



**Анализ тональности отзывов на приложения в Google Play категории
«Здоровье и Фитнес» с помощью технологий больших данных**

Царькова Н.И., доцент,

Московский политехнический университет, Москва, Россия

Бова Р.Ю., магистрант,

Московский политехнический университет, Москва, Россия

Аннотация. В данной статье проводится анализ тональности отзывов в Google Play. Для анализа была построена модель классификатора на основе свёрточных нейронных сетей с помощью инструментов машинного обучения и обработки естественного языка на Python. В результате был сделан анализ комментариев нескольких из популярных приложений категории «Здоровье и Фитнес». С помощью обученной модели удалось достичь отсеивания неинформативных, эмоциональных, провокационных комментариев, что упростило процесс извлечения мнений пользователей о потенциальном приложении-конкуренте.

Ключевые слова: анализ тональности, nlp, big data, google play, отзывы

**Analysis of the tonality of reviews of applications in the Google Play category
«Health and Fitness» using big data technologies**

Tsarkova N.I., Associate Professor,

Moscow Polytechnic University, Moscow, Russia

Bova R.Yu., Master's student,

Moscow Polytechnic University, Moscow, Russia

Annotation. This article analyzes the tonality of reviews in Google Play. For the analysis, a classifier model based on convolutional neural networks was built using machine learning and natural language processing tools in Python. As a result, an analysis of the comments of several of the popular applications of the category «Health and Fitness» was made. With the help of a trained model, it was possible to filter out uninformative, emotional, provocative comments, which simplified the process of extracting users' opinions about a potential competitor application.

Key words: tonality analysis, nlp, big data, google play, reviews

Введение

При публикации приложений в Google Play важно учитывать информацию о мнении пользователей приложений. Один из наиболее простых для сбора источников являются комментарии на странице приложения. Есть потребность следить за комментариями к своим приложениям и приложениям коллег-конкурентов. Часто наиболее полезные отзывы – отрицательные. Как правило именно из них можно извлечь больше всего информации о запросах пользователей, в то время как в положительных отзывах часто никакой конкретики не пишут¹.

Из комментариев можно извлечь следующую информацию²:

- Информацию о недоработках в UI/UX дизайне и багах.
- Оценку того, как разработчики другого приложения работают со своей аудиторией, и то, как пользователи реагируют на их комментарии.
- Идеи для реализации нового уникального функционала.

Однако, стоит учесть, что не все отрицательные отзывы одинаково полезны. Существует несколько видов отзывов, полезность которых часто оказывается сомнительной²:

¹ Как и почему анализ приложений конкурентов может помочь продвинуть ваше собственное приложение [Электронный ресурс] / Режим доступа: <https://habr.com/company/appodeal/blog/298426/>

² Ушанова, М.В. Работа с негативными отзывами в сети интернет как форма продвижения компании / М.В. Ушанова // Гуманитарные технологии в современном мире: Материалы VI Международной научно-практической конференции, Калининград, 17–19 мая 2018 года. – Калининград: Западный филиал РАНХиГС, 2018. – С. 312-313.

- Эмоциональный негативный отзыв.
- «Чёрный PR».
- Троллинг.

Данные виды комментариев под приложениями мало информативны для извлечения вышеописанных данных из них. Поэтому их лучше отфильтровывать в автоматическом режиме каким-либо образом.

Таким образом, целью данной работы является анализ приложений из категории «Здоровье и Фитнес» с точки зрения наличия в них провокационной, эмоционально-негативной составляющей.

Специфика комментариев в Google Play заключается в том, что тяжело собрать частотные метрики комментатора, или собрать их вообще невозможно. Под частотными метриками в данном случае понимается, например, количество сообщений, относительный к количеству объём этих сообщений, как часто человек вступает в диалог (в виду отсутствия диалогов). В теории можно проследить, как часто человек комментирует разные приложения, но не всегда может быть возможно: в большинстве случаев мы сталкиваемся с тем, что комментатора нужно оценить по одному довольно короткому тексту.

