Лабораторная № 5. Обработка текстовых данных из открытых источников

Решение по ссылке

1. Постановка задачи

Анализ взаимосвязи между тональностью экономических и финансовых новостей в постах ВК и уровнем вовлеченности аудитории (лайки, репосты, просмотры).

2. Описание данных.

Для анализа были собраны данные в формате CSV, содержащие информацию, извлеченную из публикаций группы ВКонтакте "Тинькофф. Журнал". "Тинькофф Журнал" (Т—Ж) является медийным ресурсом, фокусирующимся на тематике личных финансов и образа жизни в России. Материалы публикаций охватывают широкий спектр вопросов, включая управление семейным бюджетом, защиту прав потребителей, процесс оформления налоговых вычетов, а также советы по выбору бытовой техники с целью избежать излишних трат на ненужные функции. Датасет содержит следующие атрибуты для каждого поста: текст публикации, количество лайков, репостов и просмотров. Эти данные предоставляют основу для комплексного анализа активности и вовлеченности аудитории группы, позволяя оценить популярность и релевантность представленного контента среди пользователей социальной сети.

3. Сбор данных.

В ходе лабораторной работы была разработана функция get_wall_posts, предназначенная для сбора данных постов со стены группы в социальной сети ВКонтакте через VK API. Функция принимает два аргумента: VK_TOKEN — токен для аутентификации и доступа к API, и group_id — идентификатор группы, передаваемый как строка с отрицательным значением. Для выполнения запросов использовались параметры, включая версию API, смещение для пагинации и количество запрашиваемых записей. Полученный ответ в формате JSON содержит детализированную информацию о каждом посте, которая затем обрабатывается и систематизируется в датафрейме data_posts. Датафрейм включает такие данные, как идентификаторы группы и поста, дату публикации, описание, заголовок, текст поста, а также количество просмотров, лайков и репостов. В процессе работы использовался механизм итерации с заданным шагом смещения для эффективного сбора всех доступных постов, учитывая лимиты API и вводя задержку между запросами для предотвращения превышения лимитов. Этот подход позволил эффективно собрать и организовать данные для последующего анализа.

4. Схема решения.

1) Предобработка данных

Для исследования текстовых данных из постов группы ВКонтакте "Тинькофф. Журнал", был выполнен анализ частотности слов с помощью визуализации в форме облака слов (см. рисунок 1).

Рисунок 1 – Облаков слов до предобработки

Изначальный анализ показал, что в текстах присутствует значительное количество союзов, предлогов и других стоп-слов, которые обычно считаются нерелевантными для аналитических целей, поскольку не несут значимой семантической нагрузки.

Учитывая это, было принято решение о проведении предобработки текстовых данных с целью улучшения качества последующего анализа. Этот этап включал удаление пунктуации и стоп-слов, а также проведение лемматизации и токенизации текста. Для выполнения этих задач были использованы функциональные возможности библиотек nltk и pymorphy2.

После предобработки текста были выделены следующие слова, представленные на рисунке 2.

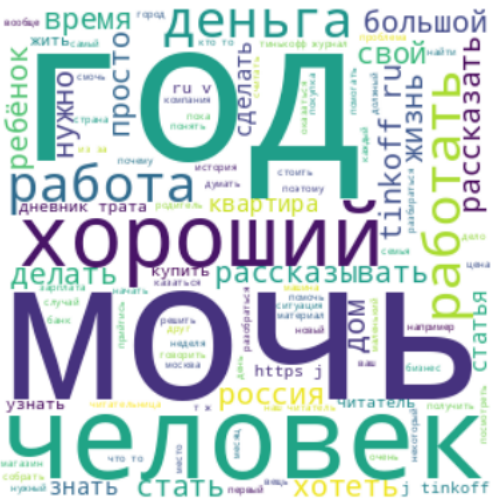

Рисунок 2 – Облаков слов после предобработки

2) Анализ тональности

Для анализа тональности текстов была выбрана модель "blanchefort/rubert-base-cased-sentiment", предоставляемая библиотекой transformers от Hugging Face. Эта модель обучена на большом корпусе русскоязычных текстов и способна классифицировать тексты по тональности.

- **Загрузка модели и токенизатора**: использовалась автоматическая загрузка модели и соответствующего токенизатора по имени модели.
- **Создание пайплайна для анализа тональности:** на основе загруженной модели и токенизатора был создан пайплайн, упрощающий процесс анализа тональности текстов.
- **Применение пайплайна к данным**: с помощью созданного пайплайна к каждому тексту в наборе данных была применена модель для определения его тональности. Результаты анализа сохранялись в новом столбце датафрейма

3) Анализ вовлеченности

После классификации текстов по тональности был проведен статистический анализ для определения влияния тональности на показатели вовлеченности аудитории. Использовался дисперсионный анализ (ANOVA) для оценки наличия статистически значимых различий в средних значениях показателей вовлеченности между группами текстов с различной тональностью.

