DOI 10.37882/2223-2966.2025.05.36

ПРИМЕНЕНИЕ ТЕОРИИ ИГР И БЯМ-АНАЛИЗА КОНТЕНТА В СОЦИАЛЬНЫХ СЕТЯХ ДЛЯ ВЫЯВЛЕНИЯ РАВНОВЕСИЯ ПО НЭШУ И КЛЮЧЕВЫХ ИГРОКОВ В УСЛОВИЯХ МАССОВЫХ БЕСПОРЯДКОВ

APPLICATION OF GAME THEORY AND LLM-BASED CONTENT ANALYSIS IN SOCIAL NETWORKS TO IDENTIFY NASH EQUILIBRIUM AND KEY PLAYERS UNDER MASS RIOTS

Urtnasan Batnasan

Summary. This article proposes a methodology that combines game theory and automated content analysis to identify Nash equilibrium and «key» players during mass riots. Using the example of the protests in Mongolia, the calculation of utility functions is based on classifying comments (P, S, N), thereby supporting managerial decision-making in crisis situations.

Keywords: game theory, mass riots, social networks, LLM (large language model), content analysis, Nash equilibrium.

Уртнасан Батнасан
Адъюнкт, Академия управления МВД России, г. Москва
batnasan.u@mail.ru

Аннотация. В статье предложена методология, сочетающая теорию игр и автоматизированный анализ контента для выявления равновесия по Нэшу и «ключевых» игроков при массовых беспорядках. На примере протестов в Монголии проведён расчёт функций полезности, основанный на классификации комментариев (P, S, N), что помогает принимать управленческие решения в кризисных ситуациях.

Ключевые слова: теория игр, массовые беспорядки, социальные сети, БЯМ (большая языковая модель), контент-анализ, равновесие по Нэшу.

Введение

В современных условиях массовые беспорядки всё чаще сопровождаются активное использованием социальных сетей в качестве пространства для координации протестных действий, формирования общественных настроений и влияния на ход событий. В условиях высокой скорости информационного обмена, мультиагентного характера взаимодействия участников существенно возрастает значение системных и количественных подходов к анализу стратегического поведения акторов, действующих в онлайн-среде.

Примером, ярко иллюстрирующим эти тенденции, являются события произошедшие в Монголии в декабре 2022 года. Тогда общественность была взбудоражена масштабной коррупцией в угольном секторе около 385000 тонн угля, предназначенных для экспорта в Китай, были похищены. Это спровоцировало массовые протесты в Улан-Баторе, преимущественно среди молодёжи, требовавшей прозрачности, в деятельности правительства, реформ в угольной отрасли. Деятельность протестующих, официальных структур, средство массовых информации, негативно настроенных пользователей в социальный сетей сформировала сложный информационный ландшафт, в котором распространение эмоционально заряженного контента тесно переплеталось с дискурсом о коррупции, ответственном управлении и экономической стабильности.

В таких условиях перед органами правительство возникает вопрос: как формализовать многостороннее стратегическое взаимодействие в динамичных онлайн-средах и количественно оценить влияние информационного контента на выбор стратегий различных групп? Для решения подобных задач целесообразно использовать синтез теоретико-игрового моделирования и автоматизированного анализа данных о контенте. Теория игр предоставляет математический аппарат для поиска равновесных сценариев (равновесий по Нэшу), устойчивых к односторонним отклонениям участников. В то же время современные большие языковые модели (БЯМ), такие как Mistral 7B, способны эффективно классифицировать комментарии, выявляя провокационный (P), поддерживающий (S) и нейтральный (N) контент, а также оценивать эмоциональную окраску (позитивную, негативную) сообщений. Эта интеграция позволяет формировать функции полезности акторов, опирающиеся на реальные данные о настроениях и реакциях аудитории.

Цель данной статьи — показать методы, сочетающие теорию игр и БЯМ-анализ контента, применяемую к данным из социальных сетей в условиях массовых беспорядков, с последующим выявлением равновесия по Нэшу и определением «ключевых» игроков, чьи стратегические сдвиги оказывают наибольшее влияние на систему. Научная новизна работы заключается в интеграции двух подходов: теории игр и автоматизированной классификации контента, что позволяет перейти от абстрактных моделей к более точным инструментам прогнозирования стратегий участников.

Практическая значимость статьи состоит в возможности поддержки принятия управленческих решений при массовых беспорядках. Понимание равновесных стратегий, сформированных на основе реальных данных (применительно к случаю протестов в Монголии), и выявление ключевых акторов помогают целенаправленно воздействовать на информационное поле, стабилизируя ситуацию и минимизируя эскалацию конфликта.

В статье рассмотрена модель, включающая пять групп игроков (правительство, официальные СМИ, авторы негативные страницы, политические деятели, блогеры /инфлюенсеры/) и две стратегии для каждой группы. Классификация текстовых данных, собранных из социальных сетей, проводилась при помощи БЯМ Mistral 7B, а результаты интегрировались в теоретико-игровую постановку для расчёта полезностей игроков. Далее из множества возможных стратегических профилей выделялось равновесие по Нэшу, анализировалась чувствительность равновесия к изменениям стратегий и определялись ключевые игроки.

Обзор литературы

Взаимодействие участников в информационном пространстве социальных сетей при массовых беспорядках является предметом интереса в ряде исследований, посвящённых теории игр, анализу данных и автоматической обработке текстов. Теоретико-игровой подход широко применялся для моделирования стратегий в сетевых системах, начиная с классических работ по анализу кооперативных и некооперативных взаимодействий [1; 2]. В контексте информационного противостояния и дезинформации теория игр использовалась для исследования стратегий агентов, распространяющих фейковые новости и пропаганду [3; 4].

Важным направлением является развитие моделей, ориентированных на выявление «ключевых» игроков и оценку влияния в социальных сетях. Так, в работах Торопова предлагается использовать теоретико-игровую центральность вершин на основе вектора Шепли, что даёт возможность учесть маржинальный вклад каждого узла в формировании коалиций [5]. Подход «экшн-модели» (actional model), предложенный Губановым и Чхартишвили, позволяет формализовать влияние пользователей, исходя из их конкретных действий и целей внешнего «агента-руководителя» [6; 7]. В дальнейшем авторы расширили модель за счёт анализа метапользователей, объединивших несколько аккаунтов или групп, для оценки совокупного влияния [7]. В монографии Чхартишвили, Губанова и Новикова рассмотрены механизмы информационного контроля и противоборства в социальных сетях, опирающиеся на теоретико-игровые принципы и модели множественных акторов [8].

Особую актуальность приобрёл вопрос о том, как формализовать влияние контента в социальных сетях на решения акторов. Здесь на помощь приходят методы обработки естественного языка (NLP) и большие языковые модели (БЯМ). Применение БЯМ для сентимент-анализа и классификации содержания сообщений доказало свою эффективность в работах, посвящённых автоматической модерации и обнаружению токсичности в онлайн-дискуссиях [9; 10]). В последние годы большые языковые модели, такие как BERT, GPT-3, BLOOM, Mistral и др., активно интегрируются в социально-политические исследования для анализа настроений публики и определения типа комментариев [11; 12].

Применительно к условиям массовым беспорядкам, в которых социальные сети становятся ареной конкурирующих нарративов, существует потребность объединения теоретико-игрового моделирования с автоматической оценкой качественных характеристик контента. Предыдущие исследования показали, что учет эмоциональной окраски и типа комментариев (провокационные, поддерживающие, нейтральные) повышает точность прогнозов о поведении аудитории и устойчивости информационных стратегий [13; 14]. Влияние негативных и провокационных сообщений на эскалацию конфликтов в сетях, а также роль «ключевых» акторов, распространяющих подобный контент, анализировалось в контексте обнаружения сетевых лидеров, посредников и «воспламеняющих» узлов [15; 16]).

Однако в большинстве упомянутых работ теоретико-игровые методы и NLP-подходы рассматривались относительно независимо: либо анализировались стратегии агентов с заданными абстрактными полезностями, либо оценивался контент без строгой стратегической интерпретации. Объединение теории игр с БЯМ-анализом контента представляет собой перспективное направление, позволяющее связать результаты сентимент-анализа и классификации комментариев с формализацией функций полезности акторов, что может привести к более точному определению равновесных стратегий и выявлению ключевых игроков [17].

Таким образом, анализ литературы показывает, что в современной научной среде существует методологи-

ческая основа как для применения теории игр к социальным сетям, так и для использования БЯМ для аналитики контента. Но комплексный подход, объединяющий эти два направления для изучения массовых беспорядков и информационных стратегий, требует дальнейших разработок. Данное исследование отвечает на этот вызов, интегрируя теоретико-игровую постановку с автоматизированной классификацией комментариев и сентимент-анализа для выявления равновесия по Нэшу и определению ключевых акторов в условиях кризисных онлайн-ситуаций.

