РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНСТИТУТ НАУЧНОЙ ИНФОРМАЦИИ
ПО ОБЩЕСТВЕННЫМ НАУКАМ

РЕЧЕВАЯ КОММУНИКАЦИЯ В СЕТЕВЫХ СТРУКТУРАХ: МЕЖДУ ГЛОБАЛЬНЫМ И ЛОКАЛЬНЫМ

Сборник научных трудов

МОСКВА
2022

УДК 81
ББК 81
Р 46

Серия
«Теория и история языкознания»

Отдел языкознания

Редакционная коллегия сборника:
Потапов В.В. – докт. филол. наук (отв. ред.),
Казак Е.А. – канд. филол. наук (отв. ред.)

Р 46 **Речевая коммуникация в сетевых структурах: между глобальным и локальным.** Сб. науч. трудов / РАН. ИНИОН. Отдел языкознания; отв. ред. Потапов В.В., Казак Е.А. – М., 2022. – 280 с.

ISBN 978-5-248-01015-8

Рассматриваются инновационные процессы, происходящие в Интернете и иных электронных средствах массовой коммуникации, которые привели к возникновению нового подхода к пониманию общественной жизни как существующей в условиях глобального информационного пространства.

Анализируются причины обращения к интернет-коммуникации, её особенности и отличия от традиционного прямого общения.

Особое внимание уделяется исследованию методов общения в сетевых структурах, при которых передача информации происходит с использованием различных форм её обмена и представления.

Сборник адресован лингвистам, специалистам по теории коммуникации и медиалингвистики.

УДК 81
ББК 81

© ФГБУН «Институт научной информации по общественным наукам РАН», 2022

ISBN 978-5-248-01015-8

СОДЕРЖАНИЕ

ПОСЛЕДСТВИЯ ВОЗДЕЙСТВИЯ ИНТЕРНЕТ-КОММУНИКАЦИИ НА ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ

Потапова Р.К., Потапов В.В. Аддитивная трансформация личности в эпоху цифровой коммуникации 5
Атабекова А.А. Коммуникация в социальных сетях в период пандемии коронавируса: тенденции и социальная значимость анализа 28

ПРОБЛЕМЫ ТЕОРИИ ИНТЕРНЕТ-КОММУНИКАЦИИ

Дедова О.В. Актуальные направления интернет-лингвистики: язык в Интернете vs Интернет в языке 40
Сидорова М.Ю. Субъект и субъектность в интернет-коммуникации: 20 лет спустя 65
Максименко О.И. Язык Интернета: динамика развития 106
Кучина С.А. Электронный художественный текст: основные понятия, критерии фикциональности, лингвосемиотическая специфика 125
Моисеенко Л.В. Интернет-мем как новая модальность социальной интеракции 147
Гусейнова И.А. «Скрытые угрозы» современной немецкоязычной интернет-коммуникации 167
Ахренова Н.А. Интернет-коммуникация: от личного письма до информационной войны 181

ПРИКЛАДНЫЕ АСПЕКТЫ ИНТЕРНЕТ-КОММУНИКАЦИИ

Кедрова Г.Е. Особенности интернет-опосредованной коммуникации в социальных сетях в мультикультурном полиэтническом мультиязыковом виртуальном сообществе ... 199
Хитина М.В. Исследование русскоязычного Интернета с помощью разных методов и программ 216
Комалова Л.Р., Кулагина Д.А. Восприятие вербализованной агрессии в контексте и вне контекста сообщений социальной интернет-сети «Твиттер»..... 229
Гордеев Д.И., Филиппов Д.В., Рей А.И. Специфика интернет-коммуникации в социальных сетях сообществ Севера России..... 241
Вишневская Г.М., Прохорова А.А. Интернет-коммуникация в профессиональном общении на иностранном языке: аспекты теории и практики 256
Курьянова И.В., Лопаткин С.Н. Глобальное цифровое пространство как способ и средство совершения преступлений..... 267

УДК: 81 DOI: 10.31249/rech/2022.00.13

Д.И. Гордеев[*/**/***]**, Д.В. Филиппов**[*/**]**, А.И. Рей**[*]

