DOI: 10.25728/mlsd.2022.1470

# ВЛИЯНИЕ ТРОЛЛЕЙ НА УРОВЕНЬ АГРЕССИИ В СОЦИАЛЬНЫХ СЕТЯХ: АГЕНТНО-ОРИЕНТИРОВАННАЯ МОДЕЛЬ[1]

**Стукал Д.К.**
*Национальный исследовательский университет «Высшая школа экономики»,*
*Россия, г. Москва, ул. Мясницкая, д.20*
dstukal@hse.ru

*Аннотация: На основе агентно-ориентированной модели исследуется влияние троллей на уровень агрессии в социальной сети при наличии и отсутствии способности пользователей идентифицировать троллей. Модель исследуется в ходе вычислительных экспериментов. Калибровка модели на данных из сети ВКонтакте позволяет выявить величину вероятного эффекта.*

Ключевые слова: тролли, агрессия, вычислительная модель, ВКонтакте.

## Введение

Исследования в области онлайн коммуникации обращают все большее внимание на активность особой группы пользователей, остававшихся на протяжении длительного времени в тени исследовательского интереса, - неаутентичных пользователей, онлайн-персона которых не совпадает с личностью человека, ведущего аккаунт социальной сети. Помимо обычных пользователей, желающих по определенным причинам скрыть свою персональную информацию, к числу таких пользователей относятся также т.н. боты и тролли. Концептуализация и исследования первой из этих групп породили достаточно большую научную литературу в области компьютерных и вычислительных социальных наук. В рамках существующей литературы боты определяются как автоматизированные аккаунты, управляемые программным кодом, имитирующим поведение обычных пользователей. С учетом такого определения, возникла большая научная литература, посвященная детекции автоматизированных аккаунтов [1-4], исследованию стратегий их использования в разных странах мира [5-6], измерению эффектов офлайн и онлайн среды на активность этих ботов [7], а также в определенной степени измерению эффектов самих ботов на коммуникацию в социальных медиа [8-9]. В то же время исследования троллинга, с одной стороны, менее многочисленны, а с другой – сталкиваются с существенными методологическими ограничениями, обусловленными неоднозначностью этого понятия. Результатом сложившейся ситуации оказывается как преобладание не претендующих на обобщение кейсовых описаний, так и острый дефицит работ, предлагающих количественную оценку степени влияния троллей на различные аспекты коммуникации в социальных медиа. В данной работе предпринимается попытка предложить модельное решение для измерения влиянию троллинга на уровень агрессии в публикациях – важную характеристику коммуникации в онлайн среде.

Решение этой задачи требует, прежде всего, концептуализации понятия троллинга. Один из существующих подходов получил широкое распространение в американском политологическом сообществе после президентских выборов 2016 г., в которых – по заверениям местных средств массовой информации – высокую политическую активность проявили пользователи, получавшие финансовое вознаграждение за свою деятельность [10]. Именно таких пользователей в американской политологии принято именовать троллями [11]. При этом исследование определенных таким образом троллей со всей очевидностью сталкивается с серьезными ограничениями, связанными с необходимостью детекции троллинга, зачастую опирающейся в эмпирических исследованиях либо на утечки в данных, либо на ручную разметку исследователей в соответствии с авторскими методиками.

Альтернативные подходы к определению троллинга акцентируют внимание на некоторых наблюдаемых характеристиках поведения аккаунтов социальных медиа и маркируют эти аккаунты как троллей в случае наличия таких признаков. К числу таких признаков относится распространение спама [12], шаблонность ответов или участие в скоординированном распространении определенных сообщений [13], высмеивание других пользователей [12] или выражение агрессии в отношении собеседников [14]. В данной работе мы опираемся на определение троллинга в соответствии с последним из указанных признаков. Это обусловлено несколькими причинами. Во-первых, первые два признака не позволяют разграничить понятия троллей и ботов, поскольку – как показано в

[1] *Исследование выполнено за счет гранта Российского научного фонда № 21-78-00079, https://rscf.ru/project/21-78-00079/*

предшествующих исследованиях – автоматизированные аккаунты широко вовлечены в распространение спама и публикацию однотипных, шаблонных сообщений [15]. Во-вторых, высмеивание других пользователей в публикациях или комментариях превращает понятие тролля в чрезвычайно широкую группу пользователей, в которую с высокой вероятностью в тот или иной момент может попасть любой пользователь социальной сети. По этим причинам мы выделяем агрессию в качестве определяющего признака троллинга.