Модель и методология

В данном статья рассматривается упрощённая теоретико-игровая модель стратегического взаимодействия между пятью группами акторов, действующими в условиях массовых беспорядков и связанных посредством социальных сетей:

1. Правительственные организации (*Government, G*): их основная задача — поддержание порядка и снижение деструктивного эмоционального фона.
2. Официальные СМИ (*Media, M*): стремятся к максимизации охвата и формированию позитивной или, по крайней мере, нейтральной информационной повестки.
3. Негативные страницы (*Negative, Neg*): провокаторы, распространяющие деструктивную, провокационную информацию с целью усилить хаос.
4. Политические акторы (*Politics, Pol*): заинтересованы в мобилизации сторонников и укреплении политического влияния.
5. Инфлюенсеры (*Influencers, Inf*): фокусируются на повышении вовлечённости аудитории, стремятся укрепить свою узнаваемость и позицию в медиапространстве.

Каждый актор имеет две стратегии. Например:

- *G: G1* — умеренная, стабильная подача информации; *G2* — жёсткая риторика и усиление контроля.
- *M: M1* — нейтральная, сбалансированная подача; *M2* — более позитивный уклон, смягчение конфликта.
- *Neg: Neg1* — провокационный шок-контент; *Neg2* — сенсационная, но менее агрессивная подача.
- *Pol: Pol1* — конструктивный, мобилизующий тон; *Pol2* — политизация с меньшей поддержкой.
- *Inf: Inf1* — провокация для вовлечения; *Inf2* — попытка поддержать вовлечённость с меньшим накалом.

Таким образом, всего имеется $2^2 = 32$ возможных комбинаций стратегий. Чтобы оценить полезности каждого игрока, необходимо количественно связать их цели с метриками активности аудитории (комментарии, реакции), а также с качественными характеристиками контента (позитив, негатив, провокации, поддержка).

Определение метрик и применение БЯМ:

Для вычисления функций полезности необходимы данные о реакции аудитории. Предположим, что изначально собран набор сообщений, реакций (*like, love, angry, sad*) и комментариев, поступивших в социальных сетях. Комментарии, по своему содержанию, разделяются на три типа:

- *P* (*провокационные*): подогревают конфликт, стимулируют негативные эмоции.
- *S* (*поддерживающие*): укрепляют позитивный настрой, лояльность к конкретному актору.
- *N* (*нейтральные*): отражают интерес, но без выраженных позитивных или негативных эмоций.

Аналогично, реакции пользователей классифицируются на позитивные (например, *like, love*), негативные (*angry, sad*) и оценивается общее число вовлечённости (*total reactions, total comments*).

Для определения типа комментариев и проведения сентимент-анализа используется крупная языковая модель БЯМ Mistral 7B. Модель автоматически классифицирует контент сообщений на *P, S, N* и оценивает общую эмоциональную окраску (*Pos, Neg*).

Используемый промт для классификации комментариев был следующим:

Отнесите следующий комментарий к одной из трех категорий: нейтральный (N), поддерживающий (S), провокационный (P).

Комментарий: «{comment}»

Категории:

1. *Нейтральный: Комментарий не выражает какого-либо сильного положительного или отрицательного мнения.*
2. *Поддерживающий: Комментарий выражает положительное или поддерживающее мнение.*
3. *Провокационный: Комментарий выражает отрицательное или провокационное мнение.*

Укажите категорию классификации: «»

Полученные таким образом агрегированные статистики по каждой группе акторов при каждой комбинации стратегий позволяют вычислить полезности.

Функции полезности:

Полезности каждого игрока отражает его интересы:

Таблица 1.

Комбинация (00000) *G1,M1,Neg1,Pol1,Inf1*							
Игрок	Комменты			Всего	Реакция		Всего
	N	P	S		Pos	Neg	
Government(G1)	66	37	51	154	217	87	304
Media(M1)	257	143	194	594	2895	965	3860
Negative(Neg1)	321	173	281	775	3186	4063	7249
Politics(Pol1)	158	66	132	356	731	269	1000
Influencers(Inf1)	479	238	423	1140	1379	648	2027
Комбинация 01001 (*M2+Inf2*)							
Игрок	Комменты			Всего	Реакция		Всего
	N	P	S		Pos	Neg	
Government	66	37	51	154	221	83	304
Media(M2)	257	143	214	624	3474	869	4053
Negative	321	173	281	775	3249	3860	7110
Politics	158	66	132	356	746	256	1000
Influencers(Inf2)	407	202	360	798	1241	518	1420
Комбинация 11111 (*G2+M2+Neg2+Pol2+Inf2*)							
Игрок	Комменты			Всего	Реакция		Всего
	N	P	S		Pos	Neg	
Government(G2)	66	37	41	146	174	96	289
Media(M2)	257	143	214	624	3474	869	4053
Negative(Neg2)	321	173	281	698	3345	3250	6524
Politics(Pol2)	174	66	112	338	658	242	950
Influencers(Inf2)	407	202	360	798	1241	518	1420

