ПРОГНОЗИРОВАНИЕ ДИНАМИКИ ОБЩЕСТВЕННОГО МНЕНИЯ ПО ЛОНГИТЮДНЫМ ДАННЫМ ВЫСОКОЙ ГРАНУЛЯРНОСТИ: МОДЕЛЬ АБЕЛЬСОНА, РЕГРЕССИОННЫЕ МОДЕЛИ И АНСАМБЛИ МОДЕЛЕЙ[1]

Бузиков М. Э.[2]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)
Петелина Ю. А.[3]
(*Ozon Tech, Москва*)
Красоткин С. А.[4], **Рыжов М. С.**[5]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)
Козицин И. В.[6]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Московский физико-технический институт, Москва*)

Рассматривается задача прогнозирования динамики общественного мнения на лонгитюдных данных высокой гранулярности, добытых из социальной сети ВКонтакте. Данная задача была предложена участникам Хакатона «УБС Challenge'2024» как одного из досуговых мероприятий XX Всероссийской школы-конференции молодых ученых «Управление большими системами» (УБС), проводившейся в г. Новочеркасске в 2024 году. Настоящая работа посвящена подробному описанию проведенного Хакатона и решений, предложенных его участниками. Для выборки из $N = 1\,648\,829$ пользователей на основании двух снимков их мнений, сделанных с разницей в полгода (в феврале и июле 2018 года), было необходимо построить прогноз распределения общественного мнения в декабре 2018 года. В распоряжении участников также была информации о структуре дружеских связей пользователей. Наибольшую точность прогноза показал ансамбль из двух моделей – модели Абельсона, усиленной при помощи оценки влиятельности по собственному вектору, и модели постоянного тренда.

---

[1] *Авторы выражают признательность С.А. Красновой и Д.А. Новикову.*
[2] *Максим Эмонайевич Бузиков, к.ф-м.н., с.н.с.* (*me.buzikov@physics.msu.ru*).
[3] *Юлия Александровна Петелина, руководитель группы.* (*ptlna@yandex.ru*).
[4] *Семён Александрович Красоткин, математик.* (*semen.krassotkin@gmail.com*).
[5] *Максим Сергеевич Рыжов, н.с.* (*ryzhov@phystech.edu*).
[6] *Иван Владимирович Козицин, к.ф-м.н., с.н.с.* (*kozisin.ivan@mail.ru*).

Ключевые слова: модели социального влияния, искусственный интеллект, социальные сети, центральность по собственному вектору, ансамбли моделей.

1. Введение

Моделирование человеческого поведения является актуальной проблемой современной науки, находящейся на стыке таких дисциплин как теория управления, экономика, сетевой анализ и социальная психология [26]. Понимание особенностей формирования мнений людей в социальных группах необходимо для прогнозирования динамики и разработки эффективных стратегий управления социальными системами. Вместе с тем необходимо отметить, что на данный момент для подавляющего большинства моделей, описывающих данные явления (так называемые модели динамики/формирования мнений или модели влияния) еще даже не проведена валидация на эмпирических данных [9]. Здесь можно отметить модель Фридкина – Джонсона, которая была валидирована в рамках лабораторных экспериментов с малыми ($\sim$ 4 человека) группами [12]. При этом эмпирические исследования указывают на непредсказуемость динамики мнений людей, когда сам факт изменения мнения спрогнозировать далеко не всегда представляется возможным [18].

В рамках XX Всероссийской школы-конференции молодых ученых «Управление большими системами» (УБС), проводившейся в г. Новочеркасске в 2024 году, было организовано уже традиционное соревнование по искусственному интеллекту и анализу данных – Хакатон «УБС Challenge'2024». В этот раз участникам было предложено разработать модели для прогнозирования динамики политических взглядов пользователей социальной сети ВКонтакте. Имея в распоряжении два снимка мнений выборки $N = 1\,648\,829$ пользователей, сделанных с разницей в полгода – феврале и июле 2018 года, – а также структуру их дружеских связей, было необходимо спрогнозировать, каким будет распределение мнений этих пользователей еще через полгода – в декабре 2018 года.

В соревновании приняло участие три команды (перечислены в порядке, в котором были присланы решения):

– Максим Бузиков (научный сотрудник лаборатории №38 «Управления по неполным данным») ИПУ РАН и Юлия Петелина (team leader, Ozon Tech);

– Семен Красоткин (математик лаборатории №90 «Цифровых систем управления в агропромышленном комплексе» РАН);

– Максим Рыжов (научный сотрудник лаборатории №38 «Управления по неполным данным» РАН).

Для решения поставленной задачи участники Хакатона применили широкий спектр подходов, включая как классические агент-ориентированные модели влияния (модель Абельсона), так и различные линейные модели, а также ансамбли моделей. Настоящая статья посвящена описанию поставленной в рамках Хакатона задачи, предложенных участниками Хакатона решений, а также сравнительному анализу их точности.

Результаты, представленные в статье, ценны в первую очередь тем, что являются одной из немногих попыток построения и валидации прогноза динамики мнений людей на реальных данных высокого уровня детализации, в том числе при помощи аппарата моделей формирования мнений. С другой стороны, наши результаты наглядно демонстрируют необходимость применения модельного подхода для решения такого рода задач: прогнозы, основанные на классических моделях влияния, усиленных методами сетевого анализа (оценка влиятельности вершин при помощи центральности по собственному вектору) оказались более точными, чем линейные data-driven модели, лишенные априорных знаний об устройстве системы. При этом использование ансамблей из моделей (в частности, модели Абельсона и модели постоянного тренда) позволило превзойти по точности нелинейные модели.

Практическая ценность связана с задачами мониторинга и прогнозирования динамики общественного мнения, планированием информационных, маркетинговых и политических кампаний.

Далее в главе 2 будет приведен краткий обзор литературы по прогнозированию динамики общественного мнения, в главе 3 будет сформулирована задача, предложенная участниками, в главе 4 будут введены метрики качества и критерии точности, в главе 5 будут описаны решения, предложенные участниками. Глава 6 посвящена обсуждению полученных результатов и дальнейших направлений исследований.

2. Обзор литературы

Моделирование процессов социального влияния и изменения мнений в обществе играет важную роль в политологии и социологии, а также в смежных областях – в государственном управлении. Начиная примерно с середины 20-го века (с хрестоматийной работы Френча – ДеГроота [8, 10]), в научной литературе было предложено колоссальное количество математических моделей, количественно описывающих те или иные аспекты социального влияния и информационных взаимодействий в социальных сетях [25]. При помощи этих моделей удалось объяснить различные социальные макро- и мезо-явления, например, консенсус, фрагментацию и поляризацию общественного мнения [9], а также образование информационных пузырей и информационных петель [6].

Необходимо отметить, что при всей проработанности теоретической составляющей данных моделей вопрос их применения для описания реальных социальных процессов и, в частности, прогнозирования динамики мнений людей – одной из ключевых прикладных задач в данной предметной области, – является малоизученным, особенно если говорить про классические, агент-ориентированные модели динамики мнений, опирающиеся на признанные в социальной психологии микро-постулаты социального влияния [21].

На данный момент достаточно успешно решаются задачи мониторинга состояния общественного мнения [16], а также определения наиболее влиятельных людей в социальной сети [5][1].

[1] *Здесь под социальной сетью понимается любая социальная сетевая струк-*

С одной стороны, причина такого состояния дел кроется в трудоемкости сбора данных, ведь для построения прогноза и проверки его качества необходимо иметь хотя бы два снимка состояния социальной системы с детализацией вплоть до мнений и социальных контактов отдельных людей, так как применение агент-ориентированных моделей социального влияния требует калибровки параметров отдельных агентов [27]. На данный момент работа с эмпирическими данными, как правило, ограничивается анализом макро-состояний социальных систем (к примеру, распределений общественного мнения), которые можно получать путем проведения обезличенных социальных опросов, их темпоральных паттернов (опять же, обезличенных) и ответом на вопрос «какие микро-механизмы социального влияния могли бы привести к таким распределениям мнений и/или их динамике?» [9].

В противовес классическим агент-ориентированным моделям можно поставить data-driven модели [19, 28], в основе которых лежат технологии искусственного интеллекта. Такие модели агностичны к механизмам формирования мнений, не имеют никакой априорной информации о том, как человек должен менять свои взгляды под действием внешних стимулов. Эти паттерны выявляются моделями на стадии обучения и уже после этого строится прогноз.

Вместе с тем, последние исследования предлагают комбинировать методы и подходы из машинного обучения с концептами из социальной психологии – см. работу [21].

3. Описание данных, обозначения и постановка задачи

В распоряжении участников Хакатона был фрагмент данных, добытых из социальной сети ВКонтакте путем повторяющихся во времени измерений крупномасштабной выборки, включающей

тура, в которой вершинам соответствуют люди, а ребрам – различного рода связи между ними [20].

$N = 1\,648\,829$ российских пользователей, которые были получены из генеральной совокупности путем наложения естественных фильтров с последующим случайным отбором – см. [18] для более подробного ознакомления с технологии отбора. В феврале (T_1), июле (T_2) и декабре (T_3) 2018 года для данной выборки пользователей были выгружены подписки на информационные источники (публичные страницы и блогеров[1]), на основании которых при помощи алгоритма искусственного интеллекта, описанного в работе [16], были построены оценки мнений относительно политической повестки. Оценка мнения x_i пользователя i является логистической функцией от свертки информационных источников, на которые он подписан, взятых с весами, идентифицированными на стадии обучения модели. Данная оценка $x_i \in [0, 1]$ показывает, насколько оппозиционен данный пользователь по отношению к действующей в Российской Федерации власти: $x_i = 0$ – полная поддержка оппозиции, $x_i = 0{,}5$ – нейтральная позиция, $x_i = 1$ – приверженность провластным взглядам.

Снимок мнений пользователей в момент времени $t \in \{T_1, T_2, T_3\}$ будем далее обозначать при помощи вектора

$$\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t)).$$

Помимо этого, в июле 2018 года были выгружены дружеские связи пользователей (связь типа «дружба» предусмотрена функционалом ВКонтакте и образуется, когда пользователь подписывается на другого пользователя путем нажатия клавиши «добавить в друзья», а второй отвечает на эту заявку). Полученный в итоге граф дружеских связей является связным (была проведена предварительная фильтрация вершин, не входящих в гигантскую компоненту связности – доля таких вершин составила менее одного процента). Данный граф можно формализовать при помощи матрицы смежности

$$\mathbf{A} \in \{0, 1\}^{N \times N},$$

[1] *Данные аккаунты являются основными продуцентами информации в ВКонтакте. В частности, аккаунты СМИ по классификации ВКонтакте являются публичными страницами.*

