

Отрицателни ефекти на автоматизираните модерационни системи: неправомерното блокиране на акаунти и последиците за онлайн присъствието на бизнеса

Negative Effects of Automated Moderation Systems: Unjust Account Blocking and the Consequences for Business Online Presence

Василена Василева¹

Резюме

С развитието на социалните мрежи и техния нарастващ принос към дигиталната икономика, платформите като Facebook и Instagram играят ключова роля в съвременния маркетинг. Те предоставят на бизнеси и създатели на съдържание възможността да достигнат до милиони потребители по цял свят, като същевременно изискват поддържането на високи стандарти за безопасност и отговорност. За следенето на съдържанието по цял свят групата на META използва автоматизирани системи, които не винаги правят правилна оценка за наличното съдържание и по този начин редица профили, страници и акаунти биват блокирани неправомерно.

Ключови думи: дигитален маркетинг, онлайн бизнес, онлайн присъствие

Abstract

With the development of social networks and their growing contribution to the digital economy, platforms like Facebook and Instagram play a key role in modern marketing. They provide businesses and content creators with the opportunity to reach millions of users worldwide, while also requiring the maintenance of high standards for safety and responsibility. To achieve this, Meta Group employs automated systems, which do not always correctly assess the available content, leading to numerous profiles, pages, and accounts being unjustly blocked.

Keywords: digital marketing, online business, online business presence

Въведение

Социалните мрежи не са просто дигитални платформи за комуникация – те се утвърдиха като един от най-значимите и динамични двигатели на съвременната икономика, който основно променя начина, по който бизнесът взаимодейства с клиентите си и изгражда своето пазарно присъствие. В ерата на глобална дигитализация около 5.52 милиарда души имат достъп до интернет (приблизително 69% от световното население) – цифра, която отразява бързото нарастване на интернет потребителите, особено в развитите региони, където достъпът до интернет е почти универсален. Същевременно, значителен брой хора – около 2.7 милиарда – все още остават офлайн, предимно в по-слабо развитите райони на Азия и Африка, където разходите и липсата на цифрови умения продължават да ограничават разширяването на интернет достъпа (DataReportal – Global Digital Insights, 2024).

Социалните платформи като Facebook, Instagram и LinkedIn вече не са само места за забавление и лични контакти, а се превръщат в стратегически инструменти за маркетинг, продажби и

¹ Докторант към катедра „Индустриален бизнес“, УНСС, e-mail: vasilena.vasileva@unwe.bg

създаване на дългосрочна потребителска лоялност. Тези платформи предоставят на компаниите безпрецедентни възможности за прецизен достъп до глобални пазари, недостижим за традиционните медии. С увеличената интернет свързаност и разширяването на мобилните технологии (около 70.3% от хората използват мобилни телефони, като 87% от тях са смартфони), социалните мрежи предлагат все по-голям потенциал за ангажиране на потребители по целия свят.

В дигиталната икономика, където всяко взаимодействие е измеримо, социалните мрежи предоставят инструменти за прецизно таргетиране на аудитории и детайлен анализ на маркетинговите кампании, като позволяват на компаниите не само да следят ефективността на своите кампании в реално време, но и да адаптират стратегиите си въз основа на реални данни за ангажираността и възвръщаемостта на инвестициите (ROI). Това е трансформация, която не само променя комуникационните канали, но и същността на глобалния бизнес, премахвайки географските бариери и предлагайки неограничени възможности за растеж и развитие.

Пример за мащаба на социалните платформи от групата на Meta, са официалните данни, публикувани във финансовите им отчети за 2023 г., според които компанията отчита годишни приходи от около 117 милиарда долара, от които приблизително 113 милиарда са дошли от реклами. (META Investor Relations , 2023) Рекламата остава основният източник на доход за компанията, съставлявайки почти 97% от общите ѝ приходи. За последното тримесечие Meta отчете общи приходи от 40.1 милиарда долара, което е значително увеличение спрямо предходната година. Основният дял от тези приходи е генериран от реклами, които са донесли 38.7 милиарда долара, представлявайки ръст от почти 24% спрямо същия период през 2022 г. Това увеличено рекламно търсене затвърждава ролята на платформата като значим рекламен канал в дигиталната индустрия. (META Investor Relations , 2023) Традиционните медии като радио и телевизия, въпреки своя широк обхват, не могат да предложат същите механизми за измерване и адаптиране на кампаниите в реално време, каквито осигуряват социалните мрежи, като същевременно са значително по-скъпи от дигиталните рекламни канали.

Освен всички положителни страни социалните мрежи обаче създават и нови предизвикателства. Автоматизираните системи за модерирание на съдържание често водят до неправомерни блокирания и ограничения, което се отразява сериозно на малките и средни предприятия, разчитащи на тези платформи като основен канал за достигане до клиенти. В тази нова дигитална реалност липсата на ефективно присъствие в социалните мрежи води до пропуснати възможности и отслабени пазарни позиции – онлайн присъствието вече не е просто избор, а условие за успех и устойчивост.

Политиките на META

Глобалното присъствие на **META** (включително Facebook, Instagram, WhatsApp и Messenger) имат глобално присъствие и това налага спазването на различни законови изисквания и регулации в рамките на всяка юрисдикция. Това включва закони за защита на личните данни, като **GDPR** в Европейския съюз, както и правила за защита на потребителите и интернет съдържанието в различни държави. В резултат, META трябва постоянно да адаптира своите политики и да внедрява системи, които гарантират безопасността на потребителите и предотвратяват незаконни или нежелани практики.

Поради глобалния си обхват, социалните мрежи на META са принудени да въведат редица политики за регулиране на потребителското поведение, като целта е да се осигури безопасно сърфиране и използване на платформите във всяка държава, съгласно местните закони и културни норми. Това обхваща всичко – от съдържание, свързано с насилие и престъпления, до дезинформация

и онлайн тормоз, които се контролират чрез стриктни насоки и автоматизирани инструменти за модериране.

Таблица 1: Политики и правила на МЕТА. Източник: META Transparency Center

ПРАВИЛА И ПОЛИТИКИ В ПЛАТФОРМИТЕ НА МЕТА						
ПОЛИТИКА СРЕЩУ НАСИЛИЕ И ПРЕСТЪПНО ПОВЕДЕНИЕ	ПОЛИТИКА ЗА БЕЗОПАСНОСТ	ПОЛИТИКА ЗА НЕПРИЕМЛИВО СЪДЪРЖАНИЕ	ПОЛИТИКА ЗА ИНТЕГРИТЕТ И АВТЕНТИЧНОСТ	ПОЛИТИКА ЗА УВАЖАВАНЕ НА ИНТЕЛЕКТУАЛНАТА СОБСТВЕНОСТ	ПОЛИТИКА ЗА ЗАЯВКИ И РЕШЕНИЯ, СВЪРЗАНИ СЪС СЪДЪРЖАНИЕ	ДРУГИ ПОЛИТИКИ
Насилие и подстрекателство	Самоубийство и самонараняване	Омразна реч	Интегритет на акаунта и автентична самоличност	Интелектуална собственост	Потребителски заявки	Политика за рекламиране
Опасни лица и организации	Сексуална експлоатация, насилие и голота на деца	Насилствено и графично съдържание	Спам	-	Допълнителна защита на непълнолетни деца	Политика за разработчици
Координиране на вреди и насърчване на престъпността	Сексуална експлоатация на възрастни	Голота на възрастни и сексуална дейност	Киберсигурност	-	-	Политика за страници, общности и събития за брандирано съдържание
Ограничени стоки и услуги	Малтретиране и тормоз	Сексуално увещаване	Неавтентично съдържание	-	-	Насоки на общността на Instagram
Измами	Експлоатация на хора	-	Невярна информация	-	-	Правен отдел на WhatsApp
-	Нарушение на поверителността	-	Преобразуване на акаунт във везпоменателен	-	-	Правен отдел на Oculus
-	-	-	-	-	-	Насоки за препоръки