5. Результаты экспериментов.

Далее получены результаты, какие тональности (положительные, нейтральные, отрицательные) получили наибольшее количество лайков, репостов и просмотров. Это может помочь понять, какой тип контента лучше всего взаимодействует с аудиторией.

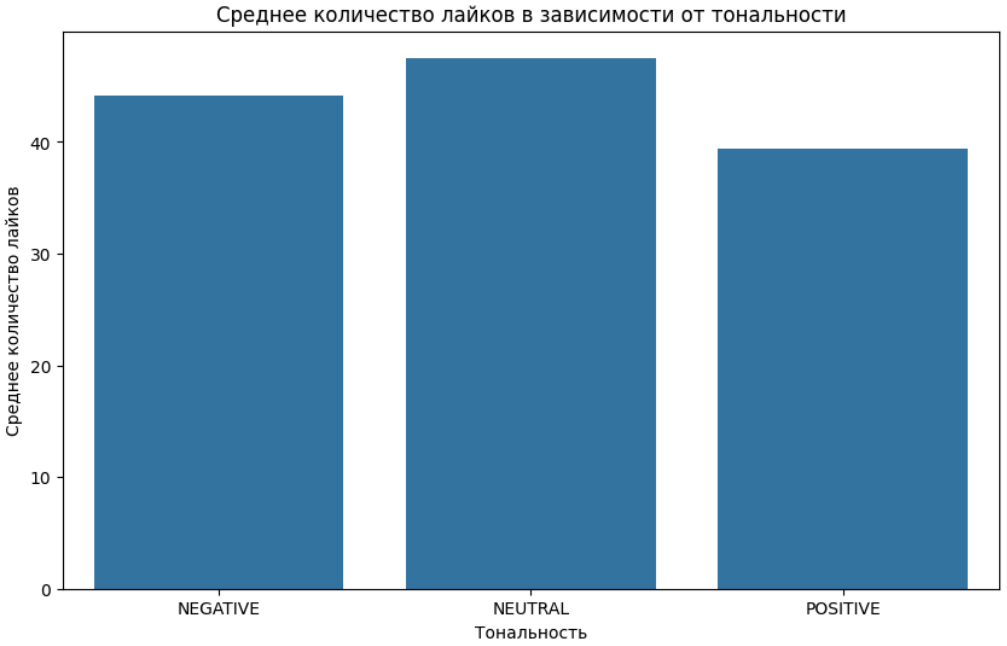

Рисунок 3 – Среднее количество лайков в зависимости от тональности

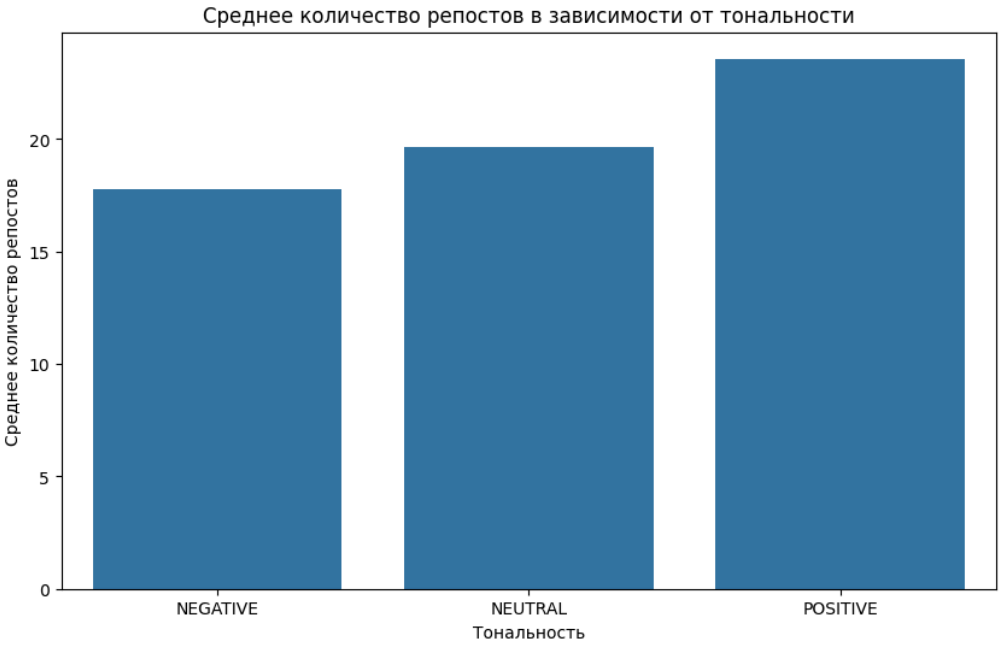

Рисунок 4 – Среднее количество репостов в зависимости от тональности

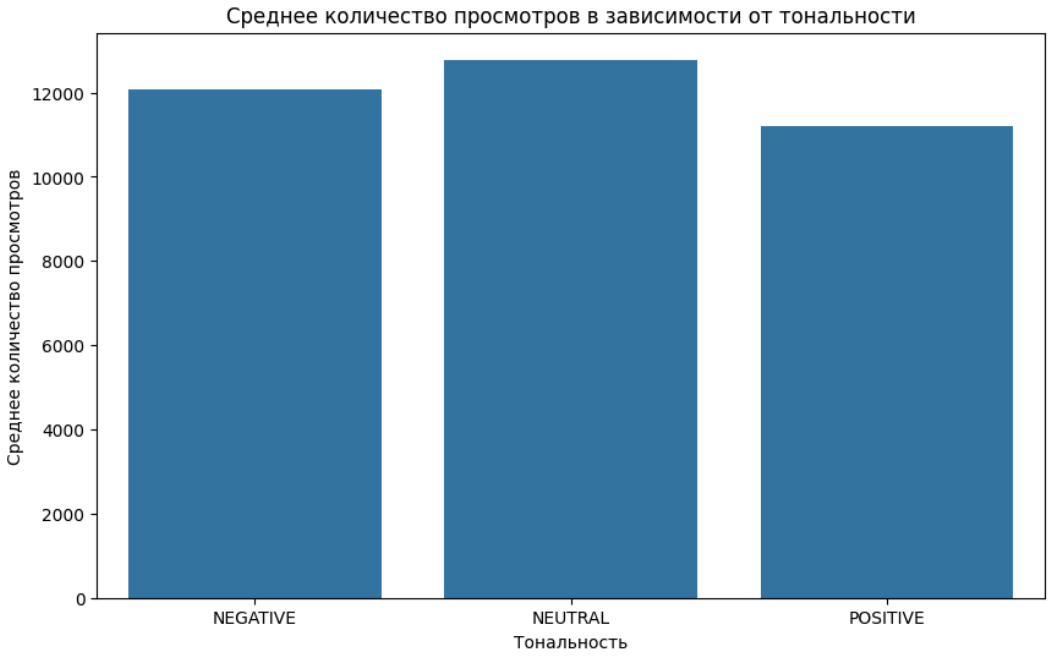

Рисунок 5 – Среднее количество просмотров в зависимости от тональности

По полученным данным видно, что позитивные посты выделяются большим количеством репостов, а нейтральные – лайков и просмотров. При этом стоит отметить, что в большей степени посты в группе Тинькофф Журнала являются нейтральными:

```
count = data['sentiment'].value_counts()

count
```

```
NEUTRAL     13916
NEGATIVE     4377
POSITIVE     2871
Name: sentiment, dtype: int64
```

Рисунок 6 – Количество постов по тональности

После проведения дисперсионного анализа (ANOVA) для оценки наличия статистически значимых различий в средних значениях показателей вовлеченности между группами текстов с различной тональностью получилось, что существует статистически значимые различия между группами по показателю "reposts" (репосты), классифицированным по тональности (положительной, нейтральной и отрицательной). Полученные результаты указывают на то, что тональность постов в группе ВК Тинькофф Журнала влияет на уровень вовлеченности аудитории в форме репостов, при этом количество лайков и просмотров практически не влияет.

6. Что можно сделать для улучшения качества решения?

Для улучшения качества решения задачи анализа взаимосвязи между тональностью экономических и финансовых новостей в постах ВК и уровнем вовлеченности аудитории можно предпринять следующие шаги:

Расширение анализа на другие показатели: кроме средних значений, можно рассмотреть использование других статистических показателей (медиана, стандартное отклонение, максимальные и минимальные значения) для более полного понимания распределения данных.

Исследование влияния времени публикации: проанализировать, как день недели, время суток и сезонность влияют на вовлеченность аудитории.

Текстовый анализ: можно также использовать методы тематического моделирования для выявления наиболее популярных тем и их связи с вовлеченностью.

7. Выводы.

Проведенный анализ показал, что тональность постов в группе ВК "Тинькофф. Журнал" статистически значимо влияет на уровень вовлеченности аудитории в форме репостов. Это указывает на то, что эмоциональная окраска контента может играть важную роль в привлечении внимания пользователей и стимулировании их к дальнейшему распространению информации.

Из данных следует, что положительные посты получают больше репостов по сравнению с нейтральными и отрицательными. Это может свидетельствовать о том, что аудитория предпочитает делиться позитивным контентом, что, в свою очередь, может способствовать формированию положительного имиджа медийного ресурса. Анализ также показал, что количество лайков и просмотров практически не зависит от тональности постов. Большинство публикаций в "Тинькофф. Журнале" имеют нейтральную тональность. Это может свидетельствовать о стремлении издания предоставлять объективную информацию и аналитику без ярко выраженной эмоциональной окраски, что важно для финансовых и экономических тем.