Government (G):

Стремится снизить негатив, увеличить долю позитивных реакций. Полезности может быть задан как:

$$U_G = \frac{Pos}{Tot} - \frac{Neg}{Tot} - \frac{P}{Comm + 1}$$

Это учитывает соотношение позитивных и негативных сигналов и эффект провокаций (P).

Media (M):

Ориентируются на охват, позитив и уменьшение негатива:

$$U_M = \frac{Pos}{Tot} - \frac{Neg}{Tot} + \frac{Comm}{Tot}$$

Negative (Neg):

Цель — максимизация негативных эмоций и провокаций:

$$U_{Neg} = \frac{Neg}{Tot} + \frac{P}{Comm + 1}$$

Politics (Pol):

Стремятся к увеличению поддерживающих (S) и позитивных реакций, снижению негатива:

$$U_{Pol} = \frac{S}{Tot} + \frac{Pos}{Tot} - \frac{Neg}{Tot}$$

Influencers (Inf):

Максимизируют общую вовлечённость (комментарии и все реакции):

$$U_{Inf} = \frac{Comm}{Tot} + \frac{Pos + Neg}{Tot}$$

Здесь *Tot* — общее число реакций, *Comm* — число комментариев, *Pos* и *Neg* — количество позитивных и негативных реакций, *P* и *S* — число провокационных и поддерживающих сигналов, определённых БЯМ.

Пример вычисления полезностов для 01001 комбинации:

U_G:

$$U_G = \frac{221}{304} - \frac{83}{304} - \frac{37}{155} =$$
$$= 0.7263 - 0.2724 - 0.2387 = 0.2152$$

U_M (M2):

$$U_M = \frac{3474}{4053} - \frac{869}{4053} + \frac{624}{4053} =$$
$$= 0.857 - 0.2145 + 0.154 = 0.7965$$

U_{Neg}:

$$U_{Neg} = \frac{3860}{7110} + \frac{173}{776} = 0.542 + 0.223 = 0.765$$

U_{Pol}:

$$U_{Pol} = \frac{132}{1000} + \frac{746}{1000} - \frac{256}{1000} =$$
$$= 0.132 - 0.746 - 0.256 = 0.622$$

U_{Inf} (Inf2):

$$U_{Inf} = \frac{798}{1420} + \frac{1241 + 518}{1420} = 0.562 + 1.239 = 1.801$$

Итог (01001): $U_G = 0.2152$, $U_M = 0.7965$, $U_{Neg} = 0.765$, $U_{Pol} = 0.622$, $U_{Inf} = 1.801$.

Процедуры определение равновесия по Нэшу и ключевых игроков:

- Рассчитать полезности для всех 32 комбинаций стратегий, подставляя значения *Pos, Neg, Comm, P, S, N* (классифицированные БЯМ и скорректированные под стратегии).
- Для каждого профиля проверить, заинтересован ли какой-либо игрок изменить свою стратегию односторонне. Если да профиль не является равновесием.
- Найти комбинацию, при которой ни один актор не хочет менять стратегию это и есть равновесие по Нэшу.
- Чтобы определить ключевых игроков, проанализировать чувствительность равновесия к изменению стратегии одного актера и оценить, как сильно это изменение влияет на полезности остальных. Если влияние существенно, актор считается ключевым.

Таким образом, метод состоит из нескольких этапов:

1. Классификация комментариев и сентимент-анализ с помощью БЯМ Mistral 7B.
2. Определение функций полезности на основе полученных метрик.
3. Моделирование 32 стратегических профилей акторов и расчет полезностов.
4. Анализ односторонних отклонений для выявления равновесия по Нэшу.
5. Проверка чувствительности равновесия и определение ключевых игроков.

Данный комплексный подход обеспечивает формализованную основу для поддержки управленческих решений в условиях массовых беспорядков, предоставляя возможность прогнозировать устойчивые паттерны поведения в соцсетях и выявлять наиболее влиятельных акторов информационного поля.

Результаты вычислений

В процессе исследования для всех $2^5 = 32$ комбинаций стратегий пяти групп игроков (*Government*, *Media*, *Negative*, *Politics*, *Influencers*) были вычислены показатели полезности. Исходные метрики вовлечённости, эмоциональной окраски (*Pos*, *Neg*), типов комментариев (*P* — провокационные, *S* — поддерживающие, *N* — нейтральные) были получены с использованием МОСИНТ база данных, скорректированных под эффект стратегий, и классифицированы посредством БЯМ Mistral 7B. Это позволило количественно оценить влияние каждой комбинации стратегий на выигрыш каждого актера.

По результатам вычислений (см. табл. 2).

Таблица 2.

Итоговые значения полезности для всех комбинаций стратегий

	Код (*G,M,Neg,Pol,Inf*)	U_G	U_M	U_{Neg}	U_{Pol}	U_{Inf}
1	00000	0.1896	0.654	0.783	0.594	1.562
2	00001	0.2073	0.67	0.773	0.609	1.801
3	00010	0.2152	0.6795	0.765	0.5558	1.563
4	00011	0.2349	0.6938	0.756	0.5558	1.801
5	00100	0.2009	0.664	0.7455	0.603	1.564
6	00101	0.2147	0.6787	0.7455	0.618	1.801
7	00110	0.227	0.6896	0.7455	0.5558	1.565
8	00111	0.2416	0.7031	0.7455	0.5558	1.801
9	01000	0.2023	0.7965	0.776	0.607	1.562
10	01001	0.2152	0.7965	0.765	0.622	1.801
11	01010	0.2278	0.7965	0.759	0.5558	1.563
12	01011	0.2433	0.7965	0.748	0.5558	1.801
13	01100	0.2147	0.7965	0.7455	0.616	1.564
14	01101	0.227	0.7965	0.7455	0.631	1.801
15	01110	0.2416	0.7965	0.7455	0.5558	1.565
16	01111	0.2547	0.7965	0.7455	0.5558	1.801
17	10000	0.0182	0.654	0.757	0.594	1.581
18	10001	0.0182	0.67	0.747	0.609	1.801
19	10010	0.0182	0.6796	0.74	0.5558	1.581
20	10011	0.0182	0.6934	0.73	0.5558	1.801
21	10100	0.0182	0.664	0.7455	0.602	1.583
22	10101	0.2147	0.6787	0.7455	0.618	1.801
23	10110	0.227	0.6896	0.7455	0.5558	1.565
24	10111	0.2351	0.7082	0.7455	0.5558	1.801
25	11000	0.0182	0.7965	0.783	0.606	1.58
26	11001	0.0182	0.7965	0.783	0.622	1.801
27	11010	0.0182	0.7965	0.783	0.5558	1.58
28	11011	≈0.2417	0.7082	≈0.73	0.5558	1.801
29	11100	0.0182	0.7965	0.7455	0.603	1.583
30	11101	0.2147	0.6787	0.7455	0.618	1.801
31	11110	0.2286	0.7965	0.7455	0.5558	1.565
32	11111	0.0182	0.7965	0.7455	0.5558	1.801

Общий характер полезностей:

- Правительственные организации (*G*) при переходе к более жёсткой стратегии (*G2*) практически всегда снижали свою полезность по сравнению с мягкой (*G1*).
- Официальные СМИ (*M*) стабильно выигрывали при переходе к более позитивной подаче (*M2*) по сравнению с нейтральной (*M1*).
- Негативные страницы (*Neg*), ориентированные на негатив и провокации, в большинстве случаев получали максимальные полезности, оставаясь на провокационной, но агрессивной стратегии (*Neg1*), нежели смягчаясь до *Neg2*.
- Политические акторы (*Pol*) предпочитали конструктивно-мобилизующий тон (*Pol1*) вместо более компромиссного варианта (*Pol2*), поскольку изменение в *Pol2* обычно снижало их итоговую полезность.
- Инфлюенсеры (*Inf*) получали наибольший выигрыш при выборе стратегии *Inf2* (повышение вовлечённости без максимальной провокации), нежели оставаться на более грубом варианте *Inf1*.

Выявление равновесия по Нэшу:

Проверка всех 32 комбинаций на предмет того, может ли один из акторов увеличить свой полезности односторонним изменением стратегии, показала, что лишь для одного профиля стратегий отсутствует стимул к одностороннему отклонению со стороны любого игрока. Таким равновесием по Нэшу оказалась комбинация:

— *Government: G1*
— *Media: M2*
— *Negative: Neg1*
— *Politics: Pol1*
— *Influencers: Inf2*

В двоичном представлении это профиль (01001). В этом состоянии:

— G достигает оптимального баланса позитив/негатив при умеренной подаче.
— M максимизирует охват при позитивной стратегии (M2).
— Neg сохраняет провокационный контент, не снижая агрессии (Neg1).
— Pol придерживается конструктивного, мобилизующего тона (Pol1).
— Inf формирует высокую вовлечённость без чрезмерной провокации (Inf2).

Ни одному игроку не выгодно единолично изменить стратегию, поскольку любая такая смена ухудшает его полезность.

Ниже приведён пример анализа односторонних отклонений от равновесного профиля (01001) — (*G1,M2,Neg1,Pol1,Inf2*). Исходные полезности в равновесии:

$$U_G = 0.2152\,,\ U_M = 0.7965\,,\ U_{Neg} = 0.765\,,\ U_{Pol} = 0.622\,,\ U_{Inf} = 1.801$$

Из таблицы 3 видно, что при попытке любого из акторов в одиночку сменить стратегию его полезности снижается. Это подтверждает, что профиль (01001) является равновесным по Нэшу, поскольку ни один игрок не улучшает свой выигрыш при одностороннем отклонении.

Определение ключевых игроков

После выявления равновесия была оценена чувствительность этого состояния к односторонним изменениям стратегий. Анализ показал, что изменение стратегии со стороны негативных страниц (*Neg*) оказывает наиболее заметное влияние на полезность некоторых других акторов, в частности СМИ. Переход *Neg* с *Neg1* к *Neg2* приводил к существенному снижению выигрышности M, тогда как другие изменения стратегий других групп не вызывали столь резких колебаний в полезностях третьих лиц. Таким образом, «Негативные страницы» проявили себя как более «ключевой» актор, изменение которого способно заметно трансформировать распределение выигрышей в информационной среде.

Таблица 3.

Таблица примера односторонних отклонений

Игроки	Текущая стратегия	Новая стратегия	Новый профиль (бинарный код)	Полезности акт. после изменения	Улучшение?
G: G1→G2	*G1 (0)*	*G2 (1)*	(11001)	$U_G = 0.0182 < 0.2152$	Нет (хуже)
M: M2→M1	*M2 (1)*	*M1 (0)*	(00001)	$U_M = 0.67 < 0.7965$	Нет (хуже)
Neg: Neg1→Neg2	*Neg1 (0)*	*Neg2 (1)*	(01101)	$U_{Neg} = 0.7455 < 0.765$	Нет (хуже)
Pol: Pol1→Pol2	*Pol1 (0)*	*Pol2 (1)*	(01011)	$U_{Pol} = 0.5558 < 0.622$	Нет (хуже)
Inf: Inf2→Inf1	*Inf2 (1)*	*Inf1 (0)*	(01000)	$U_{Inf} = 1.562 < 1.801$	Нет (хуже)

Ниже приведён пример сравнения полезностов при переходе негативных страниц (*Neg*) со стратегии *Neg1* к *Neg2* в равновесии (01001). Исходно (01001) соответствует (*G1,M2,Neg1,Pol1,Inf2*) с полезностями:

$$U_G = 0.2152\,,\ U_M = 0.7965\,,\ U_{Neg} = 0.765\,,\ U_{Pol} = 0.622\,,\ U_{Inf} = 1.801$$

При изменении *Neg1→Neg2* профиль становится (01101) = (*G1,M2,Neg2,Pol1,Inf2*), и полезносты меняются:

Таблица 4.