СПЕЦИФИКА ИНТЕРНЕТ-КОММУНИКАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ СООБЩЕСТВ СЕВЕРА РОССИИ

** Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Москва, Россия*
*** Всероссийская академия внешней торговли, Москва, Россия*
**** Московский государственный лингвистический университет, Москва, Россия {gordeev-di; filippov-dv; rey-ai}@ranepa-ru*

Аннотация. В данной работе проводится анализ сообществ социальной сети «ВКонтакте». Были построены тематические модели на основе текстовой информации. Кроме того, на основе статистических данных, характеризующих каждое сообщение и их популярность, были натренированы модели на основе алгоритма случайного леса. Модели строились для предсказания количества комментариев к сообщению, лайков, репостов, а также для времени от первого до второго комментария и времени от первого до последнего комментария. Во всех случаях метапризнаки дали прирост по сравнению с использованием только текстовой информации. Кроме того, был проведен анализ предсказаний каждой модели и основных решающих признаков.

Введение

С развитием сети Интернет начали появляться новые способы социальных взаимодействий. Интернет-дискурс обладает многими свойствами, не характерными для других типов речи. Так, сетевой дискурс обладает свойствами опосредованности общения, пространственной удаленности коммуникантов, вариативностью речевой деятельности и возможностью асинхронного общения [Поликодовая среда Интернета и проблемы валеологии, 2020; Потапова, Курьянова, 2021]. Первоначально коммуникация в сети Интернет осуществлялась посредством электронной почты. Затем

начали появляться способы массовой оперативной коммуникации, например Usenet. Со временем начали появляться сервисы и технологии для мгновенной коммуникации, например IRC, что привело к популярности чатов, которые часто локально хостились индивидуальными пользователями [Van Dijck,2013, с. 4–5]. Одной из первых социальных сетей является classmate.com, которая ставила перед собой задачу помочь людям найти бывших одноклассников и однокурсников. Появление Web 2.0 и ресурсов, направленных на генерацию контента пользователями, привели к новому витку популярности социальных сетей [Sajithra, Patil, 2013.]. Кроме того, доступ к Интернету получили многие жители развивающихся стран, среди которых была и Россия. В некоторых из этих стран локальным сервисам удалось успешно противостоять тренду растущей глобализированности Интернета [Schroeder, 2018, с. 98–99]. Одним из таких примеров является Россия и социальная сеть «ВКонтакте».

«ВКонтакте» уже не входит в топ-3 интернет-ресурсов РФ, как это было раньше [Baran, Stock, 2015]. Так, данный сервис уступает по охвату Whatsapp (согласно данным Mediascope [Mediascope], рис. 1). Однако «ВКонтакте» остается одним из важных ресурсов для общения, и в отличие от мессенджеров гораздо более открытым, что делает возможность его исследования.

Рис. 1.
Самые популярные веб-ресурсы РФ в июле 2021 г. (по данным Mediascope)

В рамках данной работы была предпринята попытка оценки того, от каких параметров могут зависеть популярность и охваты сообщений. В качестве признаков мы использовали текстовые характеристики сообщения и некоторую метаинформацию. Данную работу можно рассматривать как пример упрощенной модели распространения сообщений (без знания графовой структуры) в социальной сети и общественного резонанса.

В данной статье используются следующие термины, характеризующие специфику общения в социальной сети «ВКонтакте»:

- Пост – первоначальное сообщение в группе социальной сети
- Репост – копирование поста другим участником социальной сети
- Комментарий – последующее сообщение в ответ на пост
- Лайк – реакция «нравится» на пост

Обзор литературы

Существует много работ, посвященных анализу характеристик сообщений в социальных сетях. Так, исследователи занимались предсказанием лайков и репостов сообщений сервиса «Твиттер» [Prediction of likes and retweets using text information retrieval, 2020, p. 123–128.]. Они пробовали различные модели, наилучшие результаты показала модель на основе Doc2 vec [Le, Mikolov, 2014] векторов и RandomForest-классификатора [Liaw, Wiener, 2002] (данные предварительно были разбиты на 4 корзины). В другой работе авторы извлекали из изображений признаки посредством различных моделей, в том числе сверточных нейронных сетей, к ним добавляли TF-IDF (term frequence – inverse document frequency) текстовые признаки.

