

УДК 004.852

doi: 10.17586/2226-1494-2020-20-1-118-124

ПРОГНОЗИРОВАНИЕ РЕАКЦИИ ПОЛЬЗОВАТЕЛЕЙ В СОЦИАЛЬНЫХ СЕТЯХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Е.П. Попова, В.Н. Леоненко

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: Popova.elizaveta@list.ru

Информация о статье

Поступила в редакцию 18.10.19, принята к печати 16.12.19
Язык статьи — русский

Ссылка для цитирования: Попова Е.П., Леоненко В.Н. Прогнозирование реакции пользователей в социальных сетях методами машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 1. № 1. С. 118–124. doi: 10.17586/2226-1494-2020-20-1-118-124

Аннотация

Предмет исследования. Выполнен сравнительный анализ методов прогнозирования реакции пользователей на сообщение, опубликованное в социальных сетях, средствами машинного обучения. Предпочтение отдано методу, обеспечивающему наибольшую точность. **Метод.** Применены методы машинного обучения: метод опорных векторов, метод градиентного бустинга, случайный лес и многослойный перцептрон. В качестве референтной методики предсказания использован статистический метод регрессионного анализа — линейная регрессия. Векторизация записей для получения количественных характеристик их содержимого проведена методами «Bag of Words», TF-IDF, Word2Vec. Качество прогноза оценивалось с помощью коэффициента детерминации R². **Основные результаты.** Численный эксперимент выполнен с использованием набора данных, собранных в социальной сети «ВКонтакте». Набор содержал информацию о подписчиках сообщества, публикациях, отметках «мне нравится», «рассказать друзьям» и комментариях к публикациям. Прогнозировалось количество отметок и число комментариев под размещенной публикацией в зависимости от ее содержания. Наиболее точные результаты получены при прогнозировании числа комментариев. Качество прогнозов количества отметок «мне нравится» и «рассказать друзьям» оказалось более низким. **Практическая значимость.** Результаты работы могут найти применение при анализе влияния новостей разного характера, в том числе «фейковых» новостей, на пользователей социальных сетей. Развитие методов предсказания позволит проводить планирование мер для ускорения или сдерживания распространения сообщений.

Ключевые слова

социальные сети, прогнозирование реакции, обработка естественного языка, машинное обучение, регрессионный анализ

Благодарности

Исследование выполнено при поддержке Российского научного фонда (соглашение № 19-11-00326).

doi: 10.17586/2226-1494-2020-20-1-118-124

MACHINE LEARNING METHODS FOR FORECASTING OF SOCIAL NETWORK USERS' REACTION

E.P. Popova, V.N. Leonenko

ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: Popova.elizaveta@list.ru

Article info

Received 18.10.19, accepted 16.12.19
Article in Russian

For citation: Popova E.P., Leonenko V.N. Machine learning methods for forecasting of social network users' reaction. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 1, pp. 118–124 (in Russian). doi: 10.17586/2226-1494-2020-20-1-118-124

Abstract

Subject of Research. The paper presents comparative analysis of forecasting methods for users' response to a message emerged on social media and application of machine learning. The method that provides the highest accuracy is preferred. **Method.** The following machine learning methods were applied: support vector method, gradient boosting method,

random forest and multilayer perceptron. The statistical method of regression analysis — linear regression — was used as a reference forecasting method. Vectorization of records for obtaining quantitative characteristics of their content was carried out using such methods as: “Bag of Words”, TF-IDF and Word2Vec. The forecast quality was evaluated by R^2 determination coefficient. **Main Results.** A numerical experiment was performed using a data set collected on the VKontakte social network. The set contained information about community subscribers, publications, “I like” and “tell friends” marks and comments on publications. The number of marks and comments under the posted publication was projected, depending on its content. The most accurate results were obtained when predicting the number of comments. The quality of forecasts for the number of “I like” and “tell friends” marks turned out to be lower. **Practical Relevance.** The results of the work can be used in analyzing the effect of various news, including fake news, on users of social networks. The development of forecasting methods provides the planning of measures for acceleration or containment of the messaging distribution.

