Применение методов машинного обучения для анализа психотипов личности

О. А. Гусева

Южный федеральный университет, Ростов-на-Дону

Аннотация: В данной статье рассматривается решение задачи мониторинга контактной информации в социальной сети и определения по ней психотипа личности. В ходе работы был разработан достаточно эффективный метод сбора необходимых данных, так называемого парсинга. Собран собственный набор данных и проведены его обработка и подробный анализ. Результат исследования позволяет сделать первичное предположение о том, каким психотипом обладает тот или иной человек. Весь программный код выполнен на языке программирования Python.

Ключевые слова: кластеризация, психотип, социальная сеть, сангвиник, флегматик, холерик, меланхолик.

В современном мире благодаря быстрому распространению технологий и их внедрению в повседневную жизнь можно найти информацию почти о любом человеке, хоть раз воспользовавшемся социальными сетями. Даже при поверхностном поиске, социальные сети могут многое сказать о человеке, имеющем аккаунт и ведущем активную жизнь в Интернете. Именно поэтому довольно часто различные мессенджеры используются в научных и социологических исследованиях, в которых применяются специальные методы для подробного анализа разного рода информации. В данной работе рассматривается возможность использования машинного обучения для анализа психотипов на основе статусов. Вопросу анализа естественного языка посвящено достаточно много статей и исследований [1 – 3]. Большая часть предлагаемого разработчиками материала работает с англоязычными текстами для определения эмоциональной окраски, их классификации и, в меньшей степени, кластеризации. Целью данной работы является рассмотрение возможности использования машинного обучения для анализа психотипов на основе статусов на русском языке.

Предлагаемый в статье подход основан на использовании метода K-means для решения задачи кластеризации русскоязычных текстов и

последующем сравнении с результатами работы с теми же данными уже обученной модели.

В ходе анализа методов решения поставленной задачи, было решено использовать данные из социальной сети ВКонтакте. Такой выбор можно объяснить тем, что первой задачей в данном исследовании было собрать данные, то есть, достаточное количество статусов, подходящих для последующих обработок. Способы сбора информации из выбранного источника отличаются от других своим разнообразием и удобством в использовании. В нашей работе используется техника, применяющая методы API ВКонтакте - интерфейса, позволяющего получать информацию из базы данных vk.com с помощью http-запросов к специальному серверу. API упрощает создание кода, так как предоставляет набор готовых классов, функций или структур для работы [4].

Процесс получения таких данных имеет свои особенности. Было создано собственное Standalone-приложение. Посредством его получен API_ID, который открывает доступ ко многим возможностям метода API. Последним этапом было получение токена – некоторого ключа доступа, который передаётся на сервер вместе с запросом. В дальнейшем эти действия позволили программе просматривать страницы пользователей, считывать и сохранять нужную информацию.

Основной частью работы будет обработка текста и создание модели обучения нейронной сети. Как уже было упомянуто выше, в данной работе применимо обучение без учителя. Обучение без учителя или кластеризация является классом методов машинного обучения для поиска шаблонов в наборе данных [5]. Конкретного правильного результата, который должен быть получен после кластеризации, не существует, в отличие от обучения с учителем.