компоненты которой $a_{i,j}$ удовлетворяют условию $a_{i,j} = a_{j,i}$ (матрица симметрична) и определяются следующим образом: $a_{i,j} = 1$ тогда и только тогда, когда i и j друзья. Далее множество друзей пользователя i будем обозначать V_i:

$$V_i = \{j \in \{1, \ldots, N\} \mid a_{i,j} = 1\}.$$

Число пользователей выборки, имеющих мнение из заданного промежутка $\Delta x \subseteq [0, 1]$ в момент времени t, обозначим $y^{\Delta x}(t) \in [0, 1]$:

$$y^{\Delta x}(t) = \frac{\text{card}\,\{i \in \{1, \ldots, N\} \mid x_i(t) \in \Delta x\}}{N},$$

где $card\,\{\ldots\}$ обозначает кардинальное число множества $\{\ldots\}$.

Задача, которая была предложена участникам Хакатона, формулировалась следующим образом: на основании первых двух снимков мнений и снимка дружеских связей необходимо построить оценку распределения мнений пользователей в момент времени T_3. При этом распределение мнений задавалось путем бинирования пространства мнений $[0, 1]$ на 10 промежутков равной длины 1/10:

$$y^{[0,\frac{1}{10})}(T_3), \quad y^{[\frac{1}{10},\frac{2}{10})}(T_3), \quad \ldots, \quad y^{[\frac{9}{10},1]}(T_3).$$

Фокус данной постановки заключается в прогнозировании динамики общественного мнения при помощи атомарной информации на уровне отдельных пользователей. Было принято решение отказаться от задачи предсказания траекторий мнений отдельных пользователей как малоперспективной и труднорешаемой: несмотря на высокий уровень детализации имеющихся в распоряжении данных, в них не содержится информация, позволяющая сделать вывод об индивидуальных поведенческих особенностях пользователей. В качестве такой информации могли бы послужить данные о том, как много времени пользователь проводит онлайн, с кем из своих друзей он коммуницирует более регулярно, а с кем конфликтует [4]. Влиятельность пользователей может зависеть от их статуса[1], а доверие друг к другу может увеличиваться в случае принадлежности к одной социальной группе,

[1] *В качестве простейшей прокси статуса может выступать центральность вершины в сети, однако такой (структурный) подход игнорирует другие сферы социальных коммуникаций.*

что также не отражено в данных [11]. Полезной могла бы быть информация касательно демографических характеристик пользователей [13, 23, 24]. С другой стороны, эмпирические исследования свидетельствуют о том, что даже сам факт изменения мнения отдельного индивида (не говоря про его количественные аспекты) является достаточно сложным для прогнозирования [18, 22].

4. Метрика качества прогноза и ориентиры

Для оценки точности прогноза использовалась суммарная абсолютная ошибка AE. Пусть вектор

$$\boldsymbol{y}(T_3) = \left(y^{[0,\frac{1}{10})}(T_3), y^{[\frac{1}{10},\frac{2}{10})}(T_3), \ldots, y^{[\frac{9}{10},1]}(T_3) \right)$$

задает реальное состояние общественного мнения в момент времени T_3 (известное организаторам, но неизвестное участникам), а вектор

$$\boldsymbol{y}_{pred} = (y_1, \ldots, y_{10})$$

– его прогноз, полученный в рамках некоторой прогностической модели. Тогда суммарная абсолютная ошибка AE прогноза $\boldsymbol{y}^{pred}$ определяется согласно следующему выражению:

$$AE(\boldsymbol{y}_{pred}) = ||\boldsymbol{y}(T_3) - \boldsymbol{y}_{pred}||_1 = |y^{[0,\frac{1}{10})}(T_3) - y_1| + \ldots + |y^{[\frac{9}{10},1]}(T_3) - y_{10}|.$$

В качестве ориентиров рассматривались два базовых прогноза. Первый – это прогноз постоянного тренда, определяемый линейным соотношением

$$\boldsymbol{y}_{const} = \boldsymbol{y}(T_2) + \boldsymbol{y}(T_2) - \boldsymbol{y}(T_1).$$

Его точность составила

$$AE(\boldsymbol{y}_{const}) = 40 \times 10^{-4}.$$

Второй базовый прогноз основан на применении агент-ориентированной модели динамики мнений (SCARDO-модели [17]), и его точность составила

$$AE(\boldsymbol{y}_{SCARDO}) = 37 \times 10^{-4},$$

более подробно – см. работу [1]. Отметим, что данная модель является нелинейной.

Оба эти ориентира были озвучены участникам на установочной лекции, как и их точность.

5. Решения

Далее в данной главе будут приведены решения, предложенные участниками (в порядке, в котором они были получены).

5.1. Решение 1 (Максим Бузиков и Юлия Петелина)

Предположим, что динамика изменения мнений описывается с помощью

$$\dot{\boldsymbol{x}} = \mathbf{W}\boldsymbol{x}, \tag{1}$$

где $\mathbf{W} \in \mathbb{R}^{N\times N}$. Это так называемая модель Абельсона [2] (аналог в непрерывном времени более известной модели ДеГроота [8]). Матрица $\mathbf{W}$ в модели Абельсона отражает структуру влияний в обществе: с какой силой агенты влияют друг на друга. Данную матрицу можно идентифицировать таким образом, чтобы итоговое решение

$$\boldsymbol{x}(t) = e^{\mathbf{W}(t-t_1)}\boldsymbol{x}(t_1)$$

задачи Коши (1) с начальным условием $\boldsymbol{x}(T_1)$ было как можно ближе к $\boldsymbol{x}(T_2)$ в момент $t = T_2$. Таким образом, элементы матрицы $\mathbf{W}$ являются оптимизируемыми величинами. Без дополнительных ограничений оптимизация в пространстве $\mathbb{R}^{N\times N}$ для больших N затруднительна с вычислительной точки зрения, поэтому предпочтительно свести количество неизвестных параметров хотя бы к величине, пропорциональной N.

