифицировать темы во времени и видеть, как они развиваются с течением времени. Данная модель моделирует темы и их изменения с течением времени, принимая во внимание как шаблон совместного появления слов, так и время. То есть в этой моделии тема рассматривается как связанная с непрерывным распределением во времени;

Тематическое моделирование также состоит из четырех этапов:

1. Извлечение данных;

2. Предварительная обработка данных;

3. Предварительный анализ, который подразумевает использование методов машинного обучения для классифиыкации текста на темы;

4. Графическое представление полученных результатов с целью дальнейшего анализа.

Глава 2. Сравнительный анализ инструментов разработки

В данной главе будут рассматриваться современные инструменты разработки веб приложений. Как было сказано раннее в современном мире возникает потребность в удобном сервисе, который будет предоставлять пользователю возможность получить визуализацию дискуссии в конкретном сценарии, согласно выбранным фильтрам, а также будет предоставлять возможность проанализировать дискуссию с помощью полученной визуализации. Пользователями системы являются эксперты из разных областей, не связанных со сферой информационных технологий и областью анализа социальных сетей, а именно эксперты из области социологии и маркетинга, поэтому важно, чтобы сервис соответствовал следующим критериям:

1. Сервис должен быть интуитивно понятной для пользователей;
2. Сервис не должен требовать дополнительных знаний для работы с ним;
3. Сервис должен выдавать быстрый отклик на действия пользователей;
4. Модули сервиса должны быть автономными и не должны зависеть от других модулей;
5. Каждый программный слой должен выполнять одну обязанность.

Для того, чтобы создать сервис, который удовлетворял бы выставленным критериям, важно выбрать подходящие инструменты разработки, которые позволяют реализовать такую систему и при этом являются качественными и проверенными разработчиками в других больших проектах. Чтобы сделать правильный выбор, важно провести сравнительный анализ доступных вариантов.

2.1 Сравнительный анализ инструментов, позволяющих создать слой пользовательского представления данных

Слой представления пользовательских данных отвечает отображение данных, которые видит пользователь в браузере, а также предоставляют поль-

зователям возможность взаимодействовать с приложением в целом. Раньше слой представления пользовательских данных разрабатывали с помощью языка программирования JavaScript. Все элементы страницы создавались с помощью HTML тегов в отдельных HTML файлах, а разработчикам нужно было с помощью JavaScript найти элемент на странице, добавить на него обработчик событий и затем обработать пользовательское событие. Такой способ подходит для небольших приложений и имеет следующие проблемы:

1. Проблема масштабирования приложения;
2. Проблема поддержки большого приложения;
3. Проблема повторного использования кода.

Данные проблемы решают библиотеки и фреймворки

2.1.1 Библиотека React

Библиотека React основывается на компонентом подходе. Компонент представляет собой отдельную сущность, состоящую из HTML элементов и JavaScript кода. Компоненты удобно: комбинировать, поддерживать, переиспользовать, тестировать.

Компоненты создаются с помощью React элементов, которые представляют собой объекты, содержащие тип создаваемого HTML элемента, свойства этого элемента и дочерние элементы этого элемента. Для удобной работы с React элементами используется JSX - расширение над JavaScript, которое преобразует HTML подобный код в React элементы.

Компонент представляет собой функцию, которая на вход принимает свойства элемента – объект props и возвращает JSX разметку, которая преобразуется в React элемент. Компонент имеет свое локальное состояние, то есть данные, которые нужны для работы с этим компонентом и не доступны другим компонентам. Состояние компонента можно менять, после чего будет происходить перерендер. При изменении объекта props так же будет происходить перерендер компонента. Для обработки пользовательских действий React предоставляет специальные атрибуты элементам и компонентам, куда

нужно передавать функцию обработчик. Такой способ позволяет избежать утечек памяти, так как при удалении компонента из дерева отрисовки, обработчик удаляется вместе с ним. React так же использует виртуальное DOM дерево, которое составляется из React компонентов, затем библиотека React DOM рендерит это дерево в реальное DOM дерево в браузере. Такой подход позволяет сократить количество перерисовываний, так как React сравнивает виртуальный DOM и реальный DOM и перерисовывает только то, что нужно. Единственным минусом React является отсутствие встроенной архитектуры построения приложения в целом. React отвечает только за слой представления данных, а программист сам должен решать, как выстраивать архитектуру приложения и как вставить в эту архитектуру React.

2.1.2 Фреймворк Angular

В отличии от React, Angular – это фреймворк, то есть набор хорошо интегрированных библиотек, охватывающих широкий спектр функций: маршрутизация, управление формами, клиент-серверное взаимодействие и т. д. Angular основан на компонентном подходе, но компонент в Angular это не тоже самое, что компонент в React. Компонент в Angular – это строительный блок приложения, который состоит из двух составляющих:

- Класс компонента – это место, где хранится логика, которая нужна компоненту. Этот код должен включать функции, обработчики событий, свойства, и ссылки на сервисы;
- HTML шаблон. У каждого компонента есть HTML-шаблон, в котором определяется то, что этот компонент будет отображать. Шаблон можно задать либо строкой, либо как путь к файлу. Angular расширяет HTML дополнительным синтаксисом, который позволяет вставлять динамические данные в компонент. Angular автоматически обновляет DOM, когда состояние компонента изменяется;

Angular предоставляет динамическое изменение UI с помощью двух директив, то есть команд, которые можно использовать в HTML, чтобы повлиять на изменения в приложении:

1. Директива *ngFor – это Angular итератор, который дает возможность динамически создавать DOM элементы на основе элементов массива;

2. Директива *ngIf, используется для добавления и удаления DOM элементов в зависимости от условия;

   Angular приложения следуют паттерну MVVM, где:

- Model хранит данные приложения, в Angular это класс, созданный с помощью TypeScript, он передается в качестве параметра классу компонента;

- View отвечает за отрисовку данных, в Angular это HTML шаблон;

- View-Model содержит логику компонента и отвечает за синхронизацию Model и View, в Angular это класс компонента.

2.1.3 Итоги сравнения

	React	**Angular**
Масштабируемость	Подходит для средних и небольших команд, но при этом приложения легко масштабируются, и размер команды так же легко масштабируется	Больше подходит для больших команд, в небольших проектах в применении нет смысла из-за большого количества неуместного шаблонного кода
Скорость разработки	Разработка происходит быстрее из-за отсутствия шаблонного кода	Разработка занимает больше времени из-за необходимости соблюдать шаблоны, который предоставляет фреймворк
Область применения	Больше подходит для пользовательских приложений, в которых данные часто обновляются	Больше подходит для корпоративных приложений с большой бизнес логикой и не большим количеством обновлений пользовательских данных
DOM	Использует виртуальное DOM дерево, составленное из компонентов, затем происходит сравнение виртуального и реального DOM и перерисовывается только измененная часть	Напрямую работает с реальным DOM, это значит, что при изменении компонента происходит обновление всего DOM
Производительность	Имеет гораздо меньший вес, а так же лучшую оптимизацию перерисовок за счет виртуального DOM, поэтому приложения более производительные	Приложения имеют больший вес из-за большого количества встроенных в фреймворк библиотек, а также из-за взаимодействия с реальным DOM приложения менее производительны

Таблица 1 – Результаты сравнения библиотек

React дает больше гибкости в приложении, позволяет быстрее и легче разрабатывать приложение, обладает большой экосистемой библиотек, которые можно подключить для других ответственностей в приложении, а так же обладает лучшей производительностью за счет виртуального DOM, поэтому в качестве инструмента, позволяющего создать слой пользовательского представления данных, в проекте был выбран React.