Однако перед тем, как приступить к описанию работы программного кода, необходимо рассмотреть вопрос, касающийся психотипов, их различия и особенностей. После изучения теоретических материалов было принято решение использовать разделение по Гиппократу – на четыре группы: сангвиники, холерики, флегматики и меланхолики [6]. Приведём краткое описание каждого из них. Сангвиник стремится к частой смене впечатлений, легко и быстро отзывается на окружающие события, общителен. Эмоции у сангвиника преимущественно положительные, они быстро возникают и быстро меняются [7]. В результате исследования этой информации было предположено, что ключевыми словами, по которым можно отличить статусы сангвиника, могут быть такие слова, как, например, радость, счастье, улыбка и их производные. Также особенностями статусов людей сангвиников могут быть восклицательные знаки и большое количество красочных смайликов. Для темперамента флегматиков характерна сдержанность, медлительность психических реакций и постоянство. Такие люди надежны, ответственны, преданны. К флегматикам можно отнести выносливых, старательных и работящих людей. Они разговаривают тихо, но четко, с заметными паузами, без эмоционально-интонационных скачков. Это искренние, иногда даже чересчур прямолинейные люди [8]. При анализе статусов можно не ждать большого количества смайликов, скорее всего их не будет совсем, так же, как и восклицательных знаков. При поиске ключевых слов можно основываться на том, что флегматики замкнуты, плохо адаптируются к переменам. Их преимуществами являются постоянство, ум и дисциплина, они хорошо работают с большим количеством информации. По такому же принципу был проведён анализ психотипов холериков, отличительной чертой которых является взрывной, импульсивный характер [9] и меланхоликов, которые склонны к тревожности и печали [10].

Заметим, что к числу основных процессов данного исследования относится обработка данных. Хорошо очищенный и подготовленный текст может обеспечить более корректную работу модели. Основываясь на различных, уже предложенных ранее методах, был создан собственный алгоритм, позволяющий анализировать статусы по отдельности, не создавая слитный текст в ходе лемматизации [11-13]. Затем к отформатированным статусам, пример одного из таковых приведён в Приложении (см. рис. 1), был применён метод кластеризации K-means [14].

242149093	Возьми меня к себе. С тобой спокойно. Я обещаю больше не реветь, Спать по ночам, вести себя достойно И лишних всех из памяти стереть.	['взять', '.', 'спокойно', '.', '\n', 'обещать', 'реветь', '\n', 'спать', 'ночь', 'вести', 'достойно', '\n', 'лишний', 'память', 'стереть', '.']

Рис. 1. – Пример отформатированного статуса

Результаты его работы мы наблюдаем на графике (см. рис. 2).

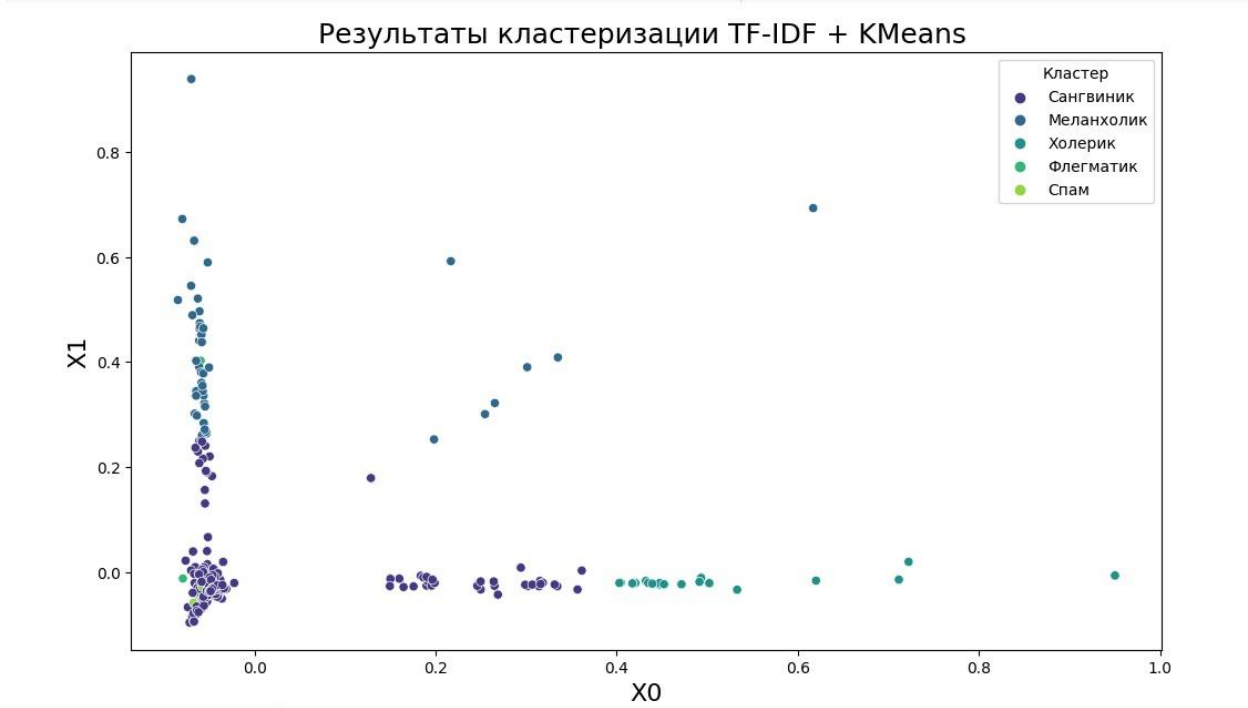

Рис. 2. – Результаты кластеризации

Можно увидеть, что, ориентируясь на статусы, большую часть пользователей социальной сети модель отнесла к сангвиникам. Более подробное

количественное описание результатов приведено в таблице (см. рис. 3).

	Без учителя	С учителем
Невозможно определить	12	382
Сангвиник	1174	299
Флегматик	13	193
Меланхолик	68	420
Холерик	65	38

Рис. 3. – Результаты работы модели

Позднее, с этими же данными, работает модель обучения с учителем «Dostoevsky», она позволяет достаточно точно определить тональность текста [15]. Все результаты для сравнения записываются в сводную таблицу.

Главным результатом проведённого исследования следует считать алгоритмизацию модели, которая была реализована в виде программного приложения. Как видно из таблицы и графика (см. рис. 3, 4), предложенный алгоритм может распознавать характерные шаблоны для четырёх различных психотипов, а также для рекламы или спама.