Сравнение полезностов акторов при изменении стратегии *Neg* с *Neg1* на *Neg2* в равновесии (01001)

Актор	Полезности при (01001)	Полезности при (01101)	Изменение
G	0.2152	0.227	+0.0118 (*небольшое увеличение*)
M	0.7965	0.6787	–0.1178 (*существенное снижение*)
Neg	0.765	0.7455	–0.0195 (*умеренное снижение*)
Pol	0.622	0.631	+0.009 (*небольшое увеличение*)
Inf	1.801	1.801	0 (*без изменений*)

Обсуждение результатов и управленческие рекомендации

Выявленное равновесие по Нэшу, соответствующее профилю (*G1,M2,Neg1,Pol1,Inf2*), показывает, что в условиях массовых беспорядков и информационного давления оптимальные стратегии акторов можно описать через умеренность правительства, позитивизацию сообщений официальных СМИ, сохранение провокационной активности негативных страниц, конструктивный политический тон и стремление инфлюенсеров к повышению вовлечённости без чрезмерного нагнетания. Данный результат указывает, что смоделированных на основе категорий комментариев (*P, S, N*) с помощью БЯМ Mistral 7B, существуют предсказуемые, устойчивые структуры взаимодействий.

Интересно отметить, что изменять свою стратегию при равновесии по Нэшу невыгодно никому. Это означает, что обнаруженный профиль стратегий стабильное состояние, к которому информационная система может склониться при отсутствии внешнего воздействия. С управленческой точки зрения данное равновесие важно: если целью являются снижение напряжённости или контроль информационного поля, то необходимо понимать, какие стратегические ходы не изменят устойчивого положения. Например, попытка правительства перейти к более жёсткой риторике (*G2*) или СМИ вернуться к нейтральности (*M1*) не приведёт к кардинальным выгодам для этих акторов.

Особый интерес представляет анализ чувствительности равновесия. Наибольший эффект на распределение выигрышей вызывает изменение стратегии со стороны негативных страниц (*Neg*). Это значит, что при прочих равных условиях именно *Neg* обладает признаками «ключевого» игрока: изменение их поведения способно сместить баланс интересов и полезностей других групп, особенно СМИ. С управленческой точки зрения данное знание означает, что именно негативные игроки информационного поля представляют собой уязвимую точку или, напротив, решающий рычаг в информационном противостоянии. Контролируя, нейтрализуя или переориентируя негативных акторов, можно повлиять на весь баланс, сместив систему из равновесного состояния в более желательное для управленцев положение.

Использование БЯМ для классификации комментариев и сентимент-анализа позволило формализовать сложные информационные сигналы и включить их в математическую модель. Это придаёт анализу дополнительную практическую ценность: применяя более точные и мощные модели, можно оперативно оценивать текущее состояние сетевой дискуссии, определять типы комментариев и уровень негативных реакций. На основе этого можно пересматривать стратегии собственных информационных кампаний или точечно воздействовать на ключевых игроков, ограничивая негативный или провокационный контент, либо стимулируя позитивные сигналы.

С точки зрения управленческих решений, данный подход предоставляет следующие рекомендации:

- Идентификация ключевых игроков: Управленцы, зная, что негативные страницы наиболее критичны для равновесия, могут приоритизировать мониторинг и противодействие именно им.
- Стратегический контроль информационного поля: Понимание равновесных стратегий помогает предвидеть, что даже при попытках изменения поведения отдельных акторов без комплексного вмешательства равновесие сложно сместить. Значит для существенного воздействия нужна целенаправленная работа с «ключевыми» элементами системы.
- Гибкая реакция на изменения: Если негативный актор меняет стратегию, реакции СМИ или правительства могут быть скорректированы заранее, опираясь на расчёт возможных полезностов. Это снижает фактор неожиданности и повышает устойчивость управленческих решений.

Таким образом, обсуждённые результаты указывают на то, что предложенная модель и методология могут

стать инструментом для понимания устойчивых паттернов поведения в социальных сетях при массовых беспорядках, а также для разработки точечных и эффективных управленческих мероприятий по стабилизации информационного пространства.

Заключение

В данной работе была предложена и апробирована комплексная методология, сочетающая теоретико-игровое моделирование и автоматизированный анализ контента с помощью большой языковой модели БЯМ (Mistral 7B), для изучения стратегических взаимодействий ключевых акторов в условиях массовых беспорядков. Через формализацию стратегий и вычисление полезностей, зависящих от типа комментариев (*P, S, N*) и эмоциональной окраски (*Pos, Neg*), удалось определить равновесный профиль стратегий по Нэшу и выявить «ключевых» игроков.

Основным результатом стало выявление уникального равновесия по Нэшу, при котором правительство остаётся умеренным (*G1*), СМИ выбирают позитивную подачу (*M2*), негативные страницы придерживаются провокационного контента (*Neg1*), политики действуют конструктивно (*Pol1*), а инфлюенсеры стремятся к максимизации вовлечённости при меньшем накале (*Inf2*). При таком сочетании стратегий ни один из игроков не имеет стимула к одностороннему изменению своей стратегии, что указывает на устойчивый характер данного «сценария» информационного взаимодействия.