Исследование социальных сетей проводится и с других аспектов. Так, активно изучаются вбросы и сообщения, которые распространяют дезинформацию и «фейковые» новости. Авторы одной из подобных работ применяют методы «опорных векторов», «случайного леса» и «наивного Байеса» для оценки фейковости сообщения в сети «Твиттер» на основе размеченных данных [Cu□maliuc, Coca, Iftene, 2018]. Из анализируемых моделей лучше всего показан метод на основе случайного леса.

Социальная сеть «ВКонтакте» также активно исследуется. Одна из подобных работ анализировала, является ли аккаунт ботом или нет [Samokhvalov, 2020]. Для этого они использовали биб-

лиотеку градиентного бустинга Catboost. Данные собирались следующим образом: каждый день скрейпер собирал информацию о 120 000 самых новых аккаунтов пользователей сети «ВК», часть из которых блокировалась администрацией. Так, авторы собрали информацию о 3,5 млн аккаунтах, 42 тыс. из которых оказались заблокированными.

Сбор данных

Для анализа была решено отобрать сравнительно непопулярные сообщества приблизительно одной тематики, для того чтобы минимизировать сетевые эффекты. Было выбрано три группы по тематике Севера России. Данные из сообществ собирались посредством API-«ВКонтакте» с использованием библиотек requests и asyncio. Собирались текст сообщения и мета-признаки: наличие или отсутствие прикрепленного изображения, дата сообщения и его комментарии. На основе этих данных статистические модели должны были осуществлять предсказания. Для этого использовались только текстовые сообщения, без анализа графов друзей и связей пользователей. Данные собирались с начала существования сообществ и по декабрь 2020 г.

Пользователи сообществ во многом пересекаются (рис. 2). Данная информация использовалась для оценки сетевого эффекта и при тренировке моделей не учитывалась. Модели тренировались для предсказания пяти показателей: числа лайков данного сообщения, числа репостов, числа комментариев для данного сообщения, а также предсказывались время до последнего комментария и время до второго комментария. Большинство сообщений во всех сообществах не имело комментариев (табл. 1). Кроме того, средние значения значительно различаются по разным сообществам, что необходимо учитывать при тренировке моделей. Мы считаем, что данные показатели позволяют оценить возможную популярность и «вирусность» сообщения.

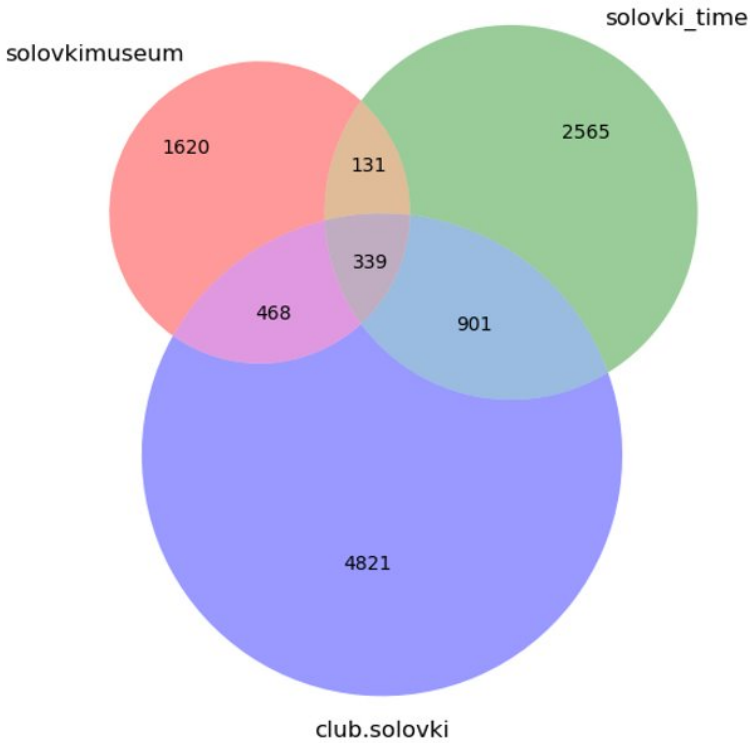

Рис. 2.
Пересечения по пользователям сообществ

Таблица 1
Количественные характеристики исследуемых сообществ

	solovkimuseum	**solovki_time**	**club.solovki**
Лайков, всего	55115	60638	472856
Постов, всего	1563	6131	10955
Среднее число лайков	35	9	43
Постов без комментариев	1213	5121	7751

Стоит также отметить, что данные предсказываемые параметры сильно зависят от даты. Так, часть набора данных приходится на карантин 2020 г., вызванный пандемией COVID-19.

В связи с этим с марта-апреля 2020 г. наблюдается резкий скачок в числе сообщений (рис. 3). Всё это, конечно же, может сказываться и на предсказываемых характеристиках данных сообщений.

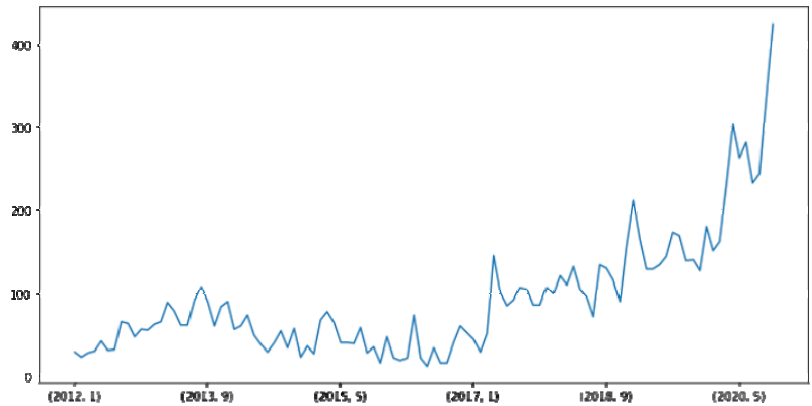

Рис. 3.
Число постов по месяцам

Однако в реальности прироста характеристик не наблюдается по большинству параметров: растет только количество комментариев, а просмотры в среднем на пост даже падают (рис. 4).

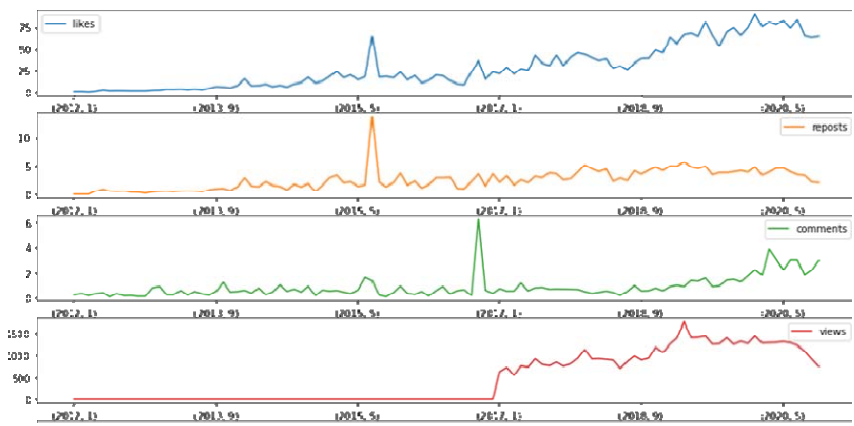

Рис. 4.
Число лайков, репостов, комментариев и просмотров в среднем за месяц

Если же данные нормализовать, то даже наблюдаемый рост по числу комментариев пропадает в период пандемии COVID-19 (рис. 5).

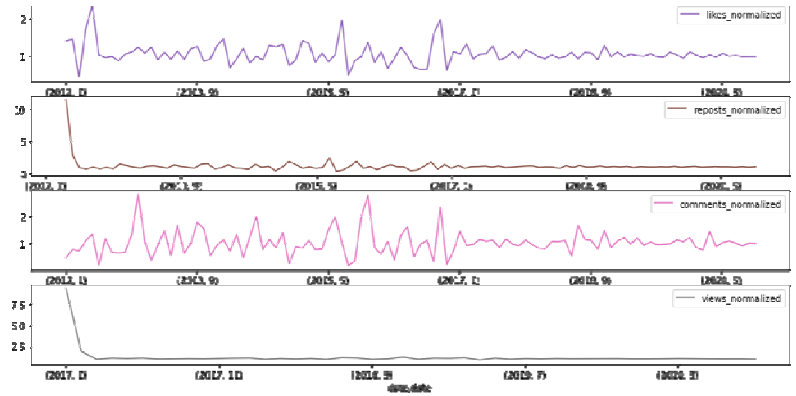

Рис. 5.
Нормализованное число лайков, репостов, комментариев и просмотров в среднем за месяц

Стоит еще отметить, что характеристики сообщений сильно зависят от времени суток и дня недели. Все действия вероятнее всего в данных сообществах в период с 20 вечера и до четырех утра по московскому времени. При этом сами посты, наоборот, в этом временном промежутке почти не появляются (рис. 6 и 7). И для предсказания временных параметров, таких как время до второго и последнего комментария, время поста, приобретает еще большее значение.