В какой мере высказываемая троллями агрессия способна оказать влияние на содержание онлайн коммуникации? Эмпирически обоснованный ответ на этот вопрос затрудняется отсутствием экспериментальных данных для оценки каузального эффекта троллей, что в значительной степени связано как с техническими, так и этическими сложностями проведения онлайн экспериментов с использованием троллинга. Этим обусловливается акцент исследователей на изучении отдельных кейсов с использованием качественной методологии, позволяющей формулировать содержательные утверждения, но ограничивающей авторов в получении количественных оценок каузального эффекта.

В данной работе предпринимается попытка предложить количественную оценку влияния троллей на уровень агрессии в социальной сети путем разработки вычислительной модели, допускающей эмпирическую калибровку на основе неэкспериментальных данных. Мы проводим серию вычислительных экспериментов, исходя из различных допущений о распределении свойств агентов. Далее гиперпараметры модели калибруются на основе эмпирических данных из социальной сети ВКонтакте, что позволяет нам предложить количественную оценку эффекта троллей на уровень агрессии с опорой на вычислительную модель.

## 1 Модель

Предлагаемая вычислительная модель воспроизводит основные черты коммуникации пользователей социальной сети через моделирование принятия решения о том, какого рода комментарии опубликовать под некоторым сообщением. Мы предполагаем существование двух групп агентов: обычных пользователей и троллей. Публикуемые комментарии могут либо содержать, либо не содержать агрессию. В соответствии с используемым в работе определением, тролли всегда публикуют агрессивные комментарии. Агрессивность комментариев обычных пользователей зависит от определенных условий.

Помимо типа (обычный пользователь или тролль), агенты наделены также позицией по обсуждаемому вопросу, условно именуемой нами идеологией. Для простоты презентации и удобства будущей калибровки модели идеология агента моделируется как дискретная случайная величина, распределенная от -2 до 2. Идеологическая позиция троллей предполагается нецентристской и принимает крайние значения от -2 до 2.

Третья характеристика агентов в предлагаемой модели – индивидуальный уровень толерантности к наблюдаемому уровню агрессии, именуемый в дальнейшем порогом толерантности. Мы предполагаем, что готовность обычного пользователя опубликовать агрессивный комментарий зависит от соотношения между удельным весом агрессии в объеме прочитанных сообщений, с одной стороны, и порогом толерантности с другой. Если доля агрессивных комментариев под прочитанным постом превышает индивидуальный порог толерантности пользователя, данный пользователь опубликует агрессивный комментарий; в противном случае пользователь публикует комментарий без агрессии. При этом мы исходим из того, что вклад разных прочитанных комментариев в накопленный пользователем уровень агрессии различается и зависит от того, насколько идеологическая позиция их авторов далека от идеологической позиции самого пользователя. В модели, таким образом, предполагается, что индивидуальный уровень агрессии возрастает в наибольшей степени при прочтении агрессивных комментариев с противоположной точной зрения; при этом агрессивные комментарии с идентичной точкой зрения могут вносить лишь небольшой вклад в индивидуальный уровень агрессии.

С учетом предлагаемого подхода к моделированию принятия решения о публикации агрессивного комментария, индивидуальные пороги толерантности моделируются как случайные величины, непрерывно распределенные от 0 до 1 в соответствии с некоторым распределением из бета-семейства. Порог толерантности для троллей, в соответствии с определением, полагается равным 0.

Наконец последняя, четвертая характеристика пользователей, – это способность к детекции троллей. Детекция троллей в данном случае понимается как выявление пользователем того, что перед ним особый тип агента, целенаправленно публикующий агрессивные комментарии. В случае детекции, агрессивное сообщение, опубликованное троллем, не вносит вклада в накопленный индивидом уровень агрессии. Иными словами, мы предполагаем, что пользователи, обладающие

навыком выявлять троллей, могут быть более устойчивы к их воздействию. Вопрос о степени устойчивости к такому воздействию может указать модель.