Keywords

social networks, reaction forecasting, natural language processing, machine learning, regression analysis

Acknowledgments

The study was supported by the Russian Science Foundation (agreement No. 19-11-00326).

Введение

Социальные сети объединяют сотни миллионов интернет-пользователей по всему миру и являются значащим каналом получения информации. В связи с этим информация о том, как будет распространяться та или иная новость, имеет важное практическое значение. Моделирование указанного процесса имеет широкий спектр применения: прогнозирование реакции пользователей, максимизация влияния, контроль слухов, оценка общественного мнения и т. д. Таким образом, прикладное значение в разработке модели распространения новостей может лежать сразу во многих сферах, таких как, например, создание эффективных маркетинговых стратегий, развитие новостных источников, выявление трендов интересов или же обнаружение «фейковых новостей» [1].

Рост социальных сетей породил множество моделей, прогнозирующих динамику процессов в киберпространстве. При этом возможно использование механистических и статистических моделей. Механистические модели являются объясняющими, т. е. опираются на механику изучаемого процесса, в то время как статистические модели при построении прогнозов основываются исключительно на исторических данных.

Одним из примеров механистических (объясняющих) моделей служат модели на основе комплексных сетей или графов. Вершины графов представляют агентов, а ребра — связи между ними [2]. Для реалистичного моделирования процессов с помощью таких моделей необходимо изучение структуры рассматриваемых социальных сетей, а также определении ключевых особенностей агентов и их взаимодействия. Многие параметры сети являются качественными, а не количественными, что затрудняет их формализацию и дальнейшее использование. Так в ряде работ исследуется проблема взаимосвязи между личными свойствами агента и характеристиками информационного процесса [3, 4].

Поскольку процесс распространения информации схож с процессом эпидемии, можно представить некую новость как вирус, который со временем заражает все большее количество людей благодаря их коммуникации друг с другом. В связи с этим для моделирования распространения информации можно использовать эпидемические модели, такие как SIR-модель (Susceptible-Infected-Recovered) на основе системы дифференци-

альных уравнений. В настоящее время данная модель пользуется все большей популярностью [5] для описания различных процессов, таких как, например, аплодисменты аудитории [6], распространение идей, распространение компьютерных вирусов, банковских операций [7] и многих других.

Для описания процесса распространения информации может применяться модель клеточного автомата [8]. Теория игр также активно используется для исследования социальных сетей [2].

В работах [9, 10] описана модель распространения тематических новостей Ландэ и Додонова. Основное отличие от предыдущих рассмотренных подходов состоит в том, что в качестве агентов выступают сообщения, а не пользователи сети.

Приведенные выше исследования посвящены либо моделированию динамики процесса распространения информации, либо поиску факторов, влияющих на интенсивность ее распространения. Описанные выше методы отличаются сложностью реализации, а также вычислительно трудоемки.

В настоящем исследовании авторов интересует в большей степени не построение механики процесса распространения реакций, а получение самих итоговых количественных показателей, т. е. количество отметок — «рассказать друзьям», «мне нравится» и комментариев за некоторый промежуток времени. В силу этого вместо построения моделей, объясняющих процесс взаимодействия пользователей, можно использовать ретроспективные агрегированные данные для предсказания, базируясь на предыдущих наблюдениях. Такая задача может быть решена статистическими методами и методами машинного обучения¹.

Примерами конкретных используемых методов являются следующие:

1) линейная регрессия (описание метода представлено в работе [12], модель прогнозирования личности на основе микроблога Sina с использованием данного метода приведена в [13]);

2) метод опорных векторов (SVM, support vector machine) (подробнее о данном методе можно прочитать

¹ Машинное обучение (machine learning) — класс методов искусственного интеллекта, характерной чертой которых является обучение в процессе применения решений множества сходных задач [11].

в [14], способ прогнозирования индекса популярности истории, опубликованной в социальной сети Digg.com, показан в [15]);

3) градиентный бустинг (метод разобран в [16]);

4) случайный лес (Random Forest) (описание представлено в [17], возможность использования данного метода в контексте прогнозирования передачи информации в социальных сетях описана в [18]);

5) сверточные нейронные сети (описание представлено в [19], метод, основанный на сверточных нейронных сетях для прогнозирования «твитов» (рассматривается социальная сеть Twitter), состоящих из текста и изображения, предложен в [20]). Среди методов нейронных сетей был выбран метод многослойного перцептрона (MLP, multilayer perceptron) [21], так как он имеет наиболее подходящую архитектуру для решения рассматриваемой задачи.

В работе приведены результаты применения описанных выше методов машинного обучения для различных методов векторизации поста (получения количественных характеристик его содержимого). Особенности их применения изложены в разделе «Применение методов».

Описание исходных данных

Рассмотрено распространение информации из сообщества социальной сети «ВКонтакте», предназначенного для сбора средств нуждающимся. В группе размещаются публикации с призывом о помощи конкретному человеку, содержащие, как правило, следующие данные:

- фамилия и имя;
- страна/город;
- диагноз человека, которому требуется помощь;
- что необходимо сделать;
- требуемая сумма;
- реквизиты для перевода средств.

На каждую из записей подписчик сообщества может отреагировать одним из следующих способов:

- проигнорировать запись;
- выразить одобрение нажатием кнопки «мне нравится»;
- оставить комментарий под записью;
- поделиться записью с друзьями нажатием кнопки «рассказать друзьям».

Размещенная повторно запись может быть снова опубликована другом подписчика или подписчиком другого сообщества. Считается, что основной процесс распространения новости происходит за счет совершивших повторную публикацию посредством нажатия «рассказать друзьям», но, поскольку в новостях пользователя может также отражаться информация о новых отметках «мне нравится», совершенных его друзьями, количество таких отметок также учитывается. В результате для каждого поста, размещенного в сообществе, имеем цепочку реакций «рассказать друзьям» и «мне нравится», а также комментарии под записью.

В работе использованы данные, собранные научной группой «Алгоритмика сложных систем» Университета ИТМО [22].

Указанный набор содержит совокупность публикаций сообщества сетевой благотворительности и их характеристик. Информация собрана за период с 1 июля 2016 года по 30 июня 2017 года и содержит сведения о:

- 805 публикациях;
- 1 159 310 отметках «мне нравится»;
- 159 608 отметках «рассказать друзьям»;
- 9 136 комментариях;
- 294 345 подписчиках сообщества.

Данные были предоставлены в виде JSON-файла, разбитого затем на отдельные txt-файлы, содержащие информацию о публикациях, об отметках «мне нравится», «рассказать друзьям», о комментариях и подписчиках сообщества (табл. 1).

Таблица 1. Структура данных

Данные	Поле	Описание
О публикациях	id	Идентификатор публикации
	date	Дата размещения публикации (timestamp)
	text	Текст публикации
Об отметках «мне нравится»	itemId	Идентификатор публикации
	likerId	Идентификатор оставившего отметку «мне нравится»
Об отметках «рассказать друзьям»	from_id	Идентификатор оставившего отметку «рассказать друзьям»
	date	Время повторного размещения публикации (timestamp)
	post_id	Идентификатор публикации
О комментариях	post_id	Идентификатор публикации
	id	Идентификатор комментария
	from_id	Идентификатор комментатора
	text	Текст комментария
	date	Время комментария (timestamp)
О подписчиках сообщества	id	Идентификатор подписчика

