

УДК 004.912, 004.89

DOI: 10.46548/21vek-2022-1157-0002

## АНАЛИЗ ПОЛЯРНОСТИ НАСТРОЕНИЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ В ПЕРИОД COVID-19

© 2022

**Зоткина Алена Александровна**, аспирант кафедры «Программирование»  
**Мартышкин Алексей Иванович**, кандидат технических наук, доцент,  
заведующий кафедрой «Программирование»  
*Пензенский государственный технологический университет*  
(440039, Россия, Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11,  
e-mails: alena.zotkina.97@mail.ru, alexey314@yandex.ru)

**Аннотация.** Обоснована методика анализа полярности настроений пользователей социальных сетей в период пандемии COVID-19 с использованием в качестве платформы для создания контента социальной сети ВКонтакте. Рассмотрены этапы анализа: сбор данных, полученных при помощи модуля для создания скриптов VK\_API, предварительная обработка данных через конвейер обработки естественного языка (токенизация, нормализация, лемматизация), создание модели и ее оценка. Для решения задачи используется высокоуровневый язык программирования Python с динамической строгой типизацией и автоматическим управлением памятью, в синтаксисе которого содержатся библиотеки NumPy, Theano, Lasagne. Представлена модель машины опорных векторов (SVM) для классификации данных, для этого используются три класса тональности текста: отрицательный, положительный и негативный. Для анализа использованы модули обработки естественного языка (NLTK) и TextBlob с различными типами ядер, а для классификаций настроений с повышенной точностью – полиномиальные и радиальные базисные функции. Показан результат тестирования классификаторов ядер, приведен пример обучения оптимизатора и его реализации. В заключении сформулированы основные выводы по проделанной работе.

**Ключевые слова:** социальные сети, вакцинация, коронавирус, ВКонтакте, SVM, TextBlob, NLTK.

## ANALYSIS OF THE SOCIAL NETWORKS USERS' MOOD POLARITY DURING THE COVID-19 PERIOD

© 2022

**Zotkina Alena Aleksandrovna**, postgraduate of sub-department «Programming»  
**Martyshev Alexey Ivanovich**, candidate of technical sciences, docent, head of sub-department «Programming»  
*Penza state technological University*  
(440039, Russia, Penza, BaydukovProyezd / Gagarin Street, 1a/11,  
e-mails: alena.zotkina.97@mail.ru, alexey314@yandex.ru)

**Abstract.** The article includes an analysis of the sentiments of social media users during the COVID-19 period. It was noted that the public network VKontakte will be used as a social platform for the collection. The structure of the analysis of the polar moods of users of the information space is given. Analysis steps included: data collection, getting help from the module to create VK\_API scripts, data preprocessing through natural language processing (tokenization, normalization, lemmatization), model creation and evaluation. It is noted that the implementation of the noted task uses the high-level Python programming language with dynamic strong typing and automatic memory management, the syntax of which contains the libraries NumPy, Theano, Lasagne. Shows how to create a support vector machine (SVM) model for data classification. It is noted that three classes of text sentiment are used for data classification: negative, positive and negative. The use of natural language processing modules (NLTK) and TextBlob is justified. The article considers various types of nuclei. It is noted that for the analysis of mood classifications, polynomial and radial basis functions will be used in order to identify the highest accuracy. The result of testing kernel classifiers is shown. An example of training the optimizer and its implementation is given. In conclusion, the main conclusions on the work done are formulated.

**Keywords:** social networks, vaccination, coronavirus, VKontakte, SVM, TextBlob, NLTK.

**Введение.** Начиная с 2020 г. по всему миру распространяется коронавирусная инфекция COVID-19, которая затронула все сферы человеческой жизни. В ежедневный обиход вошли слова «карантин», «локдаун», «чрезвычайное положение», «эпидемия», «пандемия», «коллективный иммунитет» и другие понятия. COVID-19 повлиял на глобальную экономику, мировую политику, международные отношения и изменил ценности и установки обществ, привычки и повседневные реалии. Острой проблемой для государ-

ства на данный момент является вопрос вакцинации населения, ставшей проблемой не только врачей, но и всего общества. Выявление полярности настроений поможет врачам и исследователям уяснить причины нерешительности отдельных лиц в отношении вакцинации и поправить ситуацию.