2.2 Сравнительный анализ инструментов визуализации

Так как пользователями сервиса являются эксперты, которые будут использовать визуализацию для дальнейшего анализа, то необходимо выбрать инструмент, который позволяет создать интуитивно понятную, настраиваемую виузализацию дискуссий в социальных сетях, которая при этом сможет помочь отследить эволюцию дискуссии во времени. Также инструмент должен быть совместим с бибилиотекой представления пользовательских данных React. Были расмотренны следующие инструменты:

1. Библиотека по визуализации D3. Это JavaScript библиотека для визуализации данных низкого уровня, которая перекладывает большую ответственность на разработчика. Данная библиотека предоставляет доступ к своим внутренним инструментам, которые необходимо конфугировать и настраивать нужным способом для дальнейшего использования, что с одной стороны вызывает трудности, а с другой стороны предоставляет большую гибкость. Также данная библиотка не является React ориентированной, поэтому также требует большего времени для настройки работы вместе с бибилиотекой React. Таким образом, при работе с данной бибилиотекой возникают сложности при визуализации большого числа данных по дискуссиям во времени;

2. Библиотека Echarts. Это инструмент, который содержит в себе готовые настраиваемые React компоненты графиков. Также предоставляемые компоненты графиков являются адаптивными. Производительность данной библиотеки также является ее преимуществом, несмотря на то, что с увеличением числа точек производительность снижается.

Инструмент предоставляет доступ к практически всем нужным графикам визуализации дискуссий в социальных сетях. Все графики дополнительно настраиваются с помощью различных параметров. Также Echarts предоставляет масштабирование, трансформацию во времени и динамическую подргузку данных, что очень важно при визуализации миллионов данных во времени;

Таким образом, из-за подходящего встроенного функционала, удобной настройки, совместимости с библиотекой React, а также хорошей производительности, в качестве инструмента визуализации была выбрана библиотека Echarts.

Глава 3. Разработка сервиса визуализации дискуссий в социальных сетях

Для того, чтобы реализовать качественный сервис, необходимо проанализировать уже существующие варианты.

3.1 Анализ существующих решений

В сети существуют библиотеки для построения графиков и приложения по мониторингу репутации бренда в социальных сетях, но нет библиотек или приложений, позволяющих найти дискуссию по ключевым параметрам, визуализировать ее разными способами и получить по ней анализ. Поэтому актуальность подобного сервиса не вызывает сомнений, особенно сейчас во время импортозамещения. На данный момент существуют следующие сервисы.

3.1.1 Медиалогия

Медиалогия. Сервис, позволяющий проводить мониторинг в социальных сетях, а именно оценивать мнения о бренде, изучать целевую аудиторию. Анализ строится на основе медиа индекса, который включает в себя цитируемость источника, классификацию мнений о бренде на позитивные, нейтральные и негативные, а также степень значимости упоминания. Таким образом, платформа позволяет лишь найти мнения о бренде, составить топ цитирующих СМИ и провести анализ о популярности бренда.

Из продвинутого анализа данная платформа предоставляет лишь визуализацию анализа тональности с негативным, позитивным и нейтральным оттенком в виде линейного графика, который позволяет отследить, сколько было позитивных, негативных и нейтральных упоминаний бренда в конкретную дату, а также в виде кругового графика, который показывает процентное соотношение упоминаний о бренде за выбранный период.

Таким образом, сервис не позволяет удовлетворить потребности экспертов в том, чтобы найти дискуссии по настроенным фильтрам, провести продвинутый анализ с помощью различных методов глубокого обучения и

Рисунок 1 – График анализа тональности

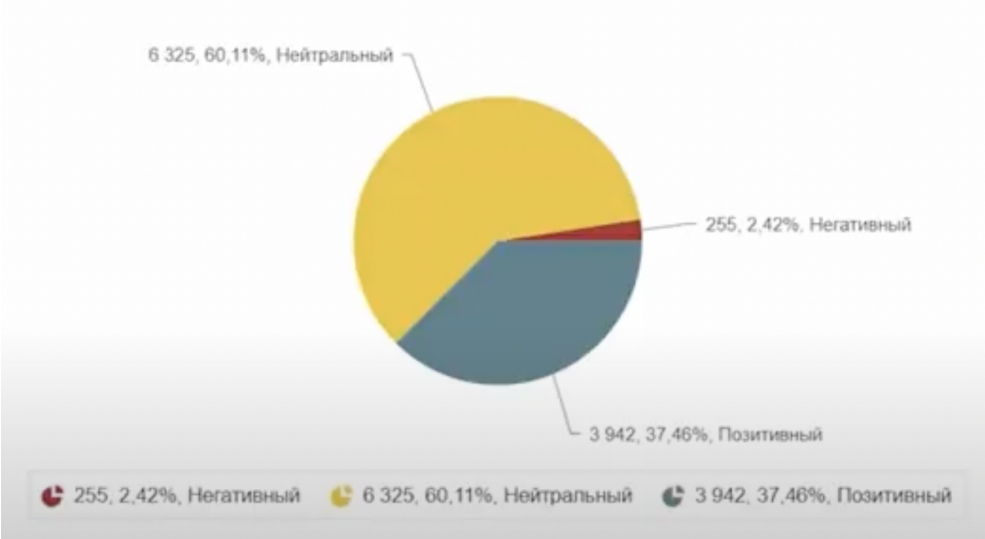

Рисунок 2 – Круговой график анализа тональности

получить их визуализацию.

3.1.2 Brand analytics

Brand analytics. Система мониторинга и анализа социальных медиа и СМИ. Сервис позволяет оперделить, что говорят о бренде и конкурентах в социальных медиа, а также что ценят клиенты в бренде и что мешает дальнейшему росту популярности бренда. Из продвинутого анализа данная платформа, как и Медиалогия, может предоставить экспертам лишь анализ тональности упоминаний о бренде с негативной и позитивной окраской. Данная платформа предоставляет лишь один линейный график сентимент анализа.