Дополнительный анализ чувствительности показал, что именно негативные акторы (*Neg*) обладают наиболее сильным влиянием на распределение итоговых полезностей остальных участников. Их переход к альтернативной стратегии способен существенно повлиять на выигрыш, например, официальных СМИ, что позволяет классифицировать негативных акторов как ключевых.

Практическая ценность исследования заключается в том, что предложенный подход может быть использован для поддержки управленческих решений во время кризисных ситуаций и массовых беспорядков. Понимание равновесных стратегий позволяет предвидеть устойчивые конфигурации информационного взаимодействия, а знание ключевых игроков нацелить меры на наиболее критичные точки влияния. Интеграция теории игр и БЯМ-анализа контента даёт инструментарий для оперативного мониторинга и моделирования онлайн-пространства, позволяя вырабатывать проактивные, научно обоснованные меры по стабилизации ситуации.

Перспективы дальнейших исследований связаны с применением других исходных данных, расширением набора стратегий, уточнением функций полезностей, применением более мощных БЯМ-моделей и добавлением динамических аспектов в модель. Это позволит повысить точность прогнозов и практическую применимость разработанных методов в реальных сценариях массовых беспорядков и иных кризисных ситуациях в информационном поле.

ЛИТЕРАТУРА

1. Myerson R.B. Game Theory: Analysis of Conflict. — Cambridge (MA): Harvard University Press, 1991. — 568 p.
2. Osborne M.J., Rubinstein A. A Course in Game Theory. — Cambridge (MA): MIT Press, 1994. — 352 p.
3. Basu A., Dickinson T., Xu H. Modeling the spread of fake news in social networks using game-theoretic frameworks // Proceedings of the AAAI Workshop on Fake News. — 2018.
4. Chen W., Hua Y., Shi J. A Fake News Game Theory Model of Two-Sided Markets // IEEE Transactions on Computational Social Systems. — 2020. — Vol. 7, no. 2. — P. 451–462.
5. Торопов Б.А. Теоретико-игровая центральность вершин в графах на основе вектора Шепли // Программные системы и вычислительные методы. — 2017. — № 2. — С. 45–54.
6. Губанов Д.А., Чхартишвили А.Г. An actional model of user influence levels in a social network // Automation and Remote Control. — 2015. — Vol. 76, no. 7. — P. 1282–1290.
7. Губанов Д.А., Чхартишвили А.Г. Influence levels of users and meta-users of a social network // Automation and Remote Control. — 2018. — Vol. 79, no. 3. — P. 545–553.
8. Чхартишвили А.Г., Губанов Д.А., Новиков Д.А. Social Networks: Models of Information Influence, Control and Confrontation. — Cham: Springer, 2019. — 247 p.
9. Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) // Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). — 2019.
10. Zhang Z., Robinson D., Tepper J. Detecting Hate Speech on Social Media using BERT and LSTM Models // ACL Workshop. — 2021.
11. Badjatiya P., Gupta S., Gupta M., Varma V. Deep learning for hate speech detection in tweets // Companion Proceedings of the Web Conference (WWW). — 2020. — P. 29–30.
12. Liu X., Zhang Y., Yu H. Large Language Models in Automated Content Moderation // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2022.
13. Ferrara E. Disinformation and social bot operations in the run up to the 2017 French presidential election // First Monday. — 2017. — Vol. 22, no. 8.
14. Kou Y., Gui L., Chen J., Zhao Y. Understanding online harassment on social media: A systematic literature review // Proceedings of the ACM on Human-Computer Interaction (PACMHCI). — 2020. — Vol. 4, no. CSCW2. — Article 158.
15. Leavitt A., Robinson J.J. The Role of Information Intermediaries in Social Media // iConference 2017 Proceedings. — 2017. — P. 714–722.
16. Carley K. M., et al. Disinformation in social media: Cognitive and predictive frameworks // IEEE Transactions on Emerging Topics in Computational Intelligence. — 2019. — Vol. 3, no. 2. — P. 106–119.
17. Ribeiro M.H., Calais P.H., Santos L.B., Almeida V.A., Meira W.Jr. Automatically Classifying News Content as Disinformation using Deep Neural Networks // Journal of Information Science. — 2021. — Vol. 47, no. 5. — P. 620–634.

Журнал «Современная наука: актуальные проблемы теории и практики»