Сделаем дополнительные предположения относительно вида матрицы $\mathbf{W}$. Будем характеризовать важность каждого пользователя с помощью центральности по собственному вектору [3]:

$$\mathbf{c} = \frac{\mathbf{A}\mathbf{c}}{\lambda_{\max}(\mathbf{A})}.$$

Здесь $\lambda_{\max}(\mathbf{A})$ соответствует максимальному по модулю собственному значению матрицы $\mathbf{A}$. Так как рассматриваемый граф является ненаправленным и связным, то его матрица смежности неразложима, а значит, в силу теоремы Фробениуса – Перрона собственное число $\lambda_{\max}(\mathbf{A})$ положительно. Соответствующий собственный вектор $\mathbf{c} \in \mathbb{R}^N$ также можно выбрать так, что все его компоненты будут положительны. Именно он определяет

важность вершин в рамках центральности по собственному вектору. Также нам понадобится вектор взвешенных мнений друзей пользователя, где в качестве весов выступают компоненты вектора центральностей $\mathbf{c}$:

$$\boldsymbol{w}(\boldsymbol{x}(t)) = \frac{\mathbf{A}(\mathbf{c} \odot \boldsymbol{x}(t))}{\mathbf{A}\mathbf{c}} \in \mathbb{R}^N.$$

Здесь $\odot$ означает произведение Адамара (покомпонентное умножение), и деление на вектор понимается также в покомпонентном смысле.

Будем считать, что скорость изменения мнения пользователя пропорциональна разнице мнения самого пользователя и средневзвешенного мнения его друзей:

$$\dot{\boldsymbol{x}} = \boldsymbol{\beta} \odot (\boldsymbol{w}(\boldsymbol{x}) - \boldsymbol{x}).$$

Здесь $\boldsymbol{\beta} \in \mathbb{R}^N$ характеризует направленность и скорость изменения мнения под влиянием друзей. Используя $\mathbf{A}(\mathbf{c} \odot \boldsymbol{x}(t)) = (\mathbf{A} \odot \mathbf{c}^\top)\boldsymbol{x}(t)$, получим:

$$\mathbf{W}(\boldsymbol{\beta}) = \boldsymbol{\beta} \odot \left(\frac{\mathbf{A} \odot \mathbf{c}^\top}{\mathbf{A}\mathbf{c}} - \mathbf{I} \right). \tag{2}$$

Таким образом, в выбранной модели есть всего N неизвестных параметров $\boldsymbol{\beta} \in \mathbb{R}^N$. Эти параметры можно выбрать, решая следующую оптимизационную задачу:

$$f(\boldsymbol{\beta}) = \|\boldsymbol{x}(T_2) - e^{\mathbf{W}(\boldsymbol{\beta})(T_2 - T_1)}\boldsymbol{x}(T_1)\|^2 \to \min_{\boldsymbol{\beta}}.$$

Суть данной оптимизации заключается в подборе таких параметров $\boldsymbol{\beta}$, чтобы решение уравнения (1) с $\mathbf{W} = \mathbf{W}(\beta)$ проходило в среднем как можно точнее через значение $\boldsymbol{x}(T_2)$.

Отметим, что градиент функции f имеет следующий вид:

$$\nabla f(\boldsymbol{\beta}) = -2(T_2 - T_1)\,(\boldsymbol{w}(\boldsymbol{x}(T_1)) - \boldsymbol{x}(T_1)) \\ \odot \left(e^{\mathbf{W}(\boldsymbol{\beta})(T_2 - T_1)}(\boldsymbol{x}(T_2) - e^{\mathbf{W}(\boldsymbol{\beta})(T_2 - T_1)}\boldsymbol{x}(T_1)) \right).$$

С вычислительной точки зрения $e^{\mathbf{W}(\boldsymbol{\beta})(T_2 - T_1)}\boldsymbol{x}(t_1)$ легче всего получить, интегрируя систему (1) от T_1 до T_2 с начальным условием $\boldsymbol{x}(T_1)$ методом Рунге – Кутты. Для вычисления $\nabla f(\boldsymbol{\beta})$ понадобится повторное интегрирование системы (1) с начальным условием $\boldsymbol{x}(T_2) - e^{\mathbf{W}(\boldsymbol{\beta})(T_2 - T_1)}\boldsymbol{x}(T_1)$.