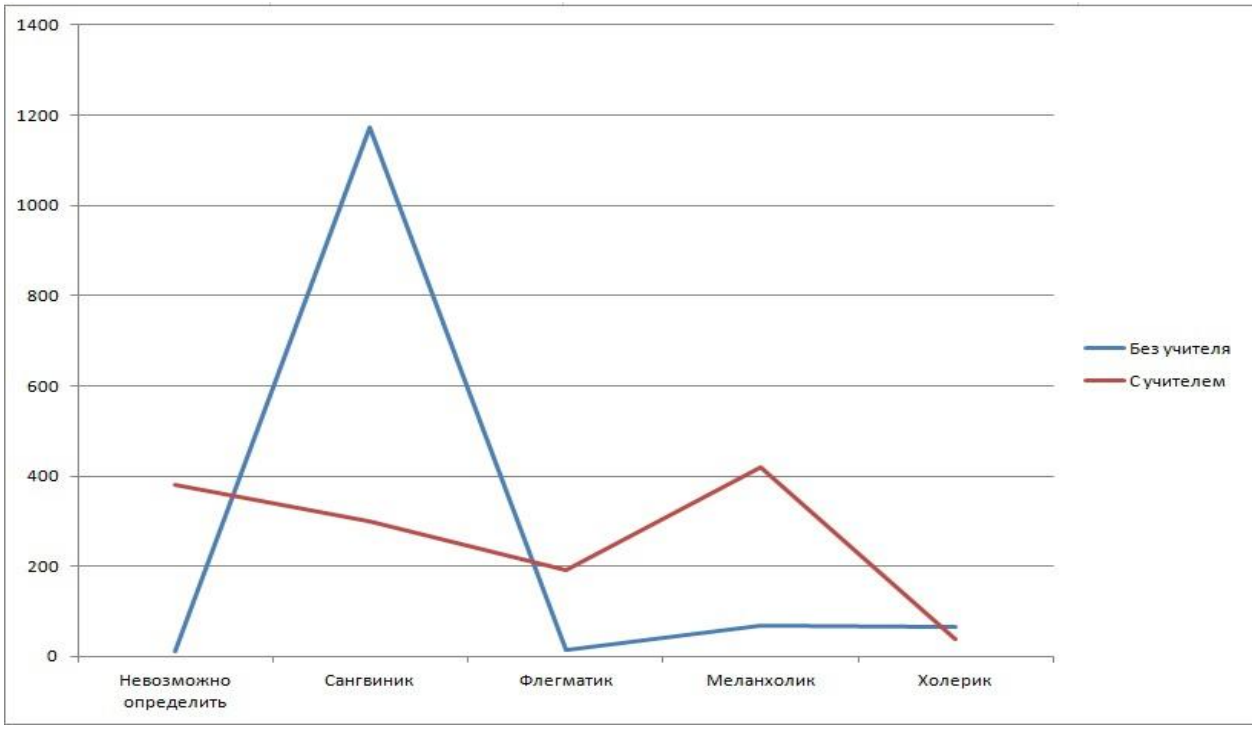

Рис. 4. – Результаты обучения моделей

Далее на основе этих шаблонов определяется, каким психотипом обладает тот или иной человек или же делается вывод о том, что текущий

статус несёт рекламный характер и не позволяет определить психотип. Точность результатов выполнения задачи определить достаточно сложно, так как уже было упомянуто, что в обучении без учителя нет заранее определённого правильного ответа, а определение психотипа является сложной задачей и с точки зрения психологии. Можно только сравнивать с результатами работы обученной модели, которая в силу сложности и глубины тематики психологии человека также может давать погрешность в оценке.

Литература

1. «Люблю» и «ненавижу»: анализ эмоциональной окраски текста с помощью Python // Proglib URL: proglib.io/p/lyublyu-i-nenavizhu-analiz-emocionalnoy-okraski-teksta-s-pomoshchyu-python-2020-11-13.
2. Clustering with maximum diameter // Github URL: github.com/antklen/diameter-clustering/blob/master/README.md.
3. Практический NLP с Python-библиотекой spaCy для SEO-задач в Google Colab // BigDataSchool URL: bigdataschool.ru/blog/spacy-library-for-nlp-in-google-colab-example.html (дата обращения: 16.01.2023).
4. Использование API // VK для разработчиков. URL: dev.vk.com/api/getting-started (дата обращения: 15.12.2022).
5. Clustering // Scikit-learn URL: scikit-learn.ru/clustering/.
6. Психотипы Людей – Классификация И Характеристика // Psylib URL: psylib.org/psikhotipy-lyudei-klassifikasija-i-prinsipy-opredelenija/.
7. Кто такой сангвиник: все об оптимистической душе компании // 4T URL: temperamenttest.org/ru-ru/sangvinik/.
8. Кто такой флегматик: ленивый ворчун или эмоционально уравновешенный логик // 4T URL: temperamenttest.org/ru-ru/flegmatik/.
9. Холерик - его сильные и слабые стороны, черты характера и поведения // Psylogik.ru URL: psylogik.ru/192-holerik.html.

10. Меланхолик как один из видов темперамента // Psylogik.ru URL: psylogik.ru/193-melanholik.html (дата обращения: 22.12.2022).

11. О сборе данных. Как собирать данные, анализировать их и грабить корованы // Хабр URL: habr.com/ru/articles/407977/.

12. Моем датасет: руководство по очистке данных в Python // Proglib URL: proglib.io/p/moem-dataset-rukovodstvo-po-ochistke-dannyh-v-python-2020-03-27.

13. Очистка данных с помощью Python и Pandas: обнаружение пропущенных значений MachineLearningMastery.ru URL: www.machinelearningmastery.ru/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b/.