Применение методов

Содержимое опубликованных записей в сообществе социальной сети представляет собой текст. Обработка естественного языка включает в себя очистку его от специальных символов и стоп-слов. Для количественного анализа содержимого записей и предсказания реакции на них методами машинного обучения требуется получить количественные характеристики текста, отражающие содержимое сообщения. Для этого производится анализ встречающихся в тексте слов и генерация численного вектора, т.е. выполняется векторизация публикации. Для того чтобы отразить разное качество прогноза на разных исходных данных, в работе применены три метода векторизации [23–25]:

1) в результате векторизации записей методом «Bag of Words» каждой публикации сопоставляется численный признак, характеризующий присутствие отдельных слов в публикации;

2) в результате векторизации записей методом TF-IDF каждой публикации сопоставляется набор численных признаков, используемый для оценки важности слова в контексте рассматриваемой записи;

3) при векторизации методом Word2Vec публикации представляют собой попарно суммированную сумму слов.

Для каждого слова в публикации проводился анализ корреляции его встречаемости с целевыми показателями. Поскольку для редко встречаемых слов невозможно установить связь с количеством реакций на публикацию, и, кроме того, наличие таких слов негативно влияет на прогностические свойства модели, было выбрано пороговое значение отсекающих слов по частоте присутствия. Были проведены эксперименты с разными частотами слов путем сравнения коэффициентов детерминации, и в результате не учитывались слова, значение соответствующего числового признака которых меньше 0,8. Таким образом, количество исходных и извлеченных из текста признаков было сокращено (табл. 2).

Завершающей стадией формирования обучающей выборки было добавление данных об отметках «мне нравится», «рассказать друзьям» и комментариях к каждому опубликованному посту в рассматриваемом сообществе в дополнение к его векторизации. Для решения задачи предсказания реакции были использованы методы линейной регрессии, опорных векторов, случайный лес и метод градиентного бустинга, вызываемые из библиотеки Scikit-learn языка программирования Python, а также метод MLP из библиотеки Keras. Методы машинного обучения применяются последовательно для прогнозирования числа отметок «мне нравится», «рассказать друзьям» и комментариев. На

вход каждому методу подается обучающая выборка векторизованных публикаций и количество соответствующих реакций под каждой из них. С использованием процедуры кросс-валидации измеряется коэффициент детерминации для оценки качества работы метода. Для оптимизации моделей был использован метод случайного поиска по узлам сетки параметров. Так, для метода опорных векторов экспериментальным образом было установлено, что применение линейного ядра демонстрирует наилучшие показатели против стандартного ядра «RBF» для метода векторизации TF-IDF. В результате использования метода MLP производится оценка общей характеристики публикации. Таким образом, при применении метода сразу ко всем трем целевым признакам, им будет учитываться взаимодействие реакций, при условии существования зависимости. Для метода MLP была использована трехслойная нейронная сеть, содержащая по 1000 нейронов на первых двух слоях и три на последнем. В результате оптимизации параметров было установлено, что наилучшие результаты соответствуют активационным функциям гиперболического тангенса на первом слое и ReLU на последних двух, при этом используется L1-регуляризация на каждом слое. Обучение проводилось на 1000 эпохах. Полученный результат был разбит на соответствующие показатели по каждой искомой характеристике.

Результатом работы методов стал прогноз реакции подписчиков сообщества на новые публикации, выраженный в количестве отметок «мне нравится», «рассказать друзьям» и числе комментариев в заданный промежуток времени, равный одному году, в предположении, что общая структура и интерес подписчиков к сообществу остаются прежними. Для оценки работы методов был использован коэффициент детерминации $R^2 = 1 - \text{Var}(y|x)/\text{Var}(y)$, где $\text{Var}(y|x)$ — условная дисперсия зависимой переменной по признакам x . Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Коэффициент $R^2 = 1$, означает отсутствие ошибки прогноза.

Основные результаты

Результаты применения комбинаций трех методов векторизации и пяти разных методов прогноза представлены на рисунке.

Как видно из полученных диаграмм, наилучшей комбинацией для предсказания числа комментариев является MLP при векторизации методом Word2Vec, который опережает по точности линейную регрессию (0,57 против 0,54 для линейной регрессии с векторизацией «Bag of Words»). Для прогнозирования числа отметок «рассказать друзьям» наилучшие результаты получены в случае применения метода MLP в сочетании с методом векторизации TF-IDF. Одинаковый коэффициент детерминации при прогнозировании числа реакций «мне нравится» получен для метода MLP при векторизации Word2Vec и для линейной регрессии при методе векторизации «Bag of Words». Прогнозирование числа отметок «мне нравится» и «рассказать друзьям» дает довольно низкое значение коэффициента детерми-

Таблица 2. Число используемых признаков

Метод векторизации	Число признаков до фильтрации	Число признаков после фильтрации
«Bag of Words»	2933	59
TF-IDF	2928	55
Word2Vec	29581	57

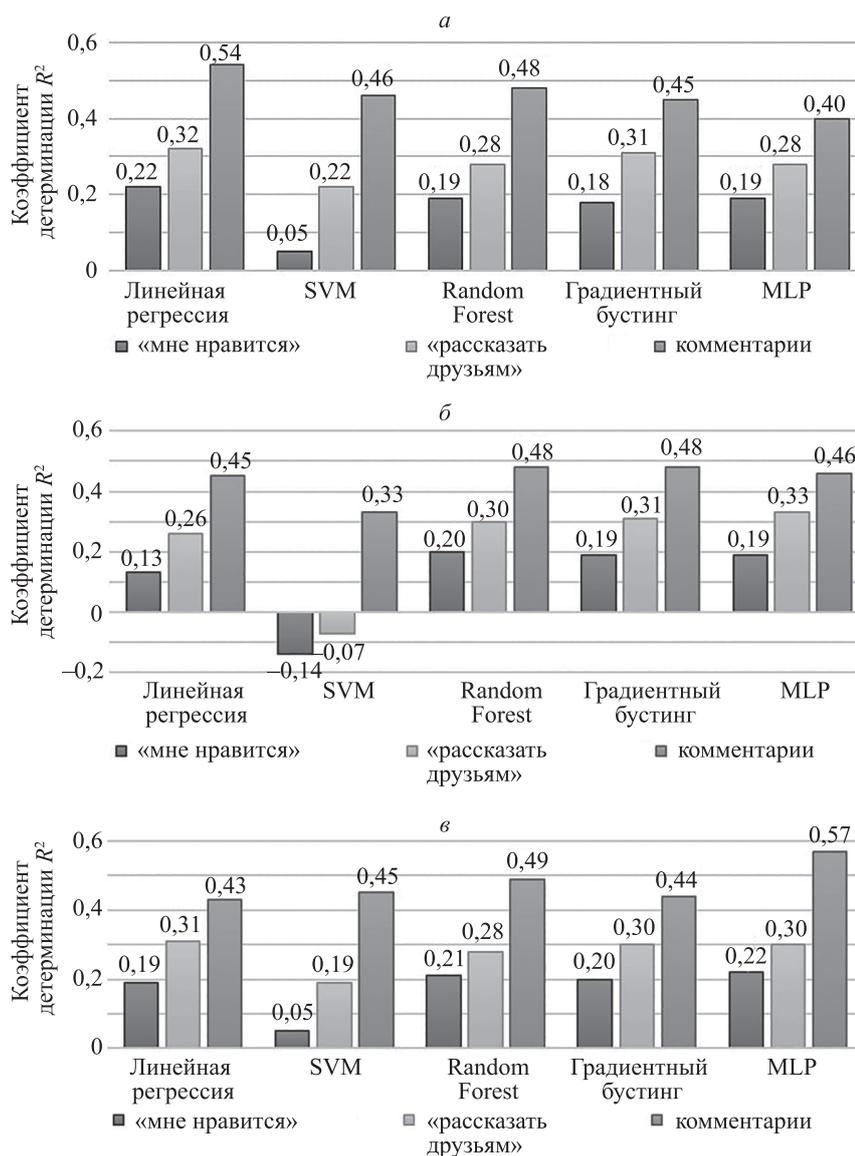


Рисунок. Качество предсказания целевых характеристик новой публикации при различных методах машинного обучения: векторизация методом «Bag of Words» (а); векторизация методом TF-IDF (б); векторизация методом Word2Vec (в)

нации, поэтому рекомендовано применение указанных методов только для прогнозирования числа комментариев под новой публикацией.