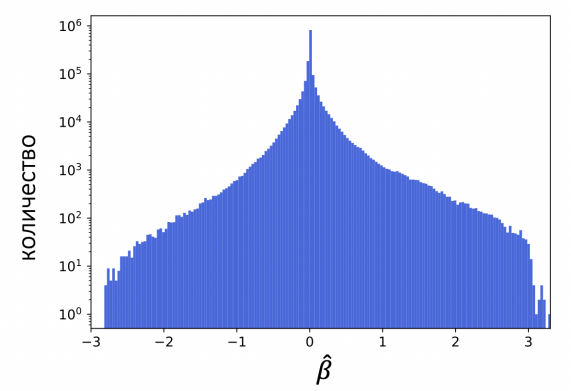

Рис. 1. Гистограмма распределения компонент вектора $\hat{\boldsymbol{\beta}}$. Вертикальная ось имеет логарифмический масштаб

Оптимизацию будем производить методом градиентного спуска:

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n - \lambda_n \nabla f(\boldsymbol{\beta}_n),$$

где $\boldsymbol{\beta}_0$, λ_0 начальные параметры. В ходе численной оптимизации было выбрано $\boldsymbol{\beta}_0 = \mathbf{0}$, а значения шага λ_n перестраивались вручную, если наблюдалось резкое повышение значения $f(\boldsymbol{\beta}_{n+1})$. Также, если шаг метода становился небольшим, то большие по модулю компоненты вектора $\boldsymbol{\beta}_n$ обнулялись для выхода из локальных минимумов. С помощью данной ручной настройки удалось подобрать такое $\hat{\boldsymbol{\beta}}$ (см. рис. 1), что $f(\hat{\boldsymbol{\beta}}) \approx 609$. Для сравнения:

$$f(\mathbf{0}) = \|\boldsymbol{x}(T_2) - \boldsymbol{x}(T_1)\|^2 \approx 2761.$$

Финальный прогноз мнений каждого из пользователей строился следующим образом:

$$\hat{\boldsymbol{x}}(T_3) = e^{\mathbf{W}(\hat{\boldsymbol{\beta}})(T_3 - T_2)} \boldsymbol{x}(T_2).$$

После этого был найден вектор $\boldsymbol{y}_{BP}$, для которого величина AE составила

$$AE(\boldsymbol{y}_{BP}) \approx 142 \times 10^{-4},$$

что превышает величину ошибки прогноза модели постоянного тренда $\boldsymbol{y}_{const}$.

Заметим, что прогноз, в котором идентификация модели Абельсона по первым двум снимкам мнений производится без использования центральности по собственному вектору (иными словами, когда вектор $\mathbf{c}$ в формуле (2) является единичным), оказался менее точен, чем с использованием центральности по собственному вектору (200×10^{-4} против 142×10^{-4}). Также необходимо отметить, что без использования центральности по собственному вектору модель Абельсона «хуже» объясняет данные при решении обратной задачи по первым двум снимкам мнений – значение оптимизируемой функции f составило примерно 750, а с использованием центральности по собственному вектору, как было отмечено выше, – 609.

Тот факт, что описанная выше модель и модель постоянного тренда построены «независимо», можно использовать для улучшения прогноза, формируя на их основании ансамбль моделей. Для этого нужно усреднить предсказания по моделям, используя веса $\alpha_{BP} = 1/AE(\boldsymbol{y}_{BP})$ (для описанной в данном разделе модели[1]), $\alpha_{const} = 1/AE(\boldsymbol{y}_{const})$ (для модели постоянного тренда[2]):

$$\hat{y}_{BP}^{\Delta x} = \left(\alpha_{const} \cdot y_{const}^{\Delta x} + \alpha_{BP} \cdot \frac{y_{BP}^{\Delta x}}{\sum_k y_{BP}^{\Delta x_k}} \right) / (\alpha_{const} + \alpha_{BP}).$$

Здесь также использована перенормировка $y_{BP}^{\Delta x}$ на $\sum_k y_{BP}^{\Delta x_k}$, так как некоторые компоненты вектора $\hat{\boldsymbol{x}}(t_3)$ могли выйти за пределы интервала $[0, 1]$. Расчеты показывают, что

$$AE(\hat{\boldsymbol{y}}_{BP}) \approx 27 \times 10^{-4},$$

т.е. прогноз существенно улучшился и в сравнении с моделью постоянного тренда, и в сравнении со SCARDO-моделью.

5.2. Решение 2 (Семен Красоткин)

Для прогнозирования распределения мнений пользователей в декабре 2018 года (вектора $\boldsymbol{y}(T_3)$) была использована модель

[1] *По правилам Хакатона, участники могли узнать точность прогноза только в момент оглашения итоговых результатов. В связи с этим для построения оценки точности данного прогноза в качестве правильного ответа был взят прогноз модели постоянного тренда.*

[2] *Точность данного прогноза была оглашена заранее на установочной лекции.*

линейной регрессии. В основе такого выбора лежало предположение о том, что краткосрочные прогнозы можно приближать линейными соотношениями. Кроме того, выбор данной модели обусловлен простотой реализации и малым использованием вычислительных ресурсов. Также предполагалось, что линейная регрессия послужит ориентиром для других решений.

В качестве регрессоров выступали мнение пользователя и мнения его друзей:

$$x_i(T_{k+1}) = \sum_{j \in V_i} \frac{\beta_1}{\operatorname{card} V_i} x_j(T_k) + \beta_2 x_i(T_k) + \beta_3, \tag{3}$$

а коэффициенты регрессии β_1, β_2 и β_3 определялись методом наименьших квадратов. При этом нормировочный множитель $1/\operatorname{card} V_i$ в первом слагаемом правой части позволял учесть тот факт, что пользователи могут иметь разное число друзей: распределение степеней вершин в рассматриваемой социальной сети, как и для большинства такого рода объектов, является степенным [18]. Фактически модель (3) является расширением модели ДеГроота [8] путем включения свободного слагаемого и снятия ограничений на коэффициенты, но с дополнительным ограничением о гомогенности весов влияний.