## 2 Результаты вычислительных экспериментов

Предложенная модель реализована в вычислительной среде R, в которой осуществлены 50 000 прогонов модели с различными наборами гиперпараметров. В ходе экспериментов мы рассматривали четыре гиперпараметра модели: долю троллей в популяции пользователей, два параметра бета-распределения порогов толерантности к агрессии и наличие/отсутствие способности к детекции ботов. Доля троллей менялась в диапазоне от 1% до 30%. Первый параметр бета-распределения был зафиксирован на значениях 1 и 2, в то время как второй менялся от 1 до 20. Такой набор параметров бета-распределения позволял рассмотреть случаи треугольного, параболического и скошенного вправо (с разной степенью скошенности) распределения.

В ходе экспериментов для каждого набора гиперпараметров осуществлялось генерирование 10 сообщений, к каждому из которых агенты модели могли публиковать комментарии. Общее число комментариев к модели определялось как пуассоновская случайная величина с математическим ожиданием 15, что соответствует среднему числу комментариев под публикациями в массиве данных, описываемом ниже.

На основе сгенерированных публикаций и комментариев к ним создавался массив данных для оценивания линейной МНК-регрессии, зависимой переменной в которой выступала доля агрессивных комментариев, опубликованных обычными пользователями, а объясняющей переменной – доля троллей в популяции агентов. В соответствии с логикой модели, мы ожидаем, что коэффициент при объясняющей переменной будет положительным, поскольку он отражает, как активность троллей способствует росту агрессии в комментариях пользователей под постами в социальных сетях. Ключевые вопросы, на которые мы стремимся получить ответы в ходе вычислительных экспериментов – это вопросы о зависимости коэффициента МНК-регрессии от 1) распределения индивидуальных порогов толерантности к агрессии и 2) способности пользователей к детекции троллей.

На рисунках 1 и 2 показана зависимость регрессионного коэффициента от второго параметра бета-распределения порогов толерантности (для первого параметра, равного 1 и 2 соответственно). При более высоких значения второго параметра математическое ожидание распределения оказывается ниже; соответственно, плотность распределения на малых значениях порогов толерантности оказывается выше. Иными словами, мы ожидаем, что при более высоких значениях второго параметра пользователи будут менее толерантны к проявлениям агрессии, а тролли будут иметь больше возможностей для провоцирования пользователей на агрессию. Таким образом, наблюдаемая закономерность (малым значениям второго параметра соответствуют меньшие значения регрессионных коэффициентов) представляется естественной.

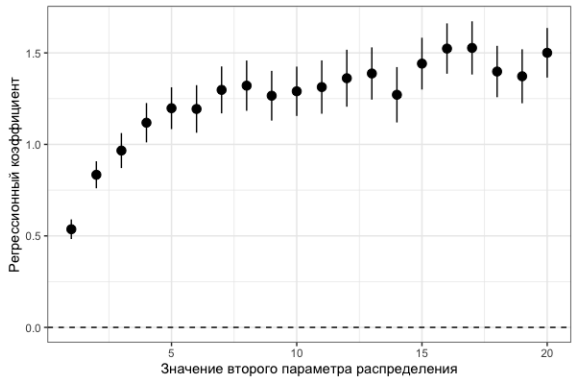


*Рис. 1. Зависимость коэффициентов регрессии от второго параметра бета-распределения ($\alpha = 2$, без детекции)*

При этом обращает на себя внимание тот факт, что регрессионный коэффициент стабилизируется в районе 1.2 для высоких значений второго параметра распределения. Иными словами, для широкого круга распределений индивидуальных порогов толерантности, рост доли троллей в популяции агентов на пять процентных пунктов приводит к росту доли агрессивных сообщений, публикуемых обычными пользователями, на шесть процентных пунктов. При этом величина влияния троллей на долю агрессивных комментариев падает до 0.5 для распределений с малыми значениями второго параметра (распределений без существенного скоса вправо).

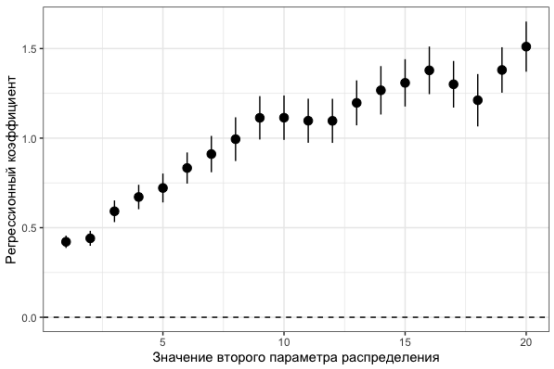


*Рис. 2. Зависимость коэффициентов регрессии от второго параметра бета-распределения (α = 1, без детекции)*

Рисунок 3 иллюстрирует ту же закономерность в ситуации, когда пользователи обладают способностью к детекции троллей. Как видно на графике, общее направление взаимосвязи остается прежним: рост значения второго параметра бета-распределения сопутствует росту регрессионного коэффициента. При этом скорость изменения регрессионного коэффициента при изменении второго параметра бета-распределения существенно снижается, что отражает способность пользователя выявить троллей с противоположной позицией и проигнорировать их эффект.