Одним из важных факторов, которых можно оценить по одному короткому тексту, является тональность этого текста. Тональность комментария – это эмоциональная оценка, которую автор демонстрирует по отношению к объекту своего комментария. Как правило, провокационные сообщения несут в себе цель вызвать определенную, негативную эмоцию у читающего. Поэтому, часто они имеют ярко выраженную негативную тональность. Этим фактом мы можем воспользоваться, в попытке отсеять провокаторов (умышленных или нет) от более рациональных комментаторов³.

В отзывах вообще, и в отзывах на Google Play в частности, комментарии как правило развернутые, что отличает их от просто сообщений в интернете. Люди, имея возможность продумать своё сообщение, часто выдают достаточно

³ Филимонов, А.В. Применение нейронных сетей для выявления троллей в социальных сетях / А.В. Филимонов, А.В. Осипов, А.Б. Климов // Нейрокомпьютеры: разработка, применение. – 2015. – № 8. – С. 87-92.

объёмный, продуманный и репрезентативный текст, по которому можно сделать анализ тональности³.

Таким образом, этапы достижения цели исследования отзывов на провокационность, можно описать следующим списком:

1. Сбор данных о комментариях пользователей.
2. Построение на них модели классификации текстов по тональности.
3. Классификация отзывов по тональности.
4. Анализ получившихся результатов

Методы и инструменты анализа тональности текстов

Тексты можно разбивать на классы тональности по-разному. Наиболее популярны следующие наборы классов: бинарный и тернарный. Бинарный состоит из позитивного и негативного классов. Тернарный состоит из трёх классов: позитивный, нейтральный и негативный. Также существуют более точные классификации, состоящие более чем из трех уровней тональности [6].

В данной работе нам нет необходимости их использовать исходя из постановки задачи, нам достаточно будет знать, что комментарий содержит негативную окраску, а позитивный и нейтральный отзыв условно можно считать позитивным. Поэтому мы воспользуемся бинарной классификацией.

Далее нужно рассмотреть сами подходы к классификации. Их выделяют 4 (табл. 1)⁴. Первые два метода могут давать высокую точность и обладают хорошей прозрачностью выводов. Однако модели для них достаточно сложно строить, требуется очень глубокий лингвистический анализ, чтобы получить хорошую точность. Так же недостаток этих методов в том, что они не универсальны. При смене области работу по созданию модели может потребоваться делать совсем заново.

Машинное обучение без учителя позволяет обучать модель практически без специальной подготовки данных. Так же модель получается достаточно

⁴ Исследование возможности снижения репутационных рисков банковских организаций на основе анализа тональности отзывов на русском языке / Воробьева А.А., Гафарова Я.К., Самоследов Г.А.

гибкой, есть возможность использовать пред обученную модель, и до обучать её для каждой конкретной области применения, что позволяет существенно снизить затраты на переиспользование модели. Однако недостатком применения таких моделей в обработке естественного языка является то, что точность таких методов может быть невысокой, и повысить её для каждой конкретной задачи может оказаться достаточно нетривиальной задачей.

Машинное обучение с учителем так же позволяет получать достаточно гибкие модели. Они обладают высокой точностью и средней сложностью построения. Минусом таких методов является то, что необходимо тратить много сил и времени на разметку наборов данных для обучения и тестирования. Не всегда удаётся найти достаточно репрезентативный набор данных, особенно в задачах обработки естественного языка на русском, поэтому приходится делать обучающие выборки вручную. Однако, так как работа с данными ведется инженером данных самостоятельно, можно собрать действительно репрезентативную выборку данных, что положительно сказывается на точности модели. Также можно добиться снижения затрат на подготовку данных с помощью пред обучения на тематически нейтральных наборах данных. Таким образом в данном случае были выбраны методы обучения с учителем.

Таблица 1

Сравнение методов классификации тональности текстов⁴

Название метода	Сложность подготовки данных для обучения	Сложность построения модели	Универсальность	Точность
На основе правил	Не требуется	Высокая	Низкая	Высокая
На основе тональных словарей	Средняя	Высокая	Низкая	Высокая
ML без учителя	Не требуется	Высокая	Высокая	Средняя
ML с учителем	Высокая	Средняя	Высокая	Высокая

Далее нужно рассмотреть сами классификаторы, относящиеся к машинному обучению с учителем. В данной работе мы выбрали две популярные

в задачах обработки естественного языка модели. Модель на основе сверточных нейросетей (CNN), такая модель позволяет учитывать контекст нескольких программ, на которые разбит текст и модель на основе рекуррентных сетей – на основе GRU-модулей, такая модель позволяет учитывать контекст предыдущих слов относительно последующих, что позволяет получать более точные результаты на длинных текстах.