Към 2024 г. МЕТА има огромна база данни от активни потребители и бизнеси в своите платформи:

- **Facebook активни потребители:** Над **2.98 милиарда** активни потребители месечно. Към тях има създадени около 200 млн. бизнес страници, като от тях 10 млн. са Активни рекламодатели в МЕТА (5%).
- **Instagram активни потребители:** Над **2.35 милиарда** активни потребители месечно.
- **WhatsApp:** Над **2.78 милиарда** активни потребители месечно.
- **Messenger:** Близко **1.3 милиарда** активни потребители месечно.

За да се справи с проследяването на спазването на всички политики в съдържанието, генерирано от около **3.5 млрд. уникални активни потребители**, МЕТА използва автоматизирани решения, базирани основно на изкуствен интелект (AI) и машинно обучение (ML), които следят и анализират всяка публикация, изображение или видео. AI алгоритмите са обучени да разпознават съдържание, което потенциално нарушава правилата на платформата, и да предприемат действия като премахване или ограничаване на достъпа до това съдържание. Основната цел на тези алгоритми е да поддържат стандартите на социалната мрежа и да създадат сигурна и приятелска среда за всички потребители, като така минимизират рисковете за онлайн обществото.

Въпреки ползите от автоматизираната модерация, внедряването на AI-базирани ботове за обработка на съдържание поставя предизвикателства пред специалистите по дигитален маркетинг и създателите на съдържание. Грешките в разпознаването на контекст или специфични културни особености могат да доведат до неправомерни блокирания или премахвания на съдържание и акаунти, които всъщност не нарушават правилата. Това създава проблеми за бизнесите и създателите, чието съдържание може да бъде погрешно идентифицирано като нарушаващо, намалявайки ефективността на маркетинговите им стратегии и ограничавайки достъпа им до аудиторията.

Предизвикателства пред изкуствения интелект (AI) и последици от автоматизираните решения за следене на съдържанието

Модерирането на съдържание чрез изкуствен интелект (AI) често води до значителни предизвикателства, особено по отношение на неправилната идентификация на нарушения на политиките на социалните мрежи. Алгоритмите, използвани от платформите като Facebook и Instagram, базират решенията си на предварително зададени критерии, но въпреки високото ниво на автоматизация, тези системи не са безупречни. Специалистите по дигитален маркетинг и създателите на съдържание често стават жертва на автоматични рестрикции, които могат да засегнат рекламни акаунти, страници и профили, дори когато те не нарушават политиките на платформите. Тези грешки водят до сериозни последици, включително блокиране или пълно изтриване на акаунти без обективна причина, което затруднява нормалната бизнес дейност. За все повече компании социалните мрежи са основен канал за генериране на приходи, както и за изграждане на дълготрайни взаимоотношения с клиентите. В този контекст, една автоматизирана грешка от страна на AI, която води до рестрикция или блокиране на профил или публикация, може да доведе до значителни загуби и пропуснати ползи.

Последиците от автоматизираните решения, основани на AI, се усещат на глобално ниво. Според данни от 2023 година, повече от 33 милиона публикации във Facebook и Instagram са били блокирани или ограничени в резултат на автоматизирани решения на системите за модериране (Content Restrictions Based on Local Law, 2023). Значителна част от тези случаи е засегнала както малки, така и големи бизнеси, които разчитат на тези платформи за маркетинг и продажби. В около 2,5% от случаите блокирането на профили или рекламни акаунти е било пълно, което е оказало сериозно въздействие върху компаниите, чиято онлайн активност е била временно или трайно спряна. Последващото възстановяване на тези акаунти често се осъществява след човешка намеса, като в голяма част от случаите се доказва, че блокировките са резултат на погрешно действие от страна на AI.