Заключение

На основе методов машинного обучения построен прогноз реакций подписчиков сообщества сетевой благотворительности социальной сети «ВКонтакте» на вновь опубликованную запись. Реакции подписчиков выражены в количестве поставленных отметок «мне нравится», «рассказать друзьям» и оставленных комментариев. Приведены результаты при различных методах векторизации публикаций.

При использовании вышеописанных методов наилучший прогноз среди представленных методов показал многослойный перцептрон при методе вектори-

зации Word2Vec, показавший значение коэффициента детерминации 0,57. Прогнозирование данными методами числа отметок «мне нравится» и «рассказать друзьям» оказывается низкоэффективным на текущем наборе данных.

Таким образом, описанные выше методы машинного обучения могут применяться для прогнозирования количества комментариев под вновь опубликованной публикацией. Для прогноза числа отметок «мне нравится» и «рассказать друзьям» рекомендуется применение других методов. Представляется возможным применить для этой задачи использование механистических моделей вместо статистических. В дальнейшем планируется попытка построения более точного прогноза числа отметок «мне нравится» и «рассказать друзьям» с помощью агрегированной эпидемической модели.

Литература

1. Суходолов А.П., Бычкова А.М. «Фейковые новости» как феномен современного медиапространства: понятие, виды, назначение, меры противодействия // Вопросы теории и практики журналистики. 2017. Т. 6. № 2. С. 143–169. doi: 10.17150/2308-6203.2017.6(2).143-169
2. Li M., Wang X., Gao K., Zhang S. A survey on information diffusion in online social networks: Models and methods // Information (Switzerland). 2017. V. 8. N 4. P. 118. doi: 10.3390/info8040118
3. Vega-Oliveros D.A., Berton L., Vazquez F., Rodrigues F.A. The impact of social curiosity on information spreading on networks // Proc. 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2017. P. 459–466. doi: 10.1145/3110025.3110039
4. Cyberemotions: Collective Emotions in Cyberspace / Ed. by J.A. Holyst. Berlin: Springer International Publishing, 2017. 318 p. (Understanding Complex Systems). doi: 10.1007/978-3-319-43639-5
5. Rodrigues H.S. Application of SIR epidemiological model: new trends // International Journal of Applied Mathematics and Informatics. 2016. V. 10. P. 92–97.
6. Mann R.P., Faria J., Sumpter D., Krause J. The dynamics of audience applause // Journal of the Royal Society Interface. 2013. V. 10. N 85. P. 20130466. doi: 10.1098/rsif.2013.0466
7. Kostylenko O., Rodrigues H.S. Torres D.F.M. Banking risk as an epidemiological model: an optimal control approach // Springer Proceedings in Mathematics and Statistics. 2018. V. 223. P. 165–176. doi: 10.1007/978-3-319-71583-4_12
8. Горковенко Д.К. Обзор моделей распространения информации в социальных сетях // Молодой ученый. 2017. № 8. С. 23–28.
9. Ландэ Д.В., Додонов В.А. Нелинейные свойства мультиагентной модели распространения новостей // Information Technology and Security. 2016. V. 4. N 2. P. 137–146.
10. Lande D.V., Hraivoronska A.M., Berezin B.O. Model of information spread in social networks // European Journal of Natural History. 2016. N 5. P. 41–44.
11. Шотыло Д.М., Крайнова В.Е., Скурыдин А.В. Тенденции развития искусственных нейронных сетей в цифровой экономике // ЭКОНОМИНФО. 2018. Т. 15. № 4. С. 65–69.
12. Seber G.A.F., Lee A.J. Linear Regression Analysis. 2nd ed. New Jersey: John Wiley & Sons, 2003. 582 p. (Wiley Series in Probability and Statistics).
13. Liu X., Zhu T. Deep learning for constructing microblog behavior representation to identify social media user's personality // PeerJ Computer Science. 2016. V. 2. P. e81. doi: 10.7717/peerj-cs.81
14. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М.: МЦМНО, 2013. 390 с.
15. Jamali S., Rangwala H. Digging digg: comment mining, popularity prediction, and social network analysis // Proc. of the 2009 International Conference on Web Information Systems and Mining (WISM). 2009. P. 32–38. doi: 10.1109/WISM.2009.15
16. Zhou Z.-H. Ensemble Methods: Foundations and Algorithms. Chapman&Hall/CRS, 2012. 236 p.
17. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. NY: Springer-Verlag, 2009. 746 p. doi: 10.1007/978-0-387-84858-7
18. Sheng Y., Subhash K. A survey of prediction using social media // Department of Computer Science, Oklahoma State University. Stillwater, Oklahoma, U.S.A. [Электронный ресурс]. URL: <https://arxiv.org/abs/1203.1647>, свободный, Яз.англ. (дата обращения: 19.08.2019).
19. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation applied to handwritten zip code recognition // Neural Computation. 1989. V. 1. N 4. P. 541–551. doi: 10.1162/neco.1989.1.4.541
20. Cai G., Xia B. Convolutional neural networks for multimedia sentiment analysis // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. V. 9362. P. 159–167. doi: 10.1007/978-3-319-25207-0_14
21. Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington: Spartan Books, 1961. 616 p.
22. Kesarev S., Severiukhina O., Bochenina K. Parallel simulation of community-wide information spreading in online social networks //