Модель (3) обучалась на всей совокупности данных и после этого использовалась для прогноза $x_i(T_3)$:

$$x_i(T_3) = \sum_{j \in V_i} \frac{\beta_1}{\operatorname{card} V_i} x_j(T_2) + \beta_2 x_i(T_2) + \beta_3$$

после чего производилось агрегирование по десяти бинам.

Для построенного прогноза $\boldsymbol{y}_K(T_3)$ точность составила

$$AE(\boldsymbol{y}_K(T_3)) \approx 171 \times 10^{-4}.$$

5.3. Решение 3 (Максим Рыжов)

В основу предлагаемого решения легла классическая линейная модель ДеГроота [8]. Предположим, что мнение $x_i(t)$ узла i зависит линейно от мнения связанных с ним узлов и его собственного мнения на момент времени $t-1$:

$$x_i(t) = \sum_{j \in V_i} \beta_{i,j} x_j(t-1) + \beta_{i,i} x_i(t-1). \tag{4}$$

В выражении (4) на коэффициенты $\beta_{i,j}$ не накладываются никакие ограничения (в отличие от модели ДеГроота, в которой $\beta_{i,j} \geqslant 0$ и $\sum_{j \in V_i} \beta_{i,j} = 1$ для любого $i \in \{1, \ldots, N\}$).

При помощи первых двух снимков мнений и матрицы смежности для каждого пользователя на основе метода наименьших квадратов были найдены коэффициенты $\beta_{i,j}$. После этого при помощи идентифицированной модели (4) были рассчитаны значения $x_i(T_3)$:

$$x_i(T_3) = \sum_{j \in V_i} \beta_{i,j} x_j(T_2) + \beta_{i,i} x_i(T_2).$$

Далее было произведено агрегирование согласно обозначенным в постановке задачи бинам.

Для построенного прогноза $\boldsymbol{y}_R(T_3)$ точность составила

$$AE(\boldsymbol{y}_R(T_3)) \approx 233 \times 10^{-4}.$$

6. Обсуждение результатов участников и выводы

В таблице 1 приведены результаты команд. Все три решения были основаны на построении прогноза изменения мнения каждого пользователя с последующим агрегированием (англ. bottom-up design), а не на использовании каких-либо макро-моделей в духе SIR-модели [15]. При этом все три решения, предложенные участниками, были основаны на линейных моделях. Решение 1 – для случая непрерывного времени, Решения 2 и 3 – для дискретного.

В рамках Решения 1 каждый пользователь-наблюдение характеризовался одним неизвестным параметром, погруженным в контекст модели Абельсона. Гетерогенность влиятельностей агентов была достигнута путем вспомогательного вычисления структурной влиятельности, определяемой центральностью по собственному вектору. В Решении 2 предполагалось, что влияние со стороны окружения гомогенно, но отличается от влияния со стороны собственного мнения. Также в этом решении предполагалось наличие свободного слагаемого, что потенциально повышало предиктивную силу модели. В рамках Решения 3 количество параметров модели для каждого пользователя определялось

числом его друзей, что делало модель склонной к переобучению. По-видимому, именно это и случилось в итоге.

Напротив, прогностическая модель с наименьшим числом параметров, учитывающая структурную влиятельность вершин (Решение 1), показала наилучший результат, если рассматривать одиночные модели (142×10^{-4}). Однако этот прогноз был менее точен, чем полученный при помощи SCARDO-модели в работе [1]. В дальнейшем использование базовой модели из Решения 1 в ансамбле с моделью постоянного тренда позволило превзойти результат SCARDO-модели (27×10^{-4} против 42×10^{-4} для SCARDO-модели). Для этого прогнозы базовой модели из Решения 1 и модели постоянного тренда были взвешены, а веса определялись точностью их прогнозов по отдельности. В то время как точность прогноза модели постоянного тренда была известна участникам Хакатона заранее, точность базовой модели из Решения 1 оставалась неизвестной. Для решения данной проблемы была применена следующая эвристика: оценка точности модели путем сравнения ее не с эталонным значением, а с искаженным, а именно – с прогнозом модели постоянного тренда, который было легко воспроизвести.

Таблица 1. Точность прогнозов моделей

Команда	Показатель (модель)
Максим Бузиков & Юлия Петелина	200×10^{-4} (модель Абельсона) 142×10^{-4} (модель Абельсона + центр. по собств. вект.) 27×10^{-4} (ансамбль моделей)
Семен Красоткин	171×10^{-4} (линейная модель с гомогенными по популяции коэффициентами и свободным членом)
Максим Рыжов	233×10^{-4} (линейная модель с гетерогенными по популяции коэффициентами и без свободного члена)

Какие выводы можно сделать по итогам Хакатона? Во-первых, «модельный» подход оказался наиболее эффективным: в то время как Решения 2 и 3 были основаны на простейших спецификациях линейной модели без каких-либо априорных предположений о механизмах влияния (в том смысле, что на коэффициенты моделей не накладывались никакие содержательные

ограничения), в основе Решения 1 лежала модель Абельсона [2] – аналог модели ДеГроота [8] в непрерывном времени, усиленная при помощи одного из классических приемов сетевого анализа – оценки влиятельности вершин сети при помощи Фробениусова вектора. При этом использование центральности по собственному вектору действительно способствовало росту предиктивной силы модели, а также ее способности описывать имеющиеся в распоряжении данные. Без этого модель Абельсона (с гомогенными весами влияний) с точки зрения точности может уступать линейным моделям, лишенным априорной информации о процессе. Открытым остается вопрос, сыграл ли свою роль тот факт, что модель, предложенная в рамках Решения 1 была основана на непрерывном времени, в отличие от Решений 2 и 3, оперирующих дискретным временем.