Таблица 2: Блокирано съдържание във Фейсбук през 2023 г. Източник: META Transparency Center

Вид съдържание	Брой блокирания
Албуми	20
Коментари	13,4 М
Събития	93
Страници и групи	40,5 хил.
Постове	33,6 М
Профили	26,7 хил.

Таблица 3: Блокирано съдържание във Инстаграм през 2023 г. Източник: META Transparency Center

Вид съдържание	Брой блокирания
Акаунти	1,15 М
Медия	50 хил.

Тази ситуация създава особено уязвимо положение както за малките, така и за големите бизнеси. Внезапната загуба на основен маркетингов канал или онлайн присъствие може да доведе до значителен спад в приходите и до загуба на конкурентни предимства. За малките предприятия това може да бъде фатално, тъй като тяхната дейност често е силно зависима от онлайн маркетинга и взаимодействието с аудиторията. Освен това онлайн присъствието им често представлява ключов елемент от техните маркетингови и бизнес стратегии. Много от тези акаунти представляват резултат от дългогодишни усилия и сериозни финансови инвестиции за изграждане на онлайн присъствие, което често включва разходи за реклами, маркетингови кампании и взаимодействие с потребителите. Когато тези акаунти бъдат неправилно премахнати, финансовите загуби могат да достигнат десетки хиляди долари, а възстановяването на изгубените потребители и репутация е труден и отнемаш време процес.

Възниква и проблемът с репутацията. Неправомерните блокирания на акаунти могат да се възприемат като липса на професионализъм или проблеми в комуникацията с клиентите, особено ако тези проблеми не бъдат адресирани своевременно. В резултат, възстановяването на репутацията и доверието на потребителите може да изисква значителни усилия, включително засилени маркетингови кампании и по-активно взаимодействие с аудиторията. Тези усилия не само са свързани с високи разходи, но и с необходимостта от дългосрочно планиране и адаптация към динамичната среда на дигиталния маркетинг.

Как функционира технологията за следене на нарушения на политиките на Meta в съдържанието

Алгоритмите за модерирание на съдържание във Facebook, които използват изкуствен интелект (AI) и машинно обучение (ML), функционират чрез сложен процес на анализ на голямо количество данни в реално време. Алгоритмите използват предварително обучени модели, които са запазени с огромни масиви от данни, включващи текстове, изображения и видеа. Тези модели се обучават да разпознават модели и специфични характеристики на съдържанието, които са свързани с нарушаване на политиките на платформата:

- 1. Анализ на текст:** Алгоритмите за обработка на естествен език (*NLP*) анализират всяка дума и израз в текстовото съдържание. Това включва разглеждане на семантични и синтактични структури, разпознаване на ключови думи и анализ на контекста, за да се открие дали текстът нарушава политиките. Сложността възниква, когато алгоритмите трябва да разпознаят и различават езикови нюанси като ирония, сарказъм или метафори.
- 2. Анализ на изображения и видеоклипове:** Алгоритмите за компютърно зрение използват модели на дълбоко обучение (*deep learning*) за разпознаване на обекти, сцени и действия в изображения и видеоклипове. Те могат да засичат конкретни визуални елементи – например оръжия, насилие, измама или символи на омразата. Проблеми възникват, когато съдържанието не е ясно дефинирано или едвам уловимо, както е в случаите с изображения, съдържащи хумор или артистични творби.
- 3. Комбиниран анализ:** За видеоклипове се комбинира и текстовият анализ (чрез разпознаване на говор) с визуалния анализ. Това позволява по-добра категоризация на съдържанието, но все пак не премахва риска от погрешно тълкуване на контекста.

Предизвикателства пред системите за проследяване на съдържание на групата на META

Въпреки че тези модели са впечатляващи по своите възможности, автоматизираните системи все още се сблъскват с няколко ключови ограничения:

1. Трудност при разпознаване на думи в контекст: Един от основните проблеми на тези системи е способността им да разберат пълния контекст на дадено съдържание. Похвати като хумора, сатирата или културните особености на даден израз могат да бъдат трудни за правилно тълкуване. Алгоритмите са склонни да действат строго по шаблони, което води до ситуации, в които безобидни изрази или „мемета“ могат да бъдат неправилно интерпретирани като нарушение на правилата.