Описание набора данных для предобучения классификаторов

Основной проблемой в данной задаче является проблема подготовки данных для обучения тонального классификатора текстов на основе нейронной сети. Для того, чтобы обучить такой классификатор нужно довольно много размеченных по тональности текстов, желательно из того же источника, к данным из которого модель будет применяться при использовании. Делать это приходится как правило вручную, и это довольно затратный по времени процесс.

Также стоит заметить, что особенность моделей обработки естественного языка заключается в том, что как правило точная модель получается очень узкоспециализированная. Даже при незначительной, казалось бы, смене области применения, может наблюдаться падение точности вплоть до неудовлетворительных уровней.

В таком случае получается проблема. Мы тратим много усилий на подготовку, очистку и разметку данных для обучения, а результат наших усилий получается высоко чувствительный к изменению области применения.

Пытаясь решить эту проблему, была исследована возможность для данного конкретного случая использования других, уже размеченных наборов данных для предобучения модели и за счет этого, снижения размерности набора данных, который мы делаем вручную. В данной работе был использован датасет твитов, собранный и размеченный Рубцовой Юлией⁵. Этот набор данных носит тематически-нейтральный характер и может подойти для предобучения

⁵ Ю.В. Рубцова. Построение корпуса текстов для настройки тонального классификатора // Программные продукты и системы, 2015, №1(109), –С.72-78

классификатора тональности отзывов на Google Play. Всего он содержит чуть более 200 тысяч текстов, поделенных на негативную и позитивную тональность примерно пополам.

Сбор и разметка данных об отзывах приложений категории «Здоровье и Фитнес» из Google Play

Для того, чтобы было легче вытаскивать данные с сайта Google Play, мы воспользовались библиотекой для Python `google-play-scraper`⁶. Данный инструмент обращается к публичному `frontend-api` Google Play и получает оттуда данные о странице приложения (в том числе и об отзывах) в JSON-формате, после чего преобразует их в Python-коллекции.

Были собраны данные по приложениям категории «Здоровье и Фитнес». Одним из самых популярных представителей этой категории является приложение `Fat Secrets`. Всего было получено данных о 21 приложении из этой категории, около 60 тысяч записей о пользователях и их отзывах для обучающей выборки и около 100 тысяч записей для целевой таблицы. Формат каждой строки данных можно увидеть в табл. 2.

Набор полей данных, следующий: дата комментария, выставленный рейтинг, публичный никнейм пользователя, ссылка на его изображение профиля, текст комментария, текст ответа разработчика (если есть), количество лайков, поставленных комментарию другими пользователями.

То есть, содержит всю информацию, которая выставлена публично на странице приложения на сайте Google Play и доступна любому, даже не авторизованному пользователю. Для того, чтобы сделать набор данных более деперсонализированным, при получении данных никнейм пользователя был заменён на уникальный хэш по алгоритму `sha256`. Обучающая часть данных была размечена на классы по тональности вручную.

⁶ Страница на Github проекта библиотеки `google-play-scraper` [Электронный ресурс] / – Режим доступа: <https://github.com/facundoolano/google-play-scraper>

Описание параметров отзывов на Google Play

Название параметра	Описание	Пример значения
reviewId	Уникальный идентификатор сообщения	gp:AOqpTOGwCsBW690fw...
userName	Никнейм пользователя	Иван Иванов
userImage	Ссылка на фото профиля пользователя	https://play-lh.googleusercontent.com/EGemoI2NTX...
content	Текст комментария	Приложение очень понравилось. Есть одна ремарочка. Зачем количество съеденных калорий уменьшается пропорционально потраченным. Хотелось бы видеть обе цифры отдельно.
score	Оценка (в звёздочках от 1 до 5)	5
thumbsUpCount	Количество лайков	2
reviewCreatedVersion	Версия комментария	3.5.2
at	Timestamp когда комментарий был оставлен	2018-06-07 03:55:05
replyContent	Текст ответа от разработчика	Добрый день, не могли бы вы написать мне на почту. Мне нужно чуть больше информации от вас чтобы исправить эту ошибку.
repliedAt	Timestamp когда ответ разработчика был оставлен	2016-12-05 07:36:07

Извлечение из текстов векторов-признаков с помощью предобученных эмбеддингов NaVec

Оба набора данных были подвергнуты типичным трансформациям для обучения моделей в задачах обработке естественного языка.

Сначала колонка с комментариями(content) была приведена к нижнему регистру с помощью средств Python. Потом над ней же была применена токенизация с помощью библиотеки nltk. В данном случае разделение было произведено по словам (nltk.tokenize.word_tokenize). После были удалены стоп-слова, так же с помощью nltk. Список стоп-слов был взят из nltk.corpus.stopwords.words(«russian»). Далее была проведена лемматизация слов. Была использована библиотека rymorphy2.

Для извлечения признаков из текста были использованы предобученные эмбединги NaVec. Данные эмбединги имеют архитектуру GloVe, и имеют две реализации: пред обученные на 150 гигабайт художественной литературы и предобученные на 50 гигабайт постов в социальных сетях. Так как данное исследование ведется с использованием данных из Google Play был использован вариант, обученный на постах из социальных сетей.

Таблица 3

Описание трансформаций исходных наборов данных для обучения модели

Этап	Инструмент
Исходные данные	-
Приведение к нижнему регистру	Python
Токенизация	NLTK (word_tokenize)
Удаление стоп-слов	NLTK (stopwords)
Лемматизация	pymorphy2
Предобученные эмбединги	NaVec

Сравнение точности моделей классификаторов на основе CNN и GRU

Для сравнения, было выбрано 2 популярные в задача анализа текстов нейросетевые модели: на основе CNN и на основе GRU. Каждая модель была обучена в трёх вариантах: пред обучена на датасете твитов, обучена на целевом датасете, пред обучена на датасете твитов и до обучена на целевом датасете из Google Play. Результаты точности на тестовой выборке целевого набора данных и набора данных для пред обучения представлены в табл. 4.

Ключевым фактором для выбора модели и её варианта является максимальная точность на тестовой целевой выборке. Ожидаемо, модели совсем без дообучения показали очень плохой результат. Без пред обучения, только на целевой выборке – средний. А наилучший результат получился с предобучением и последующим дообучением.

Согласно таблице, наиболее хорошо показала себя модель на основе CNN с дообучением. Поэтому она была выбрана в качестве классификатора для анализа.

Сравнение различных вариантов моделей классификатора текстов по тональности

Название	Точность на пред обучении	Точность на целевой выборке		
		Без дообучения	Без предобучения	С дообучением
Нейросеть на основе CNN	~75%	60,11%	73,41%	80,03%
Нейросеть на основе GRU	~78%	57,22%	69,68%	76,32%

Подведение итогов классификации

Были собраны и классифицированы комментарии от 10 приложений из категории «Здоровье и Фитнес». Были выбраны приложения, для которых можно было выгрузить все комментарии целиком в разумный срок, т.е. у которых имеется не более нескольких десятков тысяч комментариев. Так как суммарное количество в данных приложениях было разное, решено было разделить визуализацию распределения классов на 2 диаграммы. Назовём первую условно «менее популярные» (рис. 1), а вторую «более популярные» (рис. 2). На каждой гистограмме приведено сокращенное название приложения по оси X, а по оси Y – количество комментариев. Красным цветом показана доля негативных комментариев, а синим – позитивных, полученных при применении классификатора.

Проанализируем «менее популярную» категорию. Как можно увидеть у данных приложений получилось так, что положительных комментариев больше. Если проанализировать страницы приложений и посмотреть на распределение по оценкам комментариев, то можно увидеть, что оценок меньше четырёх звёзд на этих приложениях в похожем соотношении к 4 и 5 звезд, как и негативных комментариев. То есть комментариев с низкой оценкой меньше, чем с комментариями 5 звёзд.

Аналогичная ситуация и с «более популярными приложениями». В них соотношение негативных комментариев к позитивным ещё меньше чем в «менее популярной» категории.

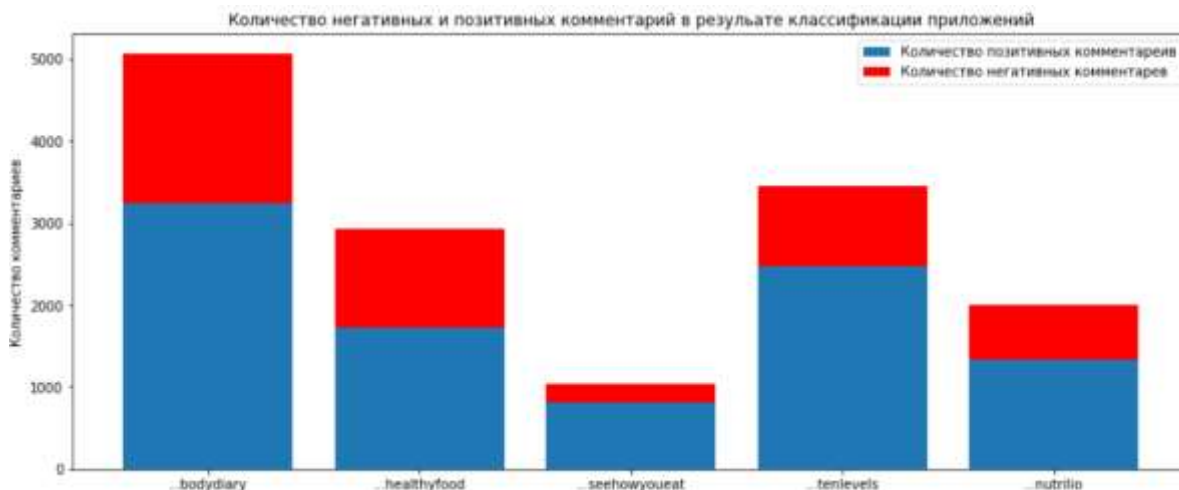


Рис. 1 – Результат работы классификатора тональности комментариев для «менее популярной» категории приложений

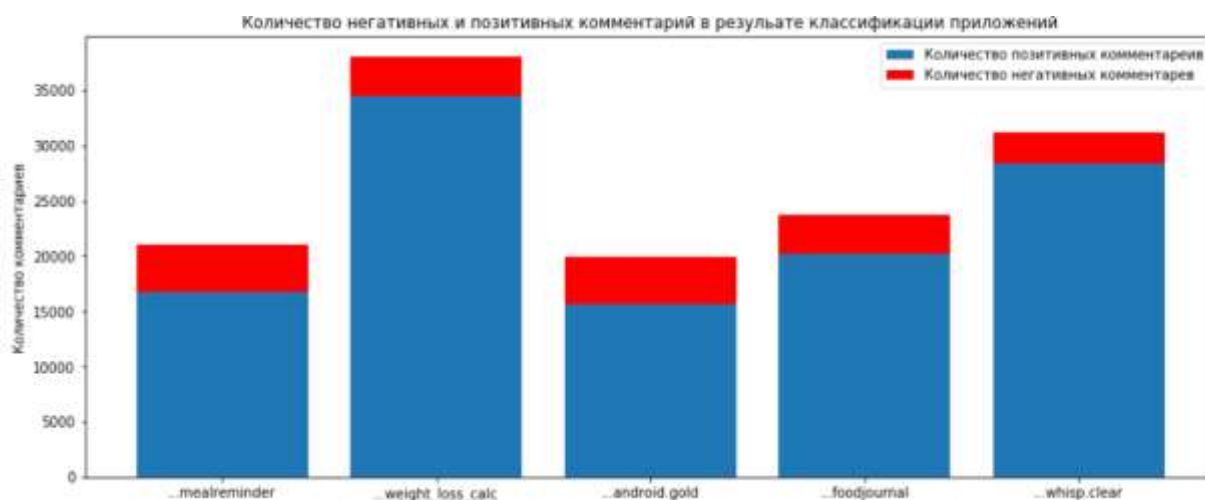


Рис. 2 – Результат работы классификатора тональности комментариев для «более популярной» категории приложений

Приведем также пример нескольких негативных комментариев, которые были правильно классифицированы моделью. Как можно заметить они действительно содержат негативный подтекст (скриншот на рис. 3).

```

1  какие-то глупые заморочки с пробным периодом – привяжите карту, чтобы начать
   пробный период... Бред... Я думала пробный потому и пробный – попробовал 7
   дней, если ок покупай, а так дурилово
2
3  Редко когда такое г... качнешь, одна реклама и никакой пользы.
4
5  Ни о чем. Просто запись веса и куча нафиг не нужной хрени.

```

Рис. 3 – Примеры верно классифицированных негативных отзывов

Однако нашлись и такие комментарии, которые явно были классифицированы неправильно. Это говорит о том, что модель можно развивать дальше, чтобы она давала большую точность, и учитывала больше особенностей комментариев.

Заключение

В ходе работы была разработана и реализована процесс обучения и применения для анализа модели определения тональности текстов для приложений Google Play. Была построена модель классификатора тональности для приложений из категории «Здоровье и Фитнес».

Полученная модель была применена для анализа тональности текстов, результаты. Применение модели показало, что с её помощью можно отсеивать эмоциональные отзывы, которые мало информативны, а зачастую и вовсе являются провокацией. Однако, не обошлось без исключений, и это говорит о том, что модель ещё можно улучшить.

Полученную модель не стоит применять для анализа приложений из других категорий, так как она скорее всего будет давать низкую точность. Однако, можно обучить новую по схожему принципу повторно, обучив заново, или дообучив существующую модель на датасете из отзывов другой категории приложений.

Библиографический список:

1. Как и почему анализ приложений конкурентов может помочь продвинуть ваше собственное приложение [Электронный ресурс] / Режим доступа: <https://habr.com/company/appodeal/blog/298426/>

2. Ушанова, М. В. Работа с негативными отзывами в сети интернет как форма продвижения компании / М. В. Ушанова // Гуманитарные технологии в современном мире : Материалы VI Международной научно-практической конференции, Калининград, 17–19 мая 2018 года. – Калининград: Западный филиал РАНХиГС, 2018. – С. 312-313.

3. Филимонов, А.В. Применение нейронных сетей для выявления троллей в социальных сетях / А.В. Филимонов, А.В. Осипов, А.Б. Климов // Нейрокомпьютеры: разработка, применение. – 2015. – № 8. – С. 87-92.

4. Исследование возможности снижения репутационных рисков банковских организаций на основе анализа тональности отзывов на русском языке / Воробьева А.А., Гафарова Я.К., Самоследов Г.А.

5. Рубцова Ю.В., Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы, 2015, – №1(109), – С.72- 78

6. Страница на Github проекта библиотеки google-play-scraper [Электронный ресурс] / – Режим доступа: <https://github.com/facundoalano/google-play-scraper>

References:

1. How and why competitor application analysis can Help promote your Own application [Electronic resource] / Access mode: <https://habr.com/company/appodeal/blog/298426/>

2. Ushanova, M.V. Work with negative reviews on the Internet as a form of company promotion / M.V. Ushanova // Humanitarian technologies in the modern world: Materials of the VI International Scientific and Practical Conference, Kaliningrad, May 17-19, 2018. – Kaliningrad: Western Branch of RANEPА, 2018. – pp. 312-313.

3. Filimonov, A.V. The use of neural networks to identify trolls in social networks / A.V. Filimonov, A.V. Osipov, A.V. Klimov // Neurocomputers: development, application. – 2015. – № 8. – pp. 87-92.

4. Research of the possibility of reducing reputational risks of banking organizations based on the analysis of the tone of reviews in Russian / Vorobyeva A.A., Gafarova Ya.K., Samosledov G.A.

5. Rubtsova Yu.V., Building a corpus of texts for tuning a tone classifier // Software products and systems, 2015, – №1(109), – Pp.72-78

6. Github page of the google-play-scraper library project [Electronic resource] /
– Access mode: <https://github.com/facundoolano/google-play-scraper>

Для цитирования: Царькова Н.И., Бова Р.Ю., Анализ тональности отзывов на приложения в Google Play категории «Здоровье и Фитнес» с помощью технологий больших данных/ Российский экономический интернет-журнал. – 2022. – № 4. URL:

© Царькова Н.И., Бова Р.Ю., Российский экономический интернет-журнал 2022, № 4.