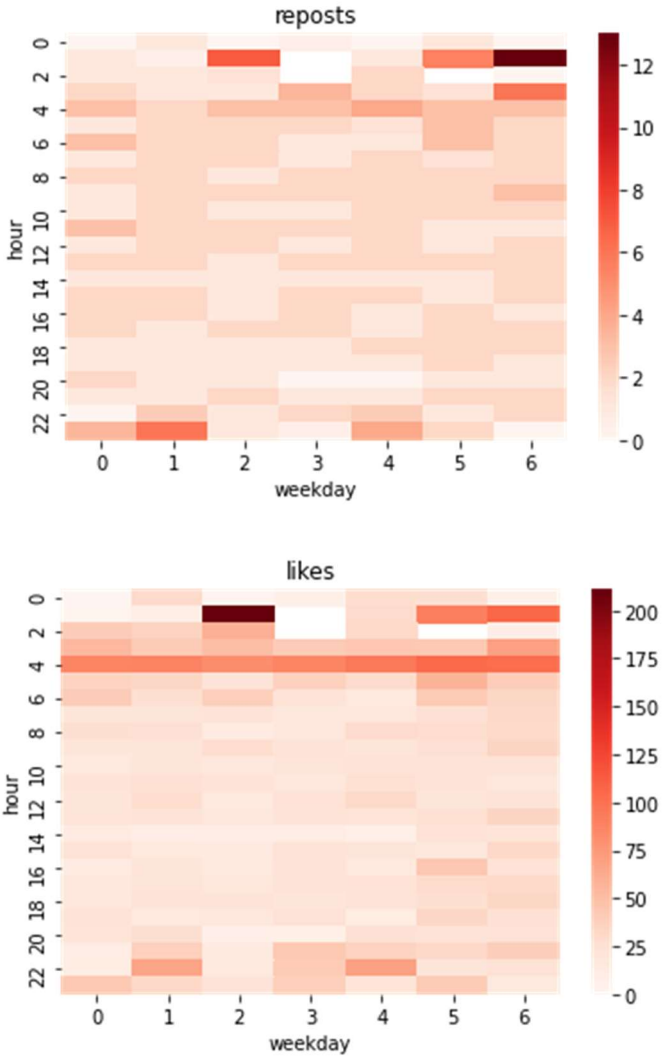

Рис. 6.

Зависимость количества репостов и лайков в зависимости от дня недели и часа. Нулевой день – это понедельник, шестой – воскресенье

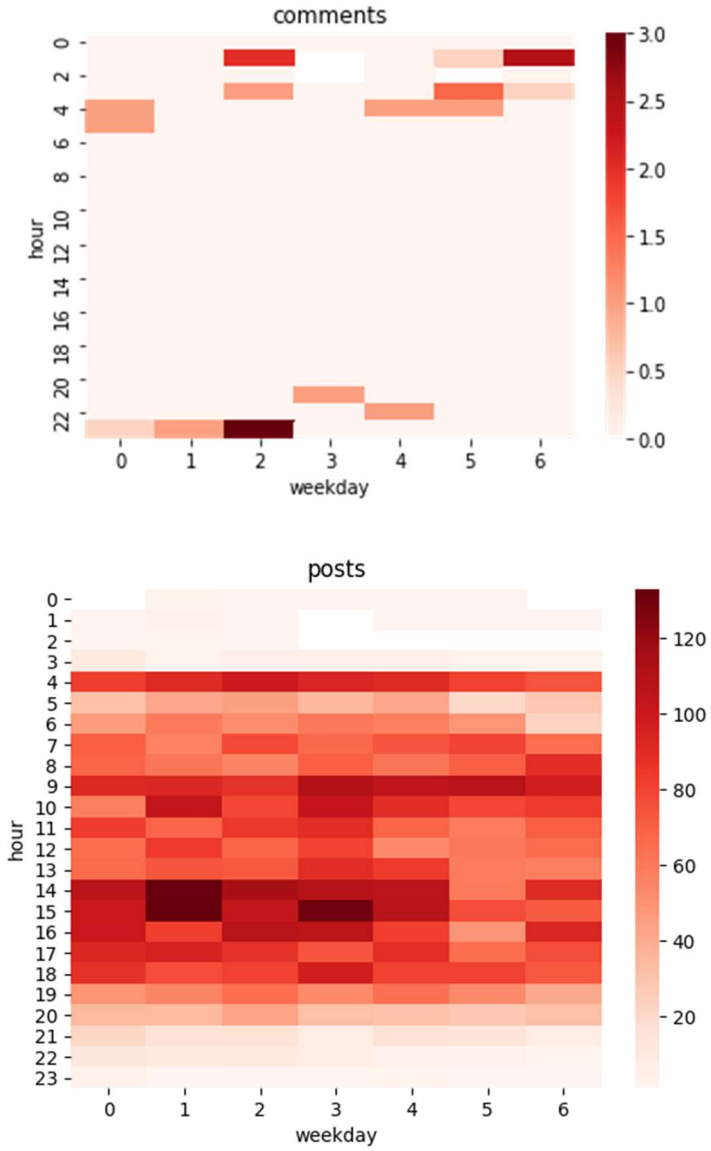

Рис. 7.

Зависимость количества комментариев и числа постов в зависимости от дня недели и часа.
Нулевой день – это понедельник, шестой – воскресенье.

Обучение моделей

В рамках данной работы была предпринята попытка предсказания пяти показателей:

1) числа лайков сообщения;
2) числа репостов;
3) числа комментариев под сообщением;
4) разница во времени от первого до последнего комментария;
5) разница во времени от первого до второго комментария.

Каждая из задач решалась как задача регрессии, где по входным текстовым и метапризнакам предсказывался численный показатель. В качестве метапризнаков использовались:

- Идентификатор автора поста (исследуемое сообщество или изначальный автор в случае репоста сообществом)
- Сообщество
- Час
- Год
- Месяц
- День недели
- Автоматически определяемая тональность с помощью библиотеки Dostoevsky [RuSemtiment: An enriched sentiment analysis dataset for social media in Russian, 2018]
- Тематики, определяемые с помощью латентного размещения Дирихле (LDA: Latent Dirichlet allocation) [Blei, Ng, Jordan, 2003].

Обучающая выборка была разбита на тренировочный, валидационный и тестовый датасеты. Тренировочный набор данных составлял 80%, а валидационный и тестовый – по 10% от всех строк датасета. Во всех случаях выборка делилась хронологически. Тестовые данные брались с мая 2020 года по ноябрь 2020 года. Валидационные – с января по апрель 2020 года.

Мы проверили несколько моделей регрессионных моделей:

- Lasso (Регрессию лассо) – линейная регрессия с L1-регуляризацией
- Ridge (Гребневую регрессию) – линейная регрессия с L2-регуляризацией
- Elasticnet (эластичная сеть) – линейная регрессия с L1 и L2 регуляризаторами
- Модель случайного леса.

Все модели брались из библиотеки scikit-learn для языка программирования Python.

В качестве основной метрики использовалась RMSE (root mean square error)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}},$$

где $\hat{y}$ – это предсказания модели, y – истинное значение для данного примера, а n – количество примеров в тестовой выборке.

Результаты

Сначала мы провели отбор лучшей модели на задаче предсказания лайков как наиболее простой. Отбор модели проводился на отдельной ограниченной выборке без хронологической привязки. Результаты представлены в таблице 2. По всем параметрам лучше всего себя показала модель «случайного леса». Для последующих задач мы использовали уже исключительно ее.

Таблица 2

Количественные характеристики исследуемых сообществ

	Lasso (alpha=0.01)	**Ridge (alpha=1)**	**Elasticnet (alpha=0.91)**	**Random forest**
RMSE (корень среднеквадратичной ошибки)	38.43	36.7	42.48	30
R^2	0.42	0.47	0.29	0.64
MAE (средняя абсолютная ошибка)	25.52	24.1	29.72	17.87
MSE (среднеквадратичная ошибка)	1477.62	1347.05	1804.34	905.92

В дальнейшем на основе алгоритма случайного леса были обучены модели для каждой из задач. Результаты данных моделей представлены в таблице 3. Параметры для каждой задачи перебирались с помощью модуля GridSearch из sklearn на основе валидационной выборки. По каждой задаче строилось две финальных

модели: одна с использованием только текстовых признаков, другая – с присутствием метапризнаков. Тем самым мы могли проверить, достаточно ли только текстовой информации для предсказания искомых характеристик. Для всех исследуемых задач метаинформация дала прирост по сравнению с использованием только текстовых данных. В целом стоит отметить, что качество моделей является сравнительно низким, и моделям не удалось надежно предсказать структурный сдвиг в общении в сети Интернет после пандемии коронавируса.

Таблица 3

Количественные характеристики исследуемых сообществ

Задача	Только текстовые признаки		Все признаки	
	RMSE	R^2	RMSE	R^2
Предсказание лайков	55.03	0.21	48.9	0.38
Предсказание репостов	4.22	0.11	4.21	0.11
Предсказание числа комментариев	1.74	0.44	1.53	0.57
Время от первого до последнего комментария (в секундах)	319629 ~89 часов	0.3	315266	0.32 ~88 часов
Время от первого до второго комментария	247697 ~69 часов	0.305	242356	0.334 ~67 часов

Кроме того, был проведен анализ вклада каждой модели и основных признаков, участвующих в предсказании тех или иных значений. Это возможно из-за сравнительно простой интерпретируемости моделей на основе случайных деревьев (рис. 8). Мы использовали важность признаков, которая показывает вклад признака в критерий информативности для всех деревьев ансамбля «случайного леса» (feature_importance из sklearn).

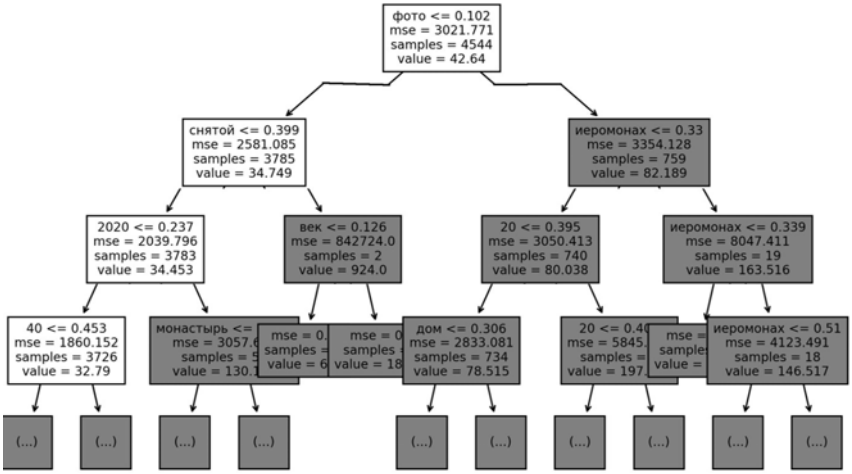

Рис. 8.
Пример натренированного классификатора-дерева из модели «случайного леса»

Важность признаков можно увидеть в таблицах 4 и 5. Как видно из данных таблиц, тональности и тематики LDA часто являются наиболее важными метапризнаками. Кроме того, при отсутствии данных о годе модели часто стремятся извлечь эти данные из текстов, что видно на примере ансамблей деревьев для предсказания числа лайков сообщения.

Таблица 4

Основные признаки для репостов и лайков

Текстовые признаки	Все признаки	Текстовые признаки	Все признаки
1) соловецкий	1) topic_0 (монастырь)	1) фото	1) sentiment_speech
2) видео	2) hour	2) соловка	2) year
3) монастырь	3) видео	3) соловецкиймузей	3) topic_4
4) соловка	4) day	4) соловецкий	4) фото
5) песнопение	5) sentiment_negative	5) 2020	5) topic_0 (монастырь)
6) ильин	6) sentiment_skip	6) монастырь	6) topic_6 (фотоконкурс)
7) документальный	7) sentiment_positive	7) 2019	7) hour
8) фото	8) sentiment_neutral	8) декабрь	8) sentiment_skip
9) песня	9) topic_12 (скит)	9) закат	9) day
10) светолетопись	10) month	10) мор	10) sentiment_negative

Таблица 5

Основные признаки для числа комментариев

Текстовые признаки	Все признаки
1) янушко	1) гулять
2) гулять	2) янушко
3) разница	3) year
4) опрос	4) разница
5) привозить	5) опрос
6) прислать	6) сформулировать
7) капитан	7) day
8) публиковать	8) привозить
9) приложение	9) кошка
10) изменить	10) капитан

Кроме того, анализ входных признаков моделей показывает, что разные признаки сообщений зависят от разных параметров. Так, количество лайков повышается для сообщений с ключевыми словами «фото» и «видео» (и, соответственно, при их наличии). А количество комментариев уже повышается от совершенно других признаков. Так, выше видно, что фамилия местного политического деятеля приводила к росту числа комментариев. Репосты во многом близки по характеристикам к лайкам, и их число возрастает для сообщений с ключевыми словами «фото» и «видео».

Заключение

В данной статье был проведен анализ сообществ Севера России социальной сети «ВКонтакте». Исследовались характеристики сообщений и были построены модели по предсказанию таких характеристик сообщений, как лайки, репосты, число комментариев, время между первым вторым и последним комментарием и первым и вторым. Данные предсказания могут использоваться для моделирования распространения и вирусности сообщений в социальных сетях. Мы выяснили, что разные виды взаимодействий с постами зависят от разных входных параметров. Так, имена политических фигур в тексте поста приводят к более высокому числу

комментариев, а упоминания фото и видео способствуют увеличению числа лайков и репостов.

Список литературы

Поликодовая среда Интернета и проблемы валеологии / Потапова Р.К., Потапов В.В., Лебедева Н.Н., Агибалова Т.В. / под ред. Р.К. Потаповой. – М. : Издательский Дом ЯСК, 2020. – 136 с.

Потапова Р.К., Курьянова И.В. Особенности исследования текста в эпоху цифровой коммуникации // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2021. – Т. 20, № 2. – С. 5–15.

Baran K.S., Stock W.G. Facebook has been smacked down. The Russian special way of SNSs: Vkontakte as a case study // Proceedings of the 2 nd European Conference on Social Media (ECSM 2015), 9. – 10. July 2015, Porto, Portugal. – Portugal, 2015. – P. 574–582.

Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // The Journal of machine Learning research. – 2003. – Vol. 3. – С. 993–1022.

Cumaliuc C.G., Coca L.G., Iftene A. Identifying Fake News on Twitter using Naive Bayes, SVM and Random Forest Distributed Algorithms // Proceedings of the 13 th Edition of the International Conference on «Linguistic Resources and Tools for Processing Romanian Language» Iasi, 22–23 November, 2018. – Iasi, 2018. – P. 177–188.

Liaw A., Wiener M. Classification and regression by randomForest // R news. – 2002. – Vol. 2. – N 3. – P. 18–22.

Le Q., Mikolov T. Distributed representations of sentences and documents // Proceedings of the 31 st International Conference on Machine Learning. – PMLR, 2014. – P. 1188–1196.

Mediascope [онлайн ресурс]. – URL: https://webindex.mediascope.net/top-resources/projects (дата обращения: 05.10.2021).

Prediction of likes and retweets using text information retrieval / Daga I., Vardhan R., Gupta A., Mukherjee P. // Procedia Computer Science. – 2020. – Vol. 168. – P. 123–128.

RuSemtiment: An enriched sentiment analysis dataset for social media in Russian / Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. // Proceedings of the 27 th international conference on computational linguistics. – 2018. – P. 755–763.

Sajithra K., Patil R. Social media–history and components // Journal of Business and Management. – 2013. – Vol. 7, N 1. – P. 69–74.

Samokhvalov D.I. Machine learning-based malicious users' detection in the VKontakte social network // Труды института системного программирования РАН. – М., 2020. – Vol. 32, N 3. – P. 109–117.

Schroeder R. Social theory after the internet: Media, technology and globalization. – UCL Press, 2018. – 210 p.

Van Dijck J. The culture of connectivity: A critical history of social media. – Oxford : Oxford University Press, 2013. – 228 p.