14. Объясните так, как будто мне 10 лет: простое описание популярного алгоритма кластеризации k-средних // Proglib URL: proglib.io/p/obyasnite-tak-kak-budto-mne-10-let-prostoe-opisanie-populyarnogo-algoritma-klasterizacii-k-srednih-2022-12-07.

15. Dostoevsky — анализ тональности в Python за 5 минут // Егоров Егор Блог о разработке на Python URL: egorovegor.ru/analiz-tonalnosti-s-python-i-dostoevsky/.

References

1. «Lyublyu» i «nenavizhu»: analiz emotsional'noy okraski teksta s pomoshch'yu Python [Love and hate: Python analysis of text emotions]. URL: proglib.io/p/lyublyu-i-nenavizhu-analiz-emocionalnoy-okraski-teksta-s-pomoshchyu-python-2020-11-13.

2. Clustering with maximum diameter. URL: github.com/antklen/diameter-clustering/blob/master/README.md.

3. Prakticheskiy NLP s Python-bibliotekoy spaCy dlya SEO-zadach v Google Colab [NLP with Python library for SEO-tasks in a Google Colab.] BigDataSchool URL: bigdataschool.ru/blog/spacy-library-for-nlp-in-google-colab-example.html (accessed: 16.01.2023).

4. Ispol'zovaniye API VK dlya razrabotchikov [API VK for developers]. URL: dev.vk.com/api/getting-started (accessed: 15.12.2022).

5. Clustering Scikit-learn URL: scikit-learn.ru/clustering/.

6. Psikhotipy Lyudey Klassifikatsiya I Kharakteristika [Psychotypes. Classification and characteristics]. Psylib URL: psylib.org/psikhotipy-lyudei-klassifikasija-i-prinsipy-opredelenija/.

7. Kto takoy sangvinik: vse ob optimisticheskoy dushe kompanii [A sanguine personality: all about the optimistic soul of the company]. 4T URL: temperamenttest.org/ru-ru/sangvinik/.

8. Kto takoy flegmatik: lenivyy vorchun ili emotsional'no uravnoveshennyy logik [A phlegmatic personality: lazy grumbler or emotionally balanced logician]. 4T. URL: temperamenttest.org/ru-ru/flegmatik/.

9. Kholerik - yego sil'n·yye i slab·yye storony, cherty kharaktera i povedeniya [Choleric his strengths and weaknesses, character traits and behavior]. Psylogik.ru URL: psylogik.ru/192-holerik.html.

10. Melankholik kak odin iz vidov temperamenta [Melancholic - one of temperaments]. URL: psylogik.ru/193-melanholik.html (accessed: 22.12.2022).

11. O sbore dannykh. Kak sobirat' dannyye, analizirovat' ikh i grabit' korovany [About data collection. How to collect, analyze data and rob caravans]. URL: habr.com/ru/articles/407977/ (accessed: 15.12.2022).

12. Moyem dataset: rukovodstvo po ochistke dannykh v Python [Cleaning dataset: Python data cleanup guide]. URL: proglib.io/p/moem-dataset-rukovodstvo-po-ochistke-dannyh-v-python-2020-03-27.

13. Ochistka dannykh s pomoshch'yu Python i Pandas: obnaruzheniye propushchennykh znacheniy [Data cleanup with Python and Pandas: detection of missing data]. URL: machinelearningmastery.ru/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b/.

14. Ob"yasnite tak, kak budto mne 10 let: prostoye opisaniye populyarnogo algoritma klasterizatsii k-srednikh [A simple description of the popular clustering algorithm k-means]. URL: proglib.io/p/obyasnite-tak-kakbudto-mne-10-let-prostoe-opisanie-populyarnogo-algoritma-klasterizacii-k-srednih-2022-12-07.

15. Dostoevsky analiz tonal'nosti v Python za 5 minut [Dostoevsky's analysis of tonality in Python in 5 minutes] Yegorov Yegor Blog o razrabotke na Python URL: egorovegor.ru/analiz-tonalnosti-s-python-i-dostoevsky/.