Круговой гарфик с общей статистикой отсутствует, вместо него сервис предлагает пользователям добавлять на график отдельную линию, отвечающую за общее количество упоминаний, что является менее удобным и наглядным вариантом.

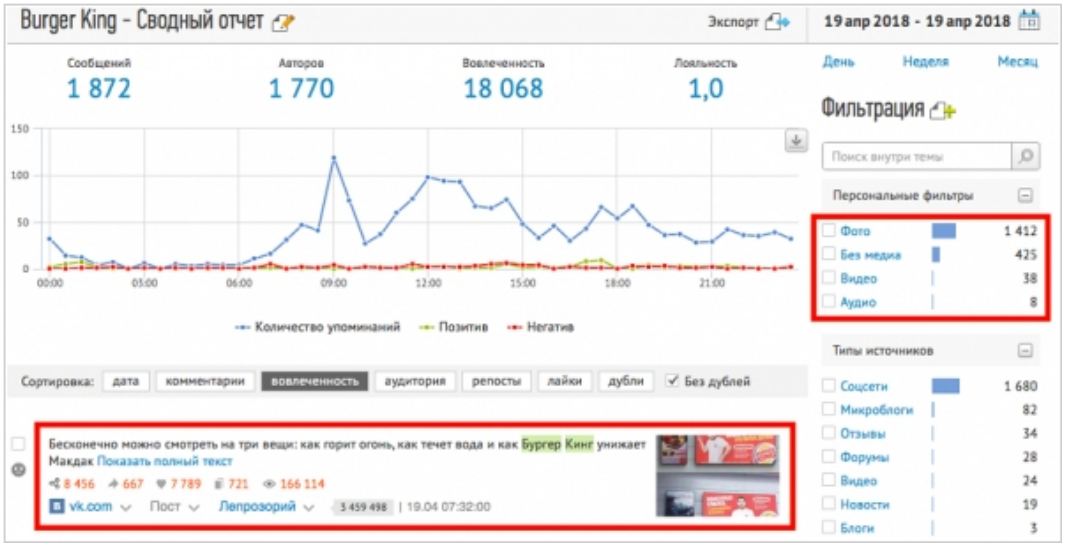

Рисунок 3 – График анализа тональности

Данный сервис аналогично не предоставляет экспертам возможность удобным и наглядным способом проанализировать любую дискуссию в социальных сетях во времени.

3.1.3 YouScan

YouScan. Это платформа также, как и описанные выше платформы предназначена для анализа бренда в социальных сетях. Отличительной особенностью данного сервиса является наличие возможности анализа изображений в социальных сетей на основе искусственного интеллекта, которая помогает предприятиям анализировать мнения потребителей, находить полезную информацию и управлять репутацией бренда. Из достоинств платформы можно выделить несколько видов визуализации анализа тональности и возможность выявления инфлюенсеров, которой нет у описанных выше платформ.