References

1. Sukhodolov A.P., Bychkova A.M. Fake News as a Modern Media Phenomenon: Definition, Types, Role of Fake News and Ways of counteracting it. *Theoretical and practical issues of journalism*, 2017, vol. 6, no 2, pp. 143–169. (in Russian). doi: 10.17150/2308-6203.2017.6(2).143-169
2. Li M., Wang X., Gao K., Zhang S. A survey on information diffusion in online social networks: Models and methods. *Information (Switzerland)*, 2017, vol. 8, no. 4, pp. 118. doi: 10.3390/info8040118
3. Vega-Oliveros D.A., Berton L., Vazquez F., Rodrigues F.A. The impact of social curiosity on information spreading on networks. *Proc. 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2017, pp. 459–466. doi: 10.1145/3110025.3110039
4. *Cyberemotions: Collective Emotions in Cyberspace* / Ed. by J.A. Holyst. Berlin, Springer International Publishing, 2017, 318 p. doi: 10.1007/978-3-319-43639-5
5. Rodrigues H.S. Application of SIR epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics*, 2016, vol. 10, pp. 92–97.
6. Mann R.P., Faria J., Sumpter D., Krause J. The dynamics of audience applause. *Journal of the Royal Society Interface*, 2013, vol. 10, no. 85, pp. 20130466. doi: 10.1098/rsif.2013.0466
7. Kostylenko O., Rodrigues H.S. Torres D.F.M. Banking risk as an epidemiological model: an optimal control approach. *Springer Proceedings in Mathematics and Statistics*, 2018, vol. 223, pp. 165–176. doi: 10.1007/978-3-319-71583-4_12
8. Horkovenko D.K. Overview of the models of information distribution in social networks. *Young Scientist*, 2017, no. 8, pp. 23–28. (in Russian)
9. Lande D., Dodonov V. Nonlinear properties of agent-based news distribution model. *Information Technology and Security*, 2016, vol. 4, no. 2, pp. 137–146. (in Russian)
10. Lande D.V., Hraivoronska A.M., Berezin B.O. Model of information spread in social networks. *European Journal of Natural History*, 2016, no. 5, pp. 41–44.
11. Shotylo D.M., Kraynova V.E., Skurydin A.V. Trends in the development of artificial neural networks in digital economy. *EKONOMINFO*, 2018, vol. 15, no. 4, pp. 65–69. (in Russian)
12. Seber G.A.F., Lee A.J. *Linear Regression Analysis*. 2nd ed. New Jersey, John Wiley & Sons, 2003, 582 p., Wiley Series in Probability and Statistics.
13. Liu X., Zhu T. Deep learning for constructing microblog behavior representation to identify social media user's personality. *PeerJ Computer Science*, 2016, vol. 2, pp. e81. doi: 10.7717/peerj-cs.81
14. Viugin V.V. *Mathematical foundations of machine learning and forecasting theory*. Moscow, MCCME Publ., 2013, 390 p. (in Russian)
15. Jamali S., Rangwala H. Digging digg: comment mining, popularity prediction, and social network analysis. *Proc. of the 2009 International Conference on Web Information Systems and Mining (WISM)*, 2009, P. 32–38. doi: 10.1109/WISM.2009.15