2. Отсъствие на емоционален или социален интелект: Алгоритмите нямат способността да разпознават нюанси и човешки емоции, което е особено важно при оценка на чувствителни теми като насилие или реч на омразата. Те не могат да оценят социалния или културен контекст на съдържанието, което често води до *фиктивни позитиви* – премахване на съдържание, което всъщност не нарушава политиките на Facebook.

3. Ограничена възможност за динамично обучение: Въпреки че машинното обучение се стреми да адаптира алгоритмите към нова информация, процесът не е инстантен. Това означава, че при големи промени в обществото или в начина, по който хората използват езика или културата, тези системи могат да останат назад и да не успеят да разберат нововъзникващи термини или контексти.

Решения за справяне с предизвикателствата, наложени от автоматизираните системи за следене на съдържание, нарушаващо политиките на META.

За да се справи с тези предизвикателства, Facebook често комбинира автоматизирани системи за модериране с човешки модератори. В случаи на спорно съдържание, където алгоритъмът не може да вземе категорично решение или има висок риск от допускане на грешка, съдържанието се предава на екип от обучени модератори. Този хибриден подход има за цел да намали броя на грешките и да подобри точността на модерирането, като вземе предвид нюансите и контекста на всяко съдържание.

Предпазването от негативните последици, произтичащи от блокиране, ограничаване или изтриване на бизнес профил в платформите на Meta (Facebook и Instagram), изисква стратегически и технически мерки. Бизнесите могат да минимизират риска и да защитят своето онлайн присъствие чрез:

1. Следване на правилата и актуализация на политиките

Едно от най-важните неща, които бизнесите могат да направят, е да се уверят, че съдържанието и рекламите, които публикуват, стриктно спазват политиките на Meta. Facebook и Instagram редовно актуализират своите правила, затова е важно да се следят промените и да се адаптира съдържанието спрямо тях. Компаниите трябва да обучат своите екипи да познават добре правилата на платформите и да избягват рискови публикации, които могат да доведат до санкции.

2. Използване на резервни канали и мултиплатформена стратегия

Едно от най-добрите решения за минимизиране на загубите при блокиране на профил в Meta е изграждането на мултиплатформено присъствие. Бизнесите не трябва да разчитат само на една платформа, а да диверсифицират онлайн присъствието си, посредством използването на други канали като Twitter, LinkedIn, TikTok, Google ads и имейл маркетинг. Това гарантира, че дори при

блокиране или изтриване на профил, те ще запазят връзка с аудиторията си и ще могат да продължат своята маркетингова и комуникационна дейност.

3. Редовни архиви и бекъп на данни

За да предотвратят загубата на съдържание или достъп до важна информация, компаниите трябва редовно да архивират своите публикации, контакти с клиенти и анализи от платформите на Meta. Това включва данни за взаимодействия с аудиторията, анализи на кампании и съдържание, публикувано в профилите. По този начин, в случай на блокиране, ще бъде по-лесно да се възстанови активността и да се възстанови загубената информация.

4. Свързване с Meta Support и установяване на контакт с представител

При възникване на проблеми, свързани с блокиране или ограничаване на акаунти, бизнесите следва да използват **Meta Business Help Center** за достъп до техническа поддръжка и подаване на обжалвания срещу решенията, взети от автоматизирани модерационни системи. В този процес е от съществено значение внимателното представяне на аргументи в подкрепа на апелацията, тъй като платформата разчита на сложни алгоритми, които не винаги отчитат контекста на съдържанието.

За по-големите организации или рекламодатели с висок рекламен бюджет е препоръчително да се осъществи директна връзка с представител на Meta. Този персонализиран контакт може да осигури ускорен процес на разрешаване на проблеми, както и по-целенасочена поддръжка, което е от особено значение в случаите на значително засягане на бизнес дейността. Взаимодействието с представител на Meta не само ускорява решаването на възникнали проблеми, но и предоставя възможност за по-тясно сътрудничество при оптимизирането на рекламните кампании и предотвратяване на бъдещи ограничения.