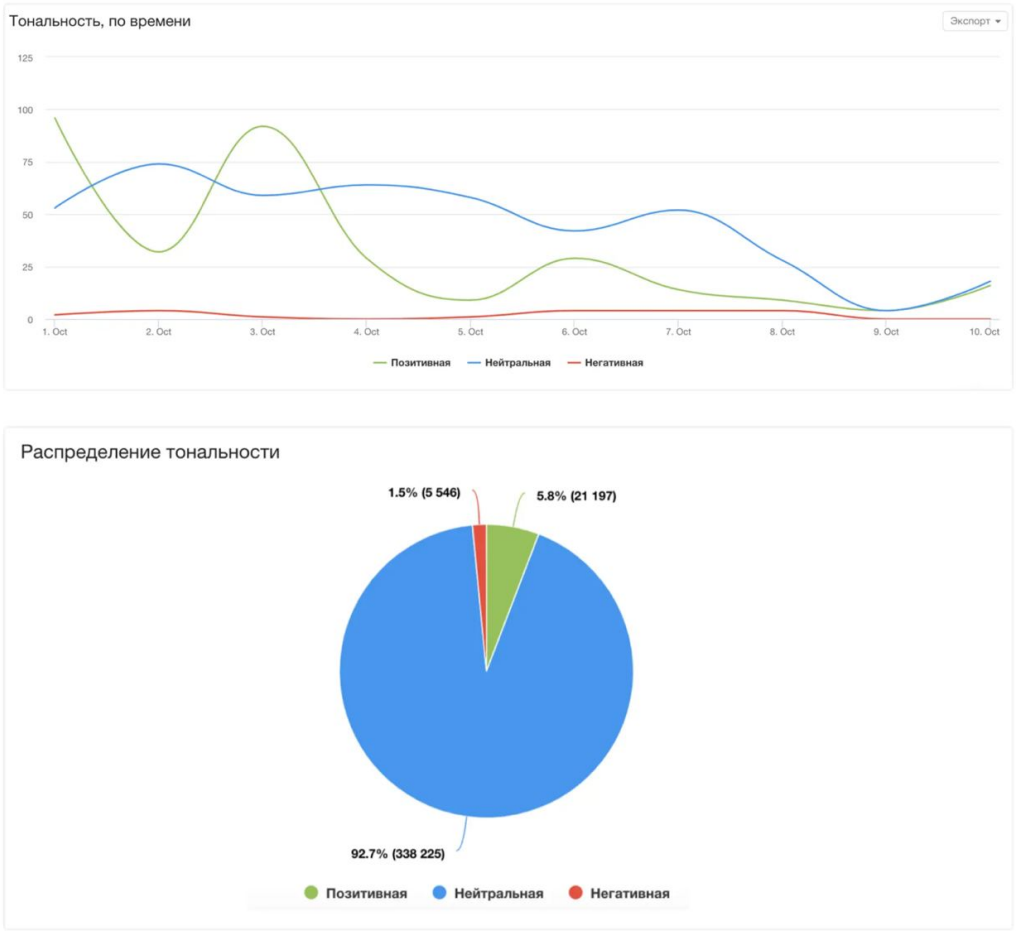

Рисунок 4 – Графики анализа тональности

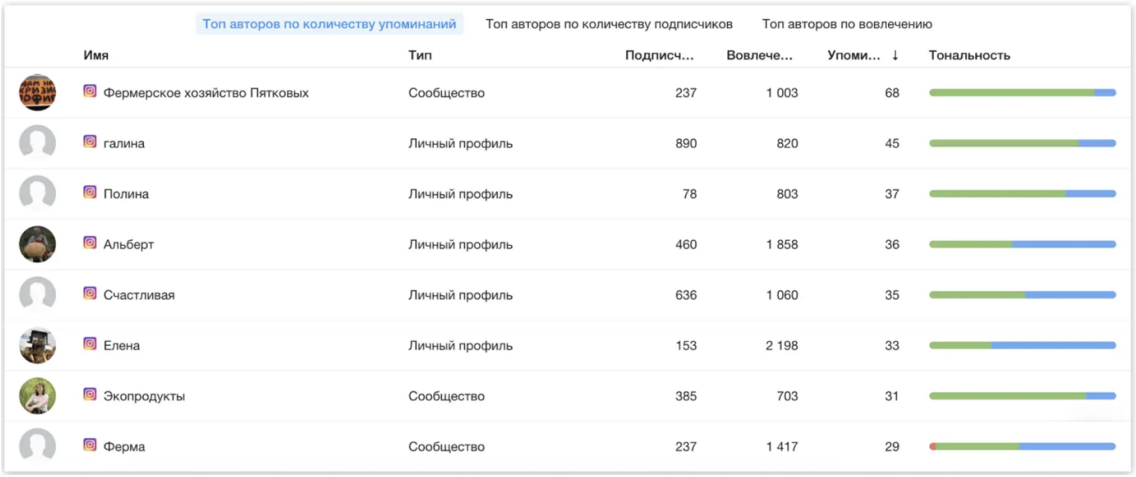

Рисунок 5 – Выявление инфлюенсеров

Данный сервис также как и выше описанные не предоставляет возможность проанализировать целую дисскуссию по ключевым словам и получить продвинутый анализ по ней, а только предлагает хорошие способы анализа конкретного бренда.

3.1.4 Brand24

Brand24 - это иностранный инструмент для анализа бренда в социальных сетях. Данный сервис помогает отслеживать по ключевым словам все, что связано с брендом: количество упоминаний, охват в социальных сетях, анализ настроений, вовлеченность, инфлюенсеров.

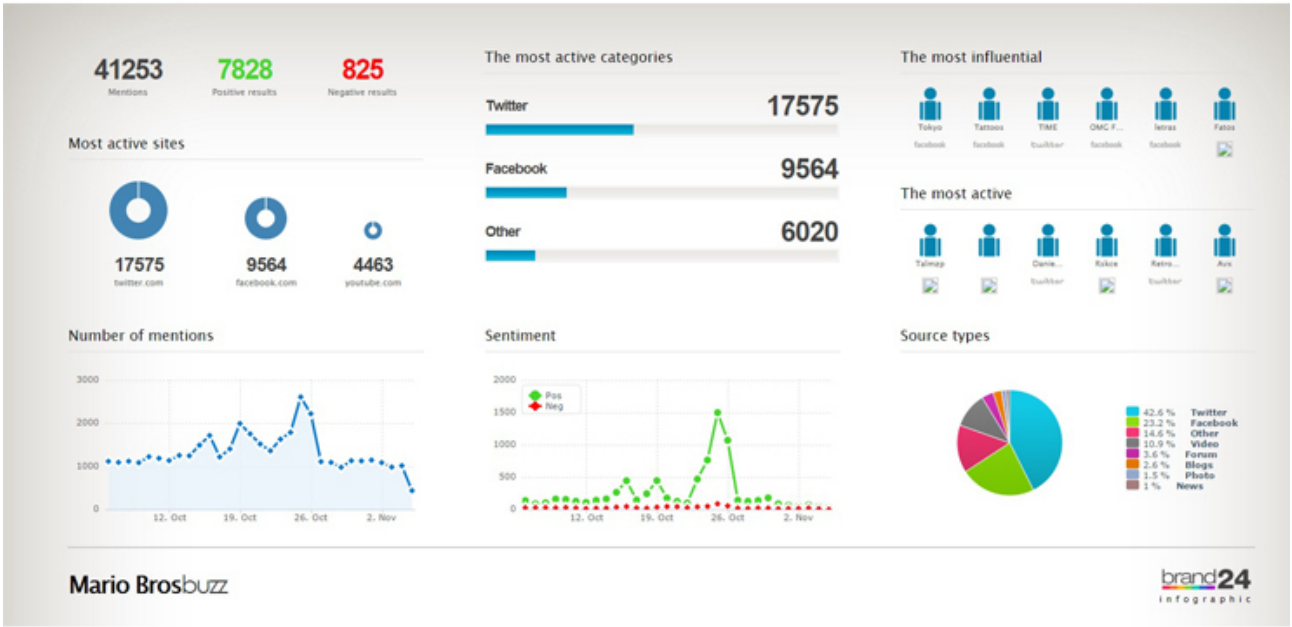

Рисунок 6 – Доступные виды анализа Brand24

Данная платформа ничем не выделяется на фоне отечественных конкурентов, а в визуализации анализа тональности уступает. Аналогично Brand24 анализирует только мнение о конкретном бренде и не предоставляет возможность проанализировать любую дискуссию.

3.1.5 Итоги

Таким образом, на данный момент существуют лишь системы, позволяющие провести анализ популярности бренда в социальных сетях. Некоторые из данных платформ обладают хорошей визуализацией анализа тональности и визуализацией найденных инфлюенсеров. Но ни один из существующих сервисов не позволяет с помощью методов глубокого анализа по выбранным фильтрам и введенным фразам найти дискуссию в социальных сетях, представить ее в графическом виде, провести анализ по ней, а также посмотреть на ее эволюцию во времени с помощью временных графиков.

3.2 Разработка сервиса

Люди всегда проводят дискуссии как при личной встрече, так и в онлайн формате. В какой-то момент в мире происходят какие-либо события и когда они произошли, люди сразу начинают обсуждать их в социальных сетях и мессенджерах. Но при этом люди так же проводили дискуссии до событий и во время событий. Поэтому возникает потребность в визуализации и анализе дискуссии во времени. Как было описано ранее, сервиса предоставляющего подобную функциональность не существует, поэтому была начата разработка подобного сервиса.

3.2.1 Функциональные требования

1. Наличие модуля создания кейса
2. Наличие модуля просмотра активных кейсов с отображением статуса кейса;
3. Наличие модуля с отображением важных инфлюенсеров с их кратким описанием;
4. Наличие модуля с продвинутым анализом, который включает в себя визуализацию тематического моделирования, аналиа тональности во времени, а также аспектную составлюящую данного анализа.

3.2.2 Модуль создания кейса

Прежде всего был реализован модуль, позволяющий эксперту задать настройки для поиска дискуссии, чтобы впоследствии ее можно было обработать, вывести статистику и визуализировать. Данный модуль предоставляет возможность эксперту идентифицировать кейс с помощью названия и описания, а также выбрать фильтры, определяющие сбор данных. Такими фильтрами являются:

- Период сбора информации, состоящий из даты начала и даты конца сбора;

- Источник сбора, то есть социальная платформа, где будут собираться данные. На данный момент система предоставляет сбор данных с видео хостинга YouTube, мессенджера Telegram, а также из социальных сетей Twitter, Вконтакте, Facebook, Одноклассники, Instagram;
- Событие, произошедшее на социальной платформе, которое описывается либо в виде тега, либо ключевым понятием, состоящим максимум из двух слов

После задания настроек кейс можно либо сохранить, либо сохранить и сразу запустить работу краулера по сбору информации. Внешний вид данного модуля представлен на Рисунке 7.

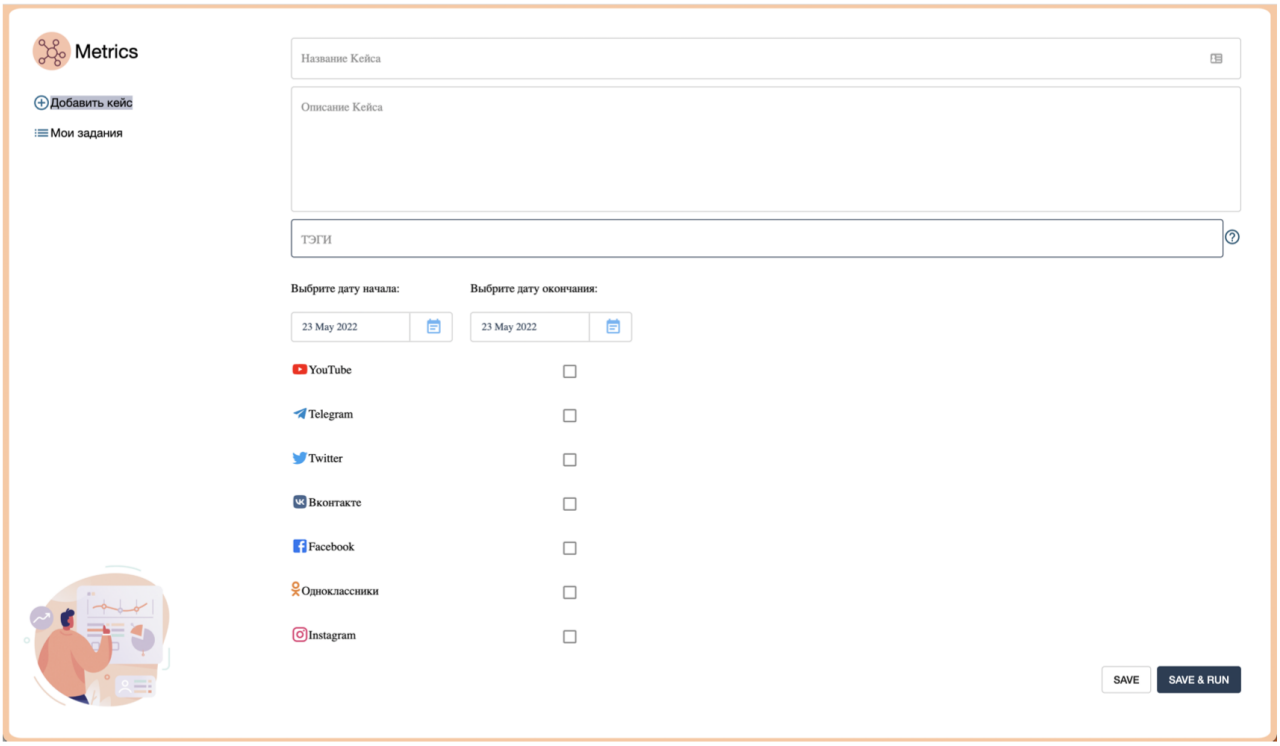

Рисунок 7 – Модуль создания кейса

После идентификации и задания нужных фильтров, кейс передается на модуль сбора данных, а затем на модуль анализа дискуссии.

3.2.3 Модуль просмотра активных кейсов

Информация в модуле анализа может обрабатываться очень долго, поэтому важно показывать эксперту статус выполнения задачи. Для этого был создан модуль просмотра активных кейсов, который позволяет эксперту отследить прогресс анализа кейса, указанный в процентах, а также статус анализа кейса, то есть этап, на котором в данный момент проходит анализ. Задача может находится в одном из следующих статусов:

- waiting to start - кейс сохранен, но работа краулера по сбору информации не запущена;
- crawling - кейс сохранен и запущена работа краулера по сбору информации;
- sentiment analysis - работа краулера завершена, проходит анализ тональности и тематическое моделирование;
- analysis is ready - анализ кейса полностью готов.
- error - на одном из этапов произошла ошибка

Когда задача полностью готова эксперт может нажать на кейс и перейти на модули с результатом анализа. Внешний вид данного модуля представлен на Рисунке 8.

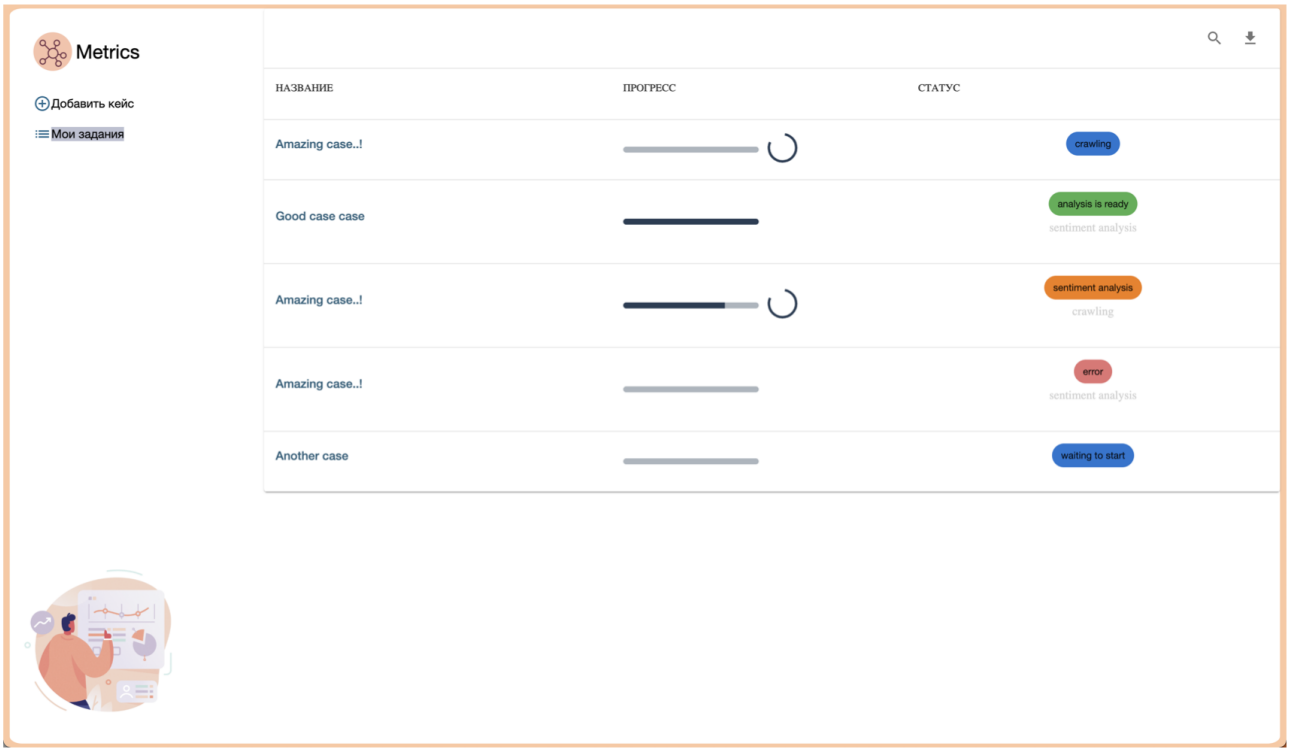

Рисунок 8 – Модуль просмотра активных кейсов

3.2.4 Модуль отображения инфлюенсеров

Часто экспертам необходимо понять, кто из участников дискуссии был самым заметным по какому-либо критерию и проанализировать причины этого. Возможно, экспертам придется перейти на страницу данного человека либо для того, чтобы найти больше факторов его заметности в дискуссии, либо для того, чтобы направить людей для сотрудничества с ним. Сервис предоставляет эксперту отображение самых популярных инфлюенсеров по одному из критериев: количество постов, количество лайков, количество репостов, количество людей вовлеченных в дискуссию с этим человеком. Внешний вид данного модуля представлен на Рисунке 9. Результаты можно экспортировать.

При выборе одного из пользователей появляется модальное окно с краткой информацией о пользователе: имя и фамилия, количественные показатели по каждому из вышеперечисленных параметров(Рисунок 10).

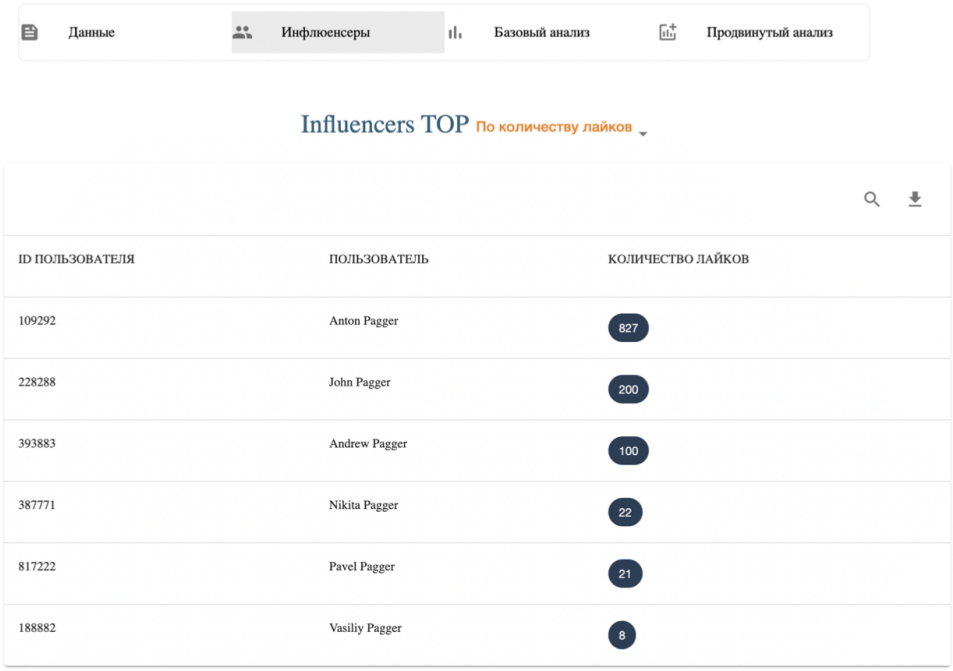

Рисунок 9 – Модуль отображения инфлюенсеров

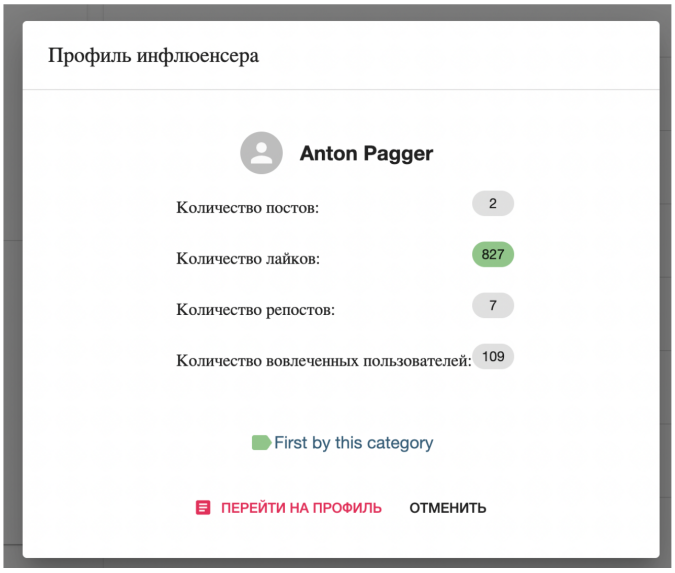

Рисунок 10 – Окно с информацией о пользователе

3.2.5 Модуль продвинутого анализа

Данный модуль предоставляет экспертам понятную и легко интерпретируемую визуализацию методов искусственного интелекта, с помощью которой они смогут как сразу получить наглядное представление результатов анализа, так и провести дальнейший анализ, используя в качестве основы данную визуализацию. Экспертам доступны следующие визуализации:

1. Визуализация анализа тональности. Экспертам доступен круговой график, который показывает какое количество постов и сообщений имели положительный, негативный или нейтральный оттенок. Также с помощью графика зависимости тональности текстовых сообщений в количественном соотношении от времени эксперты смогут понять, сколько постов и сообщений имели положительный, негативный или нейтральный окрас в конкретный момент времени или за конкретный настраиваемый период времени, то есть эксперту доступна настройка частоты обновлений данных графика по часам, дням, неделям, кварталам, годам, что позволяет экспертам проследить эволюцию дискуссии во времени(Рисунок 11). Данные графиков доступны для выгрузки.