В наше время масштабы использования интернета в качестве социальной платформы для создания контента выросли с появлением микроблогов, таких как VKontakte, Instagram и др., в них люди обмениваются

между собой информацией, публикуют свои мнения. Информация пользователя часто используется маркетологами для сбора и систематизации отзывов об их продукте в сообществе, в последнее время возрос интерес к такому анализу в связи с пандемией. Однако при стремительно возрастающем количестве блогов и постов уследить за отслеживаемой информацией затруднительно. Таким образом, реализация модели анализа настроений становится очень актуальной. Идея анализа настроений пользователей базируется на машинном обучении при обработке полученных

данных.

**Целью** данной работы является разработка модели для классификации полярности настроений с использованием метода опорных векторов (*SVM – Support Vector Machine*).

**Материалы и результаты исследования.** Для определения круга лиц, выступающих за вакцинацию или против нее, важно классифицировать мнения, высказанные пользователями социальной сети *Vkontakte*. На рисунке 1 показана структура анализа полярности настроений пользователей социальной сети.

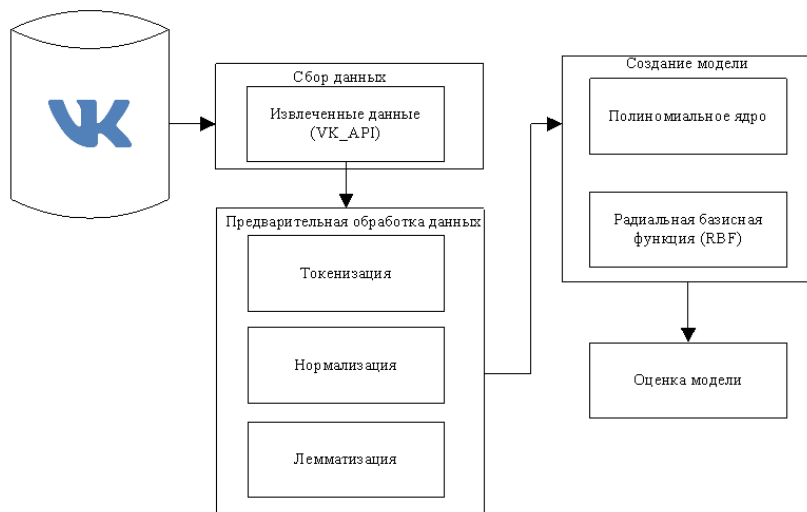


Рисунок 1 – Структура анализа полярности настроений на основе социальной сети VKontakte

Первый этап – сбор данных из социальной сети *Vkontakte*. Социальная сеть снабжена средствами для взаимодействия и извлечения информации при помощи *VK\_API* [1]. Чтобы обратиться к методу *API VKontakte*, необходимо выполнить *POST* или *GET* запрос следующего вида: *requests.get('https://api.vk.com/method/wall.get')*. Код для парсинга данных из социальной сети пишется на *Python* [2] – языке программирования, считающимся высокоуровневым. Этот язык поддерживает динамическую строгую типизацию, т.е. переменная начинает работать с типом в момент ее присваивания, что обозначает, что одна и та же переменная может принимать различные типы данных. Еще одним преимуществом использования *Python* является свойство автоматического управления памятью [3]. *Python* располагает множеством ресурсов, которые облегчают организацию машинного обучения, что выгодно отличает его от любого другого языка. Использование специальных инструментов, таких, как пакеты *pandas* (библиотека с открытым исходным кодом), предоставляющая высокопроизводительные, простые в использовании структуры данных и инструменты анализа для языка программирования *Python*, и *numpy*, позволяет достичь высокой производительности в обработке данных [4, 5].

В синтаксисе *Python* содержатся следующие библиотеки:

1) *NumPy* – поддерживает большие многомерные массивы и матрицы вместе с большой библиотекой высокоуровневых (и очень быстрых) математических

функций для операций с этими массивами. Предоставляет возможность использования генератора случайных чисел [6];

2) *Theano* – используется для быстрых численных вычислений, может быть запущена на *CPU* или *GPU*. Это ключевая базовая библиотека для глубокого обучения в *Python*. Основными качествами, послужившими в пользу ее выбора, стали интеграция с *numpy* и использование при разработке библиотеки символьного подхода, в отличие от императивного в других пакетах. Записывая вычисления в символьной парадигме, можно задать граф вычислений, который в дальнейшем будет скомпилирован и исполнен. Благодаря тому, что на этапе компиляции происходит ряд оптимизаций в коде, вычисления становятся более эффективными в части потребной памяти и скорости исполнения [7, 8];

3) *Lasagne* – используется для создания и обучения нейронных сетей в *Theano*. Она поддерживает сети передачи данных, такие как сверточные нейронные сети (*CNN*), сверточные сети, включая *Long Short-Term Memory (LSTM)* и любую их комбинацию [9]. Важно отметить, что модель *LSTM (Long Short-Term Memory)*, доминирует в большинстве задач НЛП в последние несколько лет, достигая самых высоких результатов. Подход *LSTM* считывает текст последовательно и сохраняет информацию, относящуюся к текущей задаче. В *LSTM* имеются ячейки, которые контролируют, какая информация запоминается, а что забывается. В случае анализа социальных сетей очень важно отличие и различие между «отлично» и «не очень».

*LSTM*, обученный предсказанию настроений, уяснит, что это важно, и научиться понимать, какие слова следует отрицать. В отношении чтения больших объемов текста *LSTM* можно рассматривать как «изучение» грамматических правил [10, 11].

Несколько ключевых слов, таких как вакцина, прививка, эпидемия, коронавирус используются для извлечения данных. Затем данные обрабатываются в электронных таблицах с использованием вышеуказанных библиотек. Следующим этапом является предварительная обработка данных через конвейер обработки естественного языка. Необходимые шаги включают следующие действия: токенизация, нормализация и лемматизация [12, 13].

**Токенизация.** Язык в его исходной форме не может быть точно воспринят машиной, поэтому необходимо обработать язык, чтобы машине было легче понять. Первая часть понимания данных – процесс, называемый токенизацией, или разделением строк на более мелкие части, называемые токенами. Маркер – последовательность символов в тексте, которая служит единым целым. Основной способ разбить язык на токены – разделить текст на основе пробелов и знаков препинания [14].

**Нормализация.** При данном процессе удаляются стоп-слова, знаки препинания, прописные буквы. Примерами стоп-слов являются: «если», «но», «а», «значит» и т. д. [15].

**Лемматизация.** Алгоритм лемматизации анализирует структуру слова и его контекст, чтобы преобразовать его в нормализованную форму. Очевидно, что это происходит в ущерб скорости. В отличие от стемминга лемматизация сохраняет часть речи слова без разделения суффиксов [16].

После этапа предварительной обработки данных создается модель машины опорных векторов для классификации данных. Одним из главных преимуществ использования *SVM* является то, что он уменьшает риск ошибки обобщения классификатора. Для классификации данных используют три класса: положительный, отрицательный и нейтральный. В этом процессе участвуют модули *Natural Language Tool Kit (NLTK)* и *Text Blob* [19, 20].

Для реализации метода классификации *SVM* была изображена гиперплоскость данных, разделение на классы проводилось при помощи математической функции ядра. В анализе классификации настроений можно использовать различные типы ядер, такие как линейные, сигмоидальные, радиальные базисные функции (*RBF*), также известные как гауссовы, нелинейные и полиномиальные, которые являются наиболее популярными [20]. В данном исследовании были использованы полиномиальные и радиальные базисные функции, с целью выяснения, какое из ядер может достичь наивысшей точности. На заключительном этапе модель оценивается по ее производительности.

**Результаты тестирования предложенного решения.** *Textblob* позволяет разделить комментарии пользователей по полярности на три категории: положи-

тельные, отрицательные и нейтральные. На рисунке 2 показано процентное соотношение комментариев. На следующем этапе оценивалась выбранная модель. В ходе тестирования классификаторов обоих типов ядер установлено, что использование ядра с радиальной базисной функцией дает более высокую точность классификации данных, чем полиномиальное ядро.



Рисунок 2 – Процентное соотношение комментариев в зависимости от категории

**Заключение.** В настоящее время многие виды исследований сосредоточены на анализе тональности сообщений и комментариев о вакцинации пользователей социальных сетей. Существование социальных сетей упростило получение мнений пользователей. Выполненное авторами исследование помогает создавать программное решение для анализа тональности сообщений, накопленных с помощью модуля *Textblob* и классификатора машины опорных векторов. Однако следует учитывать то, что из-за слабости предварительной фильтрации данных точность может оказаться ниже ожидаемой.

#### СПИСОК ЛИТЕРАТУРЫ:

1. Знакомство с API ВКонтакте [Электронный ресурс]. – URL: [https://vk.com/dev/first\\_guide/](https://vk.com/dev/first_guide/) (дата обращения: 21.01.2022).
2. Маккинни, У. Python и анализ данных / У. Маккинни; перевод с английского А. А. Слинкина. – 2-ое изд., испр. и доп. – Москва: ДМК Пресс, 2020. – 540 с.
3. Д. Грас. Data Science. Наука о данных с нуля / перевод с английского А.А. Логунова. – Санкт-Петербург: БХВ-Петербург, 2020. – 411 с.
4. Библиотека Pandas [Электронный ресурс]. – URL: <https://blog.skillfactory.ru/glossary/pandas/> (дата обращения: 21.01.2022).
5. Учебник по библиотеке NumPy: учитесь на примерах [Электронный ресурс]. – URL: <https://pythonist.ru/uchebnik-po-biblioteke-numpy-uchites-na-primeraх/> (дата обращения: 21.01.2022).
6. Учебник по NumPy - Визуализация примеров для быстрого изучения [Электронный ресурс]. – URL: <https://pythonscripts.com/numpy> (дата обращения: 21.01.2022).
7. GitHub - PacktPublishing/Deep-Learning-with-Theano: Deep Learning with Theano, published by Packt [Электронный ресурс]. – URL: <https://github.com/PacktPublishing/DeepLearning-with-Theano> (дата обращения: 21.01.2022).
8. Библиотеки для глубокого обучения Theano/Lasagne / Хабр [Электронный ресурс]. – URL: <https://habr.com/ru/company/ods/blog/323272/> (дата обращения: 20.01.2022).

9. Welcome to Lasagne – Lasagne 0.2. dev1 documentation [Электронный ресурс]. – URL: <https://lasagne.readthedocs.io/en/latest/> (дата обращения: 19.01.2022).
10. Conrad Tiflin. LSTM Recurrent Neural Networks for Signature Verification, 2012. 104.
11. LSTM – нейронная сеть с долгой краткосрочной памятью [Электронный ресурс]. – URL: <https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/> (дата обращения: 20.01.2022).
12. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. arXiv Report 1604.00289, 2016.
13. Курс по теоретическому глубокому машинному обучению deep learning в nlp. Лекции 1–5. [Электронный ресурс]. – URL: <https://github.com/deepmip/tdl> (дата обращения: 20.01.2022).
14. Токенизация в Python с использованием NLTK [Электронный ресурс]. – URL: <https://pythobyte.com/tokenization-in-python-using-nltk-96642092/> (дата обращения: 20.01.2022).
15. Нормализация данных в Python [Электронный ресурс]. – URL: <https://pythonist.ru/normalizacziya-dannyh-v-python/> (дата обращения: 20.01.2022).
16. Подходы лемматизации с примерами на Python [Электронный ресурс]. – URL: <https://webdevblog.ru/podhody-lemmatizacii-s-primerami-v-python/> (дата обращения: 20.01.2022).
17. SVM. Объяснение с нуля и реализация на python. Подробный разбор метода опорных векторов Python [Электронный ресурс]. – URL: <https://habr.com/ru/company/ods/blog/484148/> (дата обращения: 20.01.2022).
18. Pang B. & Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - pp.1-135.
19. Решение задач NLP с использованием TextBlob [Электронный ресурс]. – <https://egorovegor.ru/textblob-python-nlp/> (дата обращения: 20.01.2022).
20. Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. Userlevel sentiment analysis incorporating social networks // Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). 2011.

*Статья поступила в редакцию 26.01.2022*

*Статья принята к публикации 10.03.2022*