5. Проследяване и анализ на проблемни публикации

Някои автоматизирани системи на Meta може да санкционират публикации, които съдържат определени думи, изрази или изображения. Затова е важно бизнесите да проследяват и анализират кои публикации водят до рестрикции, и да ги адаптират за бъдещи кампании. Инструменти като Facebook's Ad Library също могат да помогнат за проучване на най-ефективните реклами и избягване на съдържание, което е податливо на блокиране.

Заключение

Автоматизирани системи за модерирание на съдържание в платформите на социалните мрежи, като тези, използвани от Meta (Facebook и Instagram), предлагат значителни предимства по отношение на скорост и мащаб на обработка на данни. Въпреки това, тези системи имат и съществени недостатъци, които могат да окажат негативно влияние върху бизнеса и онлайн присъствието на потребителите. Анализът на тези ефекти води до следните основни изводи и препоръки:

1. Неправомерното блокиране на акаунти и страници е един от основните проблеми, произтичащи от автоматизираната модерация. Бизнеси с дългогодишни профили и инвестиции в рекламни кампании са изправени пред риска от неочаквани рестрикции, което може да доведе до загуба на последователи и репутация.

2. Липсата на яснота и прозрачност в процеса на блокиране и ограничаване създава трудности за бизнесите, които не получават точна информация относно причините за блокирането. Това възпрепятства възможността за бърза реакция и корекция на грешките.

3. Автоматизирани алгоритми за съдържание, базирани на изкуствен интелект, често не могат да разграничат контекстуални разлики в публикациите. Това води до погрешно тълкуване на съдържание, което не нарушава правилата на платформата, и до последващи рестрикции.

4. Финансовите загуби за бизнеса могат да бъдат значителни, особено за малки и средни компании, които разчитат на социалните мрежи за привличане на клиенти и изграждане на бранд идентичност. Възможността за измерване на ефективността на кампаниите в дигитални мрежи, която е основно предимство на онлайн рекламата, се обезсмисля, когато достъпът до акаунти е ограничен.

5. Сложността на обжалване на неправомерни блокирания допълнително утежнява ситуацията. Процесите на възстановяване на акаунти често изискват време и ресурси, което затруднява бизнесите в бързо възстановяване на техния онлайн статус и комуникационни канали.

6. Влияние върху потребителската увереност – честите грешки в модерацията намаляват доверието на потребителите в платформите, което може да доведе до промяна на предпочитанията им към алтернативни канали за комуникация и реклама.

7. Необходимост от усъвършенстване на модерационните системи – ключова препоръка е въвеждането на по-добри механизми за ръчен преглед на спорни случаи и прозрачни обжалващи процеси. Това би намалило вероятността от неправомерни блокирания и би предоставило на бизнесите по-сигурна и ефективна среда за развитие.

В заключение, макар автоматизацията в модерирането да предоставя значителни възможности за социалните мрежи да обработват голямо количество данни, е наложително тези системи да бъдат допълнени с по-точни и прозрачни процеси. Балансът между автоматизирани решения и човешка намеса е от решаващо значение за намаляване на негативните последици върху бизнесите и техния онлайн имидж.

References

Content Restrictions Based on Local Law. (2023). Извлечено от META Transparency Centre: <https://transparency.meta.com/reports/content-restrictions/>

DataReportal – Global Digital Insights. (October 2024 r.). Извлечено от DataReportal : <https://datareportal.com/global-digital-overview>

How Meta enforces its policies. (н.д.). Извлечено от Transparency Centre : <https://transparency.meta.com/bg-bg/enforcement/>

META Investor Relations . (2023). Извлечено от <https://investor.fb.com/investor-news/press-release-details/2023>

Policy . (н.д.). Извлечено от Transparency Center: <https://transparency.meta.com/bg-bg/policies/>