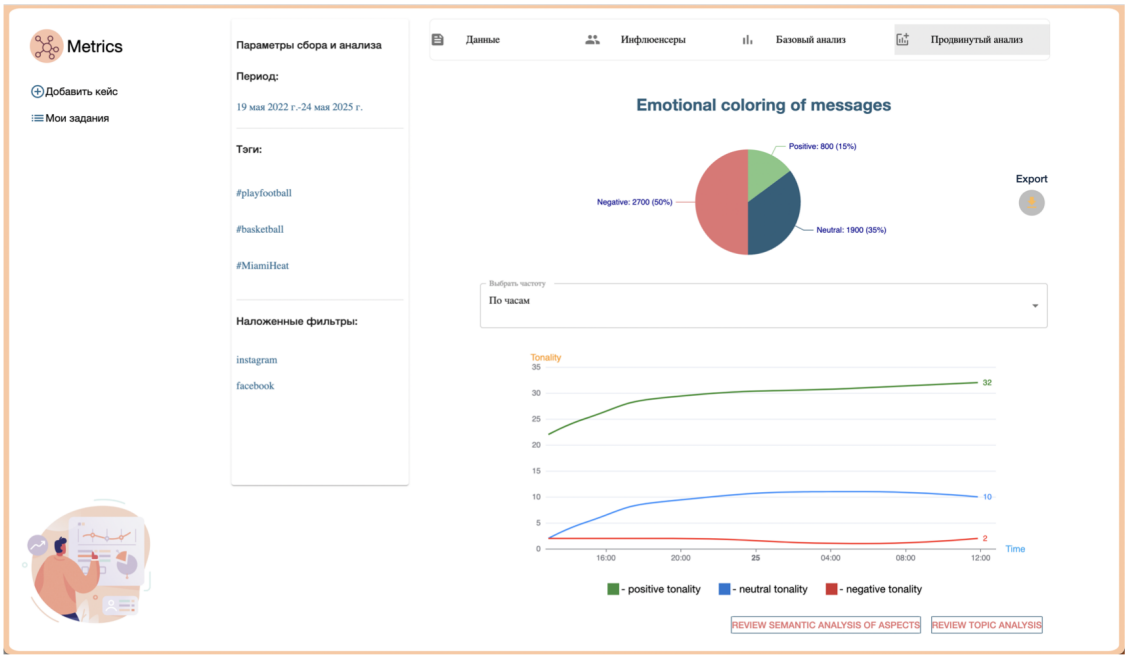

Рисунок 11 – Круговой и линейный графики анализа тональности

Также экспертам доступна возможность посмотреть аспектную составляющую анализа тональности, то есть эксперту предоставляется таблица с самыми частыми словами относительно которых пользователь выражается позитивно, негативно или нейтрально(Рисунок 12).

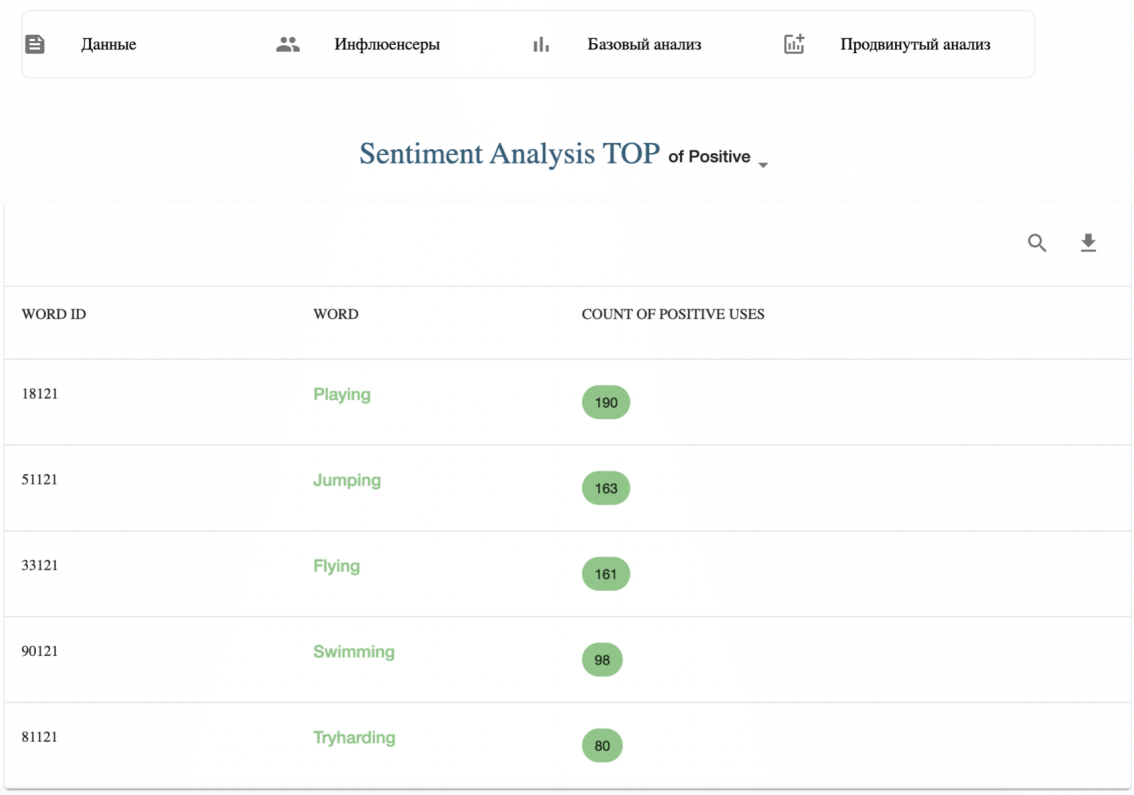

Рисунок 12 – Результаты аспектного анализа тональности

При выборе определенного слова появляется модальное окно (Рисунок 13) со списком фраз, в которых встречалось это слово с выделением самого слова для более удобного анализа экспертом. Также у эксперта есть возможность экспортировать проанализированные данные;

Рисунок 13 – Информация о выбранном слове с учетом тональности

2. Вторым анализом, визуализация которого доступна эксперту является тематическое моделирование, с помощью которого выявляются скрытые темы в дискуссиях. Возможны ситуации, когда в дискуссии миллионы сообщений, а тем всего две. Для того чтобы узнать скрытые темы, эксперту предоставляется таблица с рейтингом тем, на которые чаще происходили дискуссии в социальных сетях, найденные по заданным параметрам.

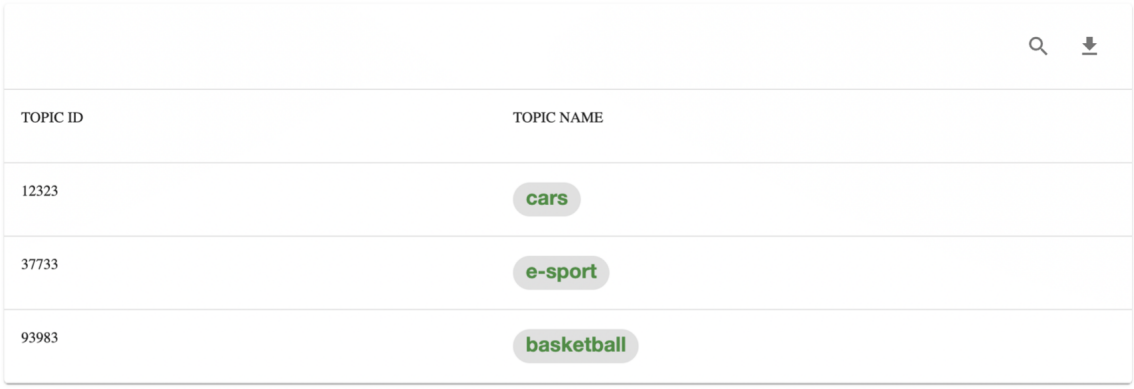

Рисунок 14 – Таблица рейтинга найденных тематик дискуссии

Для того, чтобы получить более подобрный анализ по теме, эксперту предоставляется возможность при нажатии на тему увидеть статистику, то есть самые популярные термины в этой теме. Синим цветом обозначается частота этого термина в этой теме, а голубым обозначается частота этого термина во всех темах.