16. Zhou Z.-H. *Ensemble Methods: Foundations and Algorithms*. Chapman&Hall/CRS, 2012, 236 p.
17. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. NY, Springer-Verlag, 2009, 746 p. doi: 10.1007/978-0-387-84858-7
18. Sheng Y., Subhash K. *A Survey of Prediction Using Social Media*. Department of Computer Science, Oklahoma State University. Stillwater, Oklahoma, U.S.A. Available at: <https://arxiv.org/abs/1203.1647> (accessed: 19.08.2019).
19. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1989, vol. 1, no. 4, pp. 541–551. doi: 10.1162/neco.1989.1.4.541
20. Cai G., Xia B. Convolutional neural networks for multimedia sentiment analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9362, pp. 159–167. doi: 10.1007/978-3-319-25207-0_14
21. Rosenblatt F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, Spartan Books, 1961, 616 p.
22. Kesarev S., Severiukhina O., Bochenina K. Parallel simulation of community-wide information spreading in online social networks.

- Communications in Computer and Information Science. 2019. V. 965. P. 136–148. doi: 10.1007/978-3-030-05807-4_12
23. Gali K., Venkatapathy S. Sentence realisation from bag of words with dependency constraints // Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, USA, 2009. P. 19–24.
24. Shahzad Q., Ramsha A. Text mining: Use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. V. 181. N 1. P. 25–29. doi: 10.5120/ijca2018917395
25. Jang B., Kim I., Kim J.W. Word2vec convolutional neural networks for classification of news articles and tweets // PLoS ONE. 2019. V. 14. N 8. P. e0220976. doi: 10.1371/journal.pone.0220976
- Communications in Computer and Information Science*, 2019, vol. 965, pp. 136–148. doi: 10.1007/978-3-030-05807-4_12
23. Gali K., Venkatapathy S. Sentence realisation from bag of words with dependency constraints. *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, USA, 2009, pp. 19–24.
24. Shahzad Q., Ramsha A. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 2018, vol. 181, no. 1, pp. 25–29. doi: 10.5120/ijca2018917395
25. Jang B., Kim I., Kim J.W. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS ONE*, 2019, vol. 14, no. 8, pp. e0220976 doi: 10.1371/journal.pone.0220976

Авторы

Попова Елизавета Петровна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0001-5547-5191, Popova.elizaveta@list.ru

Леоненко Василий Николаевич — кандидат физико-математических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 31767480200, ORCID ID: 0000-0001-7070-6584, vnleonenko@yandex.ru

Authors

Elizaveta P. Popova — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0001-5547-5191, Popova.elizaveta@list.ru

Vasiliy N. Leonenko — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 31767480200, ORCID ID: 0000-0001-7070-6584, vnleonenko@yandex.ru