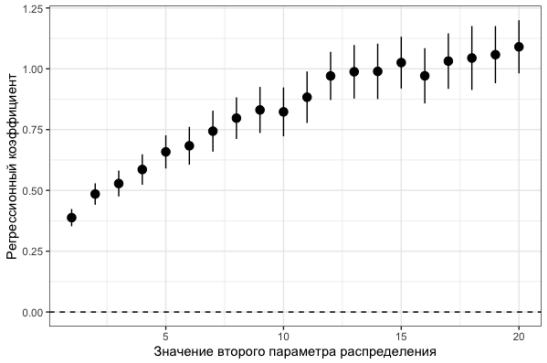


*Рис. 3. Зависимость коэффициентов регрессии от второго параметра бета-распределения (α = 1, с детекцией)*

Какие из переведенных выше графиков в наибольшей степени отражают эмпирические закономерности? Для ответа на этот вопрос и калибровки параметров модели были собраны и закодированы 2016 комментариев к случайной выборке публикаций в социальной сети ВКонтакте в 2020-2021 гг.

## 3 Эмпирическая калибровка модели

На основе набора ключевых слов о политической повестке в России в 2020-2021 гг. с помощью интерфейса разработки приложений (API) социальной сети ВКонтакте было собрано 1102 публикаций и 11582 сделанных к ним комментариев. Далее случайным образом были выбраны 50 публикаций, для которых были выгружены все связанные с ними 2016 комментариев, сделанных 832 различными пользователями. Выгруженные комментарии были использованы для ручной разметки на предмет содержания в них признаков троллинга. Мы фиксировали все указанные выше четыре признака троллинга: публикацию спама, публикацию шаблонных сообщений, высмеивание других пользователей, а также выражение агрессии. Фиксация всех указанных признаков в ходе ручной разметки позволила оценить ограничения нашей модели, опирающейся лишь на один из фигурирующих в литературе признаков – агрессию.

Поскольку выражение агрессии не является исключительной характеристикой троллей (обычный пользователь также может высказаться агрессивно – например, под влиянием активности троллей, как это предполагается в нашей модели), важная эмпирическая задача состояла в определении оснований для детекции троллей в перечне из 832 пользователей. Отталкиваясь от эмпирического

распределения числа агрессивных комментариев, сделанных пользователями в собранном массиве данных, мы определили в качестве троллей тех, кто опубликовал не менее 3 агрессивных комментариев. При таком определении доля троллей составляет в собранном массиве 3% (28 из 832 автором комментариев). Как рост этой доли мог бы сказаться на доле агрессивных комментариев?

Для калибровки разработанной вычислительной модели мы для автора каждого комментария рассчитали долю агрессивных сообщений в общем числе комментариев, прочитанных этим пользователем под рассматриваемым постом. Данный показатель выступает эмпирическим аналогом индивидуальных порогов толерантности к агрессии в нашей модели: действительно, данный показатель позволяет на основе собранных данных рассчитать, при какой доле агрессивных комментариев пользователи часто публиковали агрессивные комментарии. Рассчитанный таким образом показатель использоваться для оценивания параметров бета-распределения с помощью метода максимального правдоподобия (ММП). Полученные ММП-оценки составили $\alpha = 0.83$ и $\beta = 2.766$ для первого и второго параметров распределения соответственно.

Оценка бета-распределения порогов толерантности к агрессии позволяют калибровать предложенную вычислительную модель и указать на значение 0.5 в качестве наиболее вероятного регрессионного коэффициента, связывающего долю агрессивных комментариев с долей троллей в популяции пользователей социальной сети. Эмпирическая верификация данного результата выходит за рамки данной работы и будет являться предметом дальнейших исследований, поскольку требует уточнения методики детекции троллей в эмпирических данных.

## Заключение

Академические исследования коммуникации в социальных сетях обращают все большее внимание на активность и значение неаутентичных пользователей, заметное место среди которых занимают тролли. При этом недостаточно исследованным остается вопрос о степени влияния троллей на содержание и иные характеристики коммуникации в социальных сетях. В данной работе предпринимается попытка сделать шаг в направлении оценивания влияния троллей на уровень агрессии в публикуемых комментариях путем разработки вычислительной модели.

Проведенные вычислительные эксперименты не только указали на положительную взаимосвязь между долей троллей в популяции агентов модели и долей агрессивных комментариев, но и позволили выявить диапазон, в котором может лежать соответствующий регрессионный коэффициент при различных значениях модельных гиперпараметров. Проведенные эксперименты также указали на существенную роль распределения порогов толерантности для значения регрессионного коэффициента. Способность же пользователей социальной сети к детекции троллей, как оказалось, имеет меньшее значение, нежели распределение порогов толерантности.

Собранные эмпирические данные об онлайн дискуссиях в социальной сети ВКонтакте в 2020-2021 гг. позволили провести калибровку модели. На основе метода максимального правдоподобия была получена оценка параметров бета-распределения индивидуальных порогов толерантности к агрессии. Результатом этой калибровки стал вывод о том, что наиболее вероятное значение регрессионного коэффициента, связывающего долю агрессивных комментариев под постом с долей троллей в популяции пользователей, равно 0.5. Иными словами, десятипроцентный рост доли троллей приведет в среднем к пятипроцентному росту доли агрессивных комментариев.

В ходе дальнейших исследований предсказания данной модели могут быть уточнены на основе совершенствования эвристических правил детекции троллей в массивах собранных данных.

## Литература


1. *Ferrara E., Varol O., Davis, C.A. Menczer F., Flammini A.* The rise of social bots // Communications of the ACM. Vol.59. 2016, №7. – P.96-104.
2. *Chavoshi N., Hamooni H., Mueen A.* DeBot: Twitter bot detection via warped correlation // Proceedings of the 8th International Conference on 6th IEEE International Conference on Data Mining, ICDM 2016.
3. *Stukal D., Sanovich S., Bonneau R., Tucker J.A.* Detecting political bots on Russian Twitter // Big Data. Vol. 5. 2017, №4ю – P.310-324.
4. *Shi P., Zhang Zh., Choo K.* Detecting malicious social bots based on clickstream sequences // IEEE Access. Vol.7. – P.28855-28862.
5. *Shao Ch., Ciampaglia G.,Varol O., Yang K., Flammini A., Menczer F.* The spread of low-credibility content by social bots // Nature Communications. Vol. 9. 2018. – P.4787.
6. *Keller T., Klinger U.* Social bots in election campaigns: theoretical, empirical, and methodological implications // Political Communication. Vol.36. 2019, №1. – P.171-189.
7. *Stukal D., Sanovich S., Bonneau R., Tucker J.A.* Why botter: how pro-government bots fight opposition in Russia

// American political science review. В печати.
8. *Ross B., Pilz L., Cabrera B., Brachten F., Neubaum G., Stieglitz S.* Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks // European Journal of Information Systems. Vol. 28. 2019, №4. – P.1–19.
9. *Cheng Ch., Luo Y., Yu Ch.* Dynamic mechanism of social bots interfering with public opinion in network // Physica A: statistical mechanics and its applications. Vol. 551. 2020. – P.124163.
10. *Bail Ch.A., Guay B., Maloney E., Combs A., Sunshine Hillygus D., Merhout F.* Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017 // Proceedings of the national academy of sciences. Vol. 117. 2019, №1. – P.243-250.
11. *Sobolev A.* How pro-government trolls influence online conversations in Russia // Неопубликованная рукопись.
12. *King G., Pan J., Roberts M.E.* How the Chinese government fabricates social media posts for strategic distraction, not engaged argument // American political science review. Vol.111. 2017, №3. – P. 484-501.
13. *Keller F.B., Schoch D., Stier S., Yang, J.* Political astroturfing on Twitter: how to coordinate a disinformation campaign // Political communication. Vol.37. 2020, №2. – P.256-280.
14. *Dlala I.O., Attiaoui D., Martin A., Yaghlane B.* Trolls identification within an uncertain framework // Неопубликованная рукопись.
15. *Stukal D., Sanovich S., Bonneau R., Tucker J.A.* The use of Twitter bots in Russian political communication // PONARS Eurasia Policy Memo No. 564, 2019.