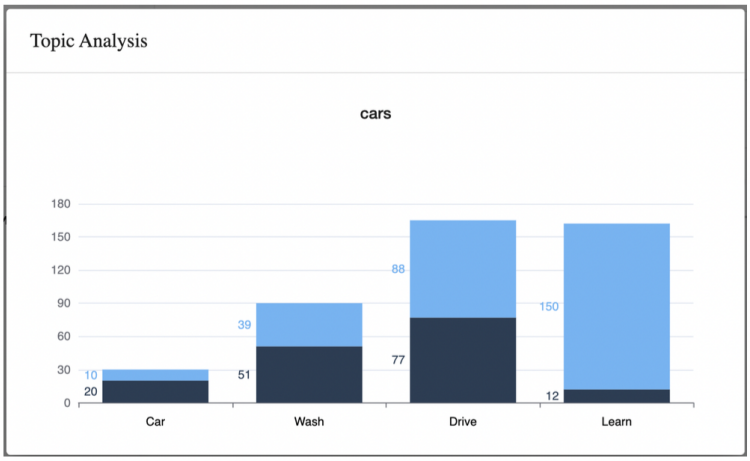

Рисунок 15 – Соотношения присвоений термина к определенной теме

3.3 Итоги

В результате проделанной работы все поставленные функциоанльные требования были выполнены, это отражено в таблице ниже.

Функциональные требования	**Результат**
Наличие модуля создания кейса	Выполнено
Наличие модуля просмотра активных кейсов	Выполнено
Наличие модуля с отображением инфлюенсеров	Выполнено
Наличие модуля с продвинутым анализом	Выполнено

Таблица 2 – Результаты выполнения функциональных требований

Выводы

В результате работы с помощью современных методов программирования пользовательских веб приложений были разработаны модули сервиса по анализу дискуссий в социальных сетях, которые позволяют экспертам из разных областей получить наглядное и легко интерпретируемое представление результатов алгоритмов машинного обучения по анализу тональности, тематическому моделлированию и нахождению инфлюенсеров, что поможет экспертам высказывать и подтверждать различные гипотезы.

Реализованный сервис полностью соответствует функциональным требованиям, описанным в подпункте 3.2.1 главы 3. Исходный код представлен в репозитории: https://gitlab.com/ilafila/metrics.

Заключение

Результаты работы

В рамках данной выпускной квалификационной работы были выполнены следующие задачи:

1. Проведен обзор существующей научной литературы и изучена проблематика современного анализа социальных сетей;
2. Проведен сравнительный анализ инструментов, позволяющих реализовать слой пользовательского представления данных в сервисе;
3. Проведен сравнительный анализ инструментов визуализации;
4. Проведен сравнительный анализ существующих сервисов по анализу социальных сетей;
5. Реализованы модули системы, позволяющие провести визуализацию и последующий анализ дискуссий в социальных сетях, исходный код которых представлен в репозитории: https://gitlab.com/ilafila/metrics.

Перспективы развития

Данная дипломная работа несет следующие перспективы развития:

- Полное перевод приложения на английский язык
- Тестирование работы сервиса на большем числе дискуссий;
- Добавление новых видов анализа и визуализации, а именно суммиразации и анализа, основанного на геопозиции;
- Усовершенстование уже существующих визуализаций анализа тональности и тематического моделирования;
- Сдача сервиса в экплуатацию экспертам.

Список литературы

[1] Документация к библиотеке React для языка программирования JavaScript [Электронный ресурс], Режим доступа – свободный: https://ru.reactjs.org/

[2] Документация к библиотеке Angular для языка программирования JavaScript [Электронный ресурс], Режим доступа – свободный: https://angular.io/

[3] Документация к библиотеке Redux для языка программирования JavaScript [Электронный ресурс], Режим доступа – свободный: https://redux.js.org/

[4] Документация к библиотеке MobX для языка программирования JavaScript [Электронный ресурс], Режим доступа – свободный: https://mobx.js.org/

[5] Social Network Analysis: Methods and Applications, Авторы: Stanley Wasserman, Katherine Faust, Stanley (University of Illinois Wasserman, Urbana-Champaign) pp 94 – 105, 177 – 188, 1999.

[6] Different Aspects of Social Network Analysis, Mohsen Jamali, IEEE 2006, DOI:10.1109/WI.2006.61.

[7] The SAGE Handbook of Social Network Analysis, John Scott, Peter J. Carrington pp 9 – 11.

[8] Social network analysis with SNA, Carter T. Butts, University of California, Journal of statistical software 24(i06), 2008, DOI:10.18637/jss.v024.i06

[9] Social network analysis: developments, advances, and prospects, John Scott, Springer-Verlag 2010, DOI:10.1007/s13278-010-0012-6

[10] Dohaiha, H.H., Prasad, P., Maag, A., et al.: ‘Deep learning for aspect-based sentiment analysis a comparative review’, Expert Syst. Appl., 2018, 118, pp. 272–299

[11] Ortis, A., Farinella, G.M., Battiato, S.: ‘An overview on image sentiment analysis: methods, datasets and current challenges’, Proc. of the 16th Int. Joint Conf. on e-Business and Telecommunications - Volume 1: SIGMAP, INSTICC, Prague, Czech Republic, 2019, pp. 290–300

[12] Alessandro Ortis , Giovanni Maria Farinella1, Sebastiano Battiato, Survey on visual sentiment analysis, Institution of Engineering and Technology, Department of Mathematics and Computer Science, University of Catania, 2020.

[13] David M. Blei Probabilistic topic models 2012

[14] Shuhui Jiang; Xueming Qian; Jialie Shen; Yun Fu; Tao Mei Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations 2015

[15] AlexandraAmado, PauloCortez, PauloRita, Sergio Moro. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis, European Research on Management and Business Economics

[16] Edward R., Tufte The Visual Display of Quantitative Information, 2001

[17] Нейтена Яу, Искусство визуализации в бизнесе, 2013

[18] David McCandless, Beautiful News: Positive Trends, Uplifting Stats, Creative Solutions, 2021