Второй вывод заключается в том, что применение техник машинного обучения, в частности ансамбля моделей, а также дополнительных эвристик позволяет существенно повысить точность прогноза. Скомбинировав две независимые модели – модель Абельсона и модель постоянного тренда – Максиму Бузикову и Юлии Петелиной удалось значительно улучшить свой прогноз, обойдя оба ориентира (со 142×10^{-4} до 27×10^{-4}). Необходимо учитывать, что второй из указанных ориентиров был получен при помощи SCARDO-модели, которая является нелинейной. Иными словами, ансамбль из двух линейных моделей оказался более успешен с прогностической точки зрения, чем нелинейная модель.

В дальнейшем интересным было бы рассмотреть альтернативные, нелинейные модели социального влияния, основанные, к примеру, на механизме ограниченного доверия [7]. Также перспективным видится тестирование иных метрик центральности [5]. Как было отмечено в работе [1], рассматриваемый набор данных, несмотря на свои внушительные размеры, характеризуется относительно невысоким уровнем изменений состояния общественного мнения. В связи с этим интерес представляет изучение аналогичных постановок на лонгитюдных данных высокой гра-

нулярности, собранных в рамках более длительных временных промежутков, состоящих из более чем трех волн. В этом случае в прогностические модели можно попытаться интегрировать эффект памяти [14] и оценить его влияние на точность прогноза.

Литература

1. КОЗИЦИН И.В. *Построение прогноза динамики общественного мнения при помощи SCARDO-модели* // Управление большими системами. – 2024. – Вып. 108. – С. 124–136. – DOI: https://doi.org/10.25728/ubs.2024.108.7.
2. ABELSON R.P. *Mathematical models of the distribution of attitudes under controversy* // Contributions to Mathematical Psychology. – 1964.
3. BONACICH P. *Some unique properties of eigenvector centrality* // Social Networks. – 2007. – Vol. 29, No. 4. – P. 555–564. – DOI: https://doi.org/10.1016/j.socnet.2007.04.002.
4. BOND R.M. et al. *A 61-million-person experiment in social influence and political mobilization* // Nature. – 2012. – Vol. 489, No. 7415. – P. 295–298. – DOI: https://doi.org/10.1038/nature11421.
5. CHEBOTAREV P., GUBANOV D.A. *How to Choose the Most Appropriate Centrality Measure? A Decision-Tree Approach* // IEEE Trans. on Systems, Man, and Cybernetics: Systems. – 2024. – DOI: https://doi.org/10.1109/TSMC.2024.3510633.
6. CINUS F. et al. *The effect of people recommenders on echo chambers and polarization* // Proc. of the Int. AAAI Conf. on Web and Social Media. – 2022. – Vol. 16. – P. 90–101. – DOI: https://doi.org/10.1609/icwsm.v16i1.19275.
7. DEFFUANT G. et al. *Mixing beliefs among interacting agents* // Advances in Complex Systems. – 2000. – Vol. 3, No. 01n04. – P. 87–98. – DOI: https://doi.org/10.1142/S0219525900000078.
8. DEGROOT M.H. *Reaching a consensus* // Journal of the American Statistical Association. – 1974. – Vol. 69, No. 345. – P. 118–121. – DOI: https://doi.org/10.1080/01621459.1974.10480137.

9. FLACHE A. et al. *Models of social influence: Towards the next frontiers* // JASSS – The Journal of Artificial Societies and Social Simulation. – 2017. – Vol. 20, No. 4. – P. 2. – DOI: https://doi.org/10.18564/jasss.3521.
10. FRENCH JR J.R.P. *A formal theory of social power* // Psychological Review. – 1956. – Vol. 63, No. 3. – P. 181.
11. FREY V. et al. *Who influences lower-status individuals more: People of higher-status outgroups or people of their lower-status ingroup? Examining the difference between matters of opinion and matters of fact* // Social Science Research. – 2024. – Vol. 123. – P. 103060. – DOI: https://doi.org/10.1016/j.ssresearch.2024.103060.
12. FRIEDKIN N.E., JIA P., BULLO F. *A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences* // Sociological Science. – 2016. – Vol. 3. – P. 444–472. – DOI: https://doi.org/10.15195/v3.a20.
13. GEZHA V.N., KOZITSIN I.V. *The effects of individuals' opinion and non-opinion characteristics on the Organization of Influence Networks in the online domain* // Computers. – 2023. – Vol. 12, No. 6. – P. 116. – DOI: https://doi.org/10.3390/computers12060116.
14. JEDRZEJEWSKI A., SZNAJD-WERON K. *Impact of memory on opinion dynamics* // Physica A: Statistical Mechanics and its Applications. – 2018. – Vol. 505. – P. 306–315. – DOI: https://doi.org/10.1016/j.physa.2018.03.077.
15. JI C., JIANG D. *Threshold behaviour of a stochastic SIR model* // Applied Mathematical Modelling. – 2014. – Vol. 38, No. 21–22. – P. 5067–5079. – DOI: https://doi.org/10.1016/j.apm.2014.03.037.
16. KOZITSIN I.V. et al.*Modeling political preferences of russian users exemplified by the social network Vkontakte* // Mathematical Models and Computer Simulations. – 2020. – Vol. 12. – P. 185–194. – DOI: https://doi.org/10.1134/S2070048220020088.
17. KOZITSIN I.V. *A general framework to link theory and*

empirics in opinion formation models // Scientific reports. – 2022. – Vol. 12, No. 1. – P. 5543. – DOI: https://doi.org/10.1038/s41598-022-09468-3.

18. KOZITSIN I.V. *Opinion dynamics of online social network users: a micro-level analysis* // The Journal of Mathematical Sociology. – 2023. – Vol. 47, No. 1. – P. 1–41. – DOI: https://doi.org/10.1080/0022250X.2021.1956917.
19. KULKARNI B. et al. *SLANT+: A nonlinear model for opinion dynamics in social networks* // IEEE Int. Conf. on Data Mining (ICDM). – IEEE, 2017. – P. 931–936. – DOI: https://doi.org/10.1109/ICDM.2017.117.
20. NEWMAN M. *Networks*. – Oxford University Press, 2018.
21. OKAWA M., IWATA T. *Predicting opinion dynamics via sociologically-informed neural networks* // Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. – 2022. – P. 1306–1316. – DOI: https://doi.org/10.1145/3534678.35392.
22. PANSANELLA V. et al. *Change my mind: Data driven estimate of open-mindedness from political discussions* // Int. Conf. on Complex Networks and Their Applications. – Cham: Springer Int. Publishing, 2022. – P. 86–97. – DOI: https://doi.org/10.1007/978-3-031-21127-0_8.
23. PESHKOVSKAYA A., BABKINA T., MYAGKOV M. *Social context reveals gender differences in cooperative behavior* // Journal of Bioeconomics. – 2018. – Vol. 20. – P. 213–225. – DOI: https://doi.org/10.1007/s10818-018-9271-5.
24. PESHKOVSKAYA A., BABKINA T., MYAGKOV M. *Gender effects and cooperation in collective action: A laboratory experiment* // Rationality and Society. – 2019. – Vol. 31, No. 3. – P. 337–353. – DOI: https://doi.org/10.1177/1043463119858788.
25. PROSKURNIKOV A.V., TEMPO R. *A tutorial on modeling and analysis of dynamic social networks. Part I* // Annual Reviews in Control. – 2017. – Vol. 43. – P. 65–79. – DOI: https://doi.org/10.1016/j.arcontrol.2017.03.002.
26. PROSKURNIKOV A.V., TEMPO R. *A tutorial on modeling*

and analysis of dynamic social networks. Part II // Annual Reviews in Control. – 2018. – Vol. 45. – P. 166–190. – DOI: https://doi.org/10.1016/j.arcontrol.2018.03.005.

27. RAVAZZI C. et al. *Learning hidden influences in large-scale dynamical social networks: A data-driven sparsity-based approach, in memory of Roberto Tempo* // IEEE Control Systems Magazine. – 2021. – Vol. 41, No. 5. – P. 61–103. – DOI: https://doi.org/10.1109/MCS.2021.3092810.
28. ZHU L., HE Y., ZHOU D. *Neural opinion dynamics model for the prediction of user-level stance dynamics* // Information Processing & Management. – 2020. – Vol. 57, No. 2. – P. 102031. – DOI: https://doi.org/10.1016/j.ipm.2019.03.010.

FORECASTING THE DYNAMICS OF PUBLIC OPINION BASED ON LONGITUDINAL DATA OF HIGH GRANULARITY: THE ABELSON MODEL, REGRESSION MODELS, AND ENSEMBLES OF MODELS

Maksim Buzikov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Cand.Sc., Senior Researcher (me.buzikov@physics.msu.ru).
Iuliia Petelina, Ozon Tech, Moscow, Team Leader (ptlna@yandex.ru).
Semen Krassotkin, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Mathematician (semen.krassotkin@gmail.com).
Maksim Ryzhov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, researche (ryzhov@phystech.edu).
Ivan Kozitsin, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, PhD, senior researcher, Moscow Institute of Physics and Technology, Moscow, assistant professor (kozitsin.ivan@mail.ru).

Abstract: We consdier the problem of forecasting the dynamics of public opinion based on longitudinal data of high granularity gleaned from the social network VKontakte. This problem was suggested to the participants of the Hackathon «UBS Challenge'2024» as one of the leisure events of the XX All-Russian School-Conference of Young Scientists «Management of Large Systems» (UBS), held in Novocherkassk in 2024. This paper is devoted to a detailed description of the Hackathon and the solutions proposed by its participants. For a sample of $N = 1\,648\,829$ *users, based on two granular snapshots of their opinions taken six months apart (in February and July 2018), participants have to elaborate on a forecast of the distribution of public opinion in December 2018. The participants also had the information about the structure of friendship ties between users. We report that the highest accuracy was achieved by an ensemble of two models – the Abelson model, enhanced by estimating users' social power via the eigenvector centrality measure, and the constant trend model.*

Keywords: social influence models, artificial intelligence, social networks, eigenvector centrality, ensembles of models.

УДК 519.7
ББК 22.18

Статья представлена к публикации
членом редакционной коллегии А.Г. Чхартишвили.

Поступила в редакцию 28.01.2025.
Дата опубликования 31.05.2025.