

Sprawozdanie dotyczące moderacji treści na Forum portalu Bankier.pl

Niniejsze sprawozdanie jest publikowane w związku z obowiązkami sprawozdawczymi w zakresie przejrzystości wprowadzonymi przez Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych)

Rozporządzenie zobowiązuje platformy do zapewnienia przejrzystości dotyczącej podejmowanych działań. Poniżej znajdują się kluczowe statystyki dotyczące zgłoszeń oraz podjętych działań moderacyjnych.

Okres sprawozdania: 02.03.2024-31.12.2024

Polityka i procedury moderacji treści

1. Cel moderacji

Moderacja treści na platformie ma na celu zapewnienie bezpiecznej i merytorycznej wymiany poglądów w zgodzie z **obowiązującym prawem** oraz **regulaminem platformy**. Proces moderacji obejmuje zarówno działania **automatycznych narzędzi**, jak i **interwencje moderatorów**.

2. Mechanizmy moderacji

- **Automatyczna moderacja** – wykorzystuje algorytmy sztucznej inteligencji do wykrywania i usuwania treści naruszających zasady. Decyzje te podlegają wyrywkowej kontroli moderatorów.
- **Moderacja ręczna** – moderatorzy analizują zgłoszenia użytkowników oraz podejmują działania wobec treści naruszających regulamin lub przepisy prawa.

3. Zgłaszanie nieodpowiednich treści

Użytkownicy mogą zgłaszać treści naruszające regulamin lub przepisy prawa.

- Zgłoszenia weryfikowane są w sposób terminowy i obiektywny.
- Decyzje o usunięciu treści są uzasadniane, a użytkownicy informowani o możliwości odwołania.
- Zgłoszenia od zaufanych podmiotów sygnalizujących rozpatrywane są priorytetowo.

4. Nakazy organów państwowych

Platforma realizuje nakazy organów państw członkowskich zgodnie z regulacjami DSA.

- Każdy nakaz podlega weryfikacji prawnej, a w razie niejasności moderatorzy mogą żądać wyjaśnień.
- W przypadku decyzji o usunięciu treści użytkownicy są informowani, jeśli platforma dysponuje ich danymi kontaktowymi.

5. Odwołania i system skarg

Użytkownicy mogą składać skargi na decyzje moderacyjne poprzez dostępne mechanizmy.

- Wewnętrzny system skarg działa zgodnie z regulaminem serwisów Regulamin serwisów internetowych Bonnier Business Polska.
- Zgłaszający oraz użytkownicy, których treści zostały usunięte, otrzymują informacje o podjętych działaniach i możliwościach odwołania.

6. Ograniczenia dostępu do platformy

- Użytkownicy, którzy nadużywają mechanizmów zgłoszeń lub publikują rażąco nielegalne treści, mogą podlegać sankcjom, w tym czasowemu lub bezterminowemu zawieszeniu dostępu.

Dzięki zastosowaniu nowoczesnych narzędzi i przejrzystych procedur, platforma dba o jakość treści i przestrzeganie regulacji prawnych oraz wewnętrznych zasad.

Wyjaśnienie powodów zgłoszeń i skarg na decyzje:

- **Spam** – autor wielokrotnie publikuje identyczne lub bardzo podobne treści w krótkim czasie.
- **Wulgaryzmy** – treść zawiera obraźliwe lub niecenzuralne słowa.
- **Nie na temat** – wpis nie dotyczy tematu dyskusji lub jest niepowiązany z tematem forum
- **Obrażanie** – wpis atakuje, poniża lub obraża inne osoby.
- **Podbija wątek** – autor celowo dodaje wpisy bez merytorycznej treści, aby przesunąć temat na górę listy.
- **Dobra osobiste** – treść narusza prywatność, godność lub prawa osobiste innych osób.
- **Fałszywe informacje** – wpis zawiera nieprawdziwe lub wprowadzające w błąd treści.
- **Reklama** – treść służy do promocji produktów, usług lub firm, bez związku z tematem dyskusji.
- **Dane osobowe/teleadresowe** – wpis zawiera prywatne informacje (np. adres, numer telefonu) bez zgody osoby, której dotyczą.
- **Samobójstwo** – treść zawiera zachęty do samookaleczenia lub samobójstwa.
- **Prawa autorskie** – wpis narusza prawa własności intelektualnej, np. poprzez publikację chronionych materiałów bez zgody.
- **Nawoływanie do przestępstwa** – treść zachęca do łamania prawa, np. oszustw lub innych działań niezgodnych z prawem.
- **Dyskryminacja** – wpis zawiera treści dyskryminujące ze względu na rasę, płeć, orientację seksualną, religię itp.
- **Nazizm i komunizm** – treść propaguje ideologie totalitarne, symbole lub hasła powiązane z nazizmem lub komunizmem.
- **Przemoc** – wpis zawiera groźby, opisy brutalnych scen lub promuje przemoc wobec innych osób.
- **Narusza moralność** – treść jest nieprzyzwoita, obraźliwa lub niezgodna z powszechnie przyjętymi normami etycznymi.
- **Podszywanie** – autor udaje inną osobę lub organizację w celu oszustwa lub wprowadzenia w błąd.
- **Wielonickowiec** – autor umieszcza wiele wpisów, podpisując się różnymi nickami, co może wprowadzać w błąd lub zakłócać dyskusję.
- **Uporczywe łamanie zasad** – autor wielokrotnie narusza regulamin, mimo wcześniejszych ostrzeżeń.

- **Nienawiść lub zagrożenie życia** – treść zawiera groźby śmierci, życzenia krzywdy wobec konkretnych osób lub grup.
- **Inne** – wpis narusza regulamin w sposób, który nie mieści się w żadnej z wymienionych kategorii.
- **Tajemnica przedsiębiorstwa** – treść zawiera poufne informacje biznesowe, które nie powinny być publicznie udostępniane.
- **Rozporządzenie MAR** – wpis narusza regulacje dotyczące rynków finansowych i obrotu papierami wartościowymi.
- **Informacja poufna** – treść zawiera tajne lub wrażliwe dane, które nie powinny być udostępniane publicznie.
- **Naruszenie przepisów** – wpis łamie obowiązujące przepisy prawa.
- **Naruszenie regulaminu forum** – treść jest sprzeczna z zasadami obowiązującymi na platformie.
- **Zgłoszenie uprawnionego** – wpis został zgłoszony przez osobę lub instytucję mającą do tego formalne uprawnienia.

- **Usunięcie wpisu** – użytkownik złożył skargę, uznając, że jego wpis został niesłusznie usunięty, mimo że według niego nie naruszał regulaminu platformy ani obowiązujących przepisów.
- **Odrzucenie prośby o usunięcie wpisu** – prośba użytkownika o usunięcie treści została odrzucona, ponieważ wpis nie naruszał zasad platformy ani przepisów prawa.
- **Tymczasowe zawieszenie usługi pisania** – użytkownik otrzymał tymczasowy zakaz publikowania treści na platformie z powodu naruszenia regulaminu.
- **Bezterminowe zawieszenie usługi pisania** – użytkownik otrzymał stały zakaz publikowania treści na platformie z powodu poważnych lub powtarzających się naruszeń regulaminu.

Zgłoszenia do moderacji

Użytkownicy zgłaszają treści, które ich zdaniem naruszają regulamin platformy. Moderacja analizuje zgłoszenia i podejmuje decyzje o ich uznaniu lub odrzuceniu.

Tabela 1: Podsumowanie zgłoszeń do moderacji

Typ zgłoszenia	Liczba zgłoszeń	% zgłoszeń
Uznane zgłoszenia	27002	28.67%
Odrzucone zgłoszenia	67174	71.33%
łącznie	94176	100.00%

Podsumowanie:

Tylko 28.67% zgłoszeń skutkowało uznaniem i podjęciem działań, co oznacza, że większość zgłoszonych treści (71.33%) została odrzucona.

Może to oznaczać, że nie wszystkie zgłoszenia użytkowników są zasadne lub że platforma stosuje **wysokie standardy weryfikacji treści** przed ich usunięciem.

Łączna liczba zgłoszeń wynosi 94 176, co potwierdza dużą aktywność użytkowników w procesie moderacji.

Rekomendacje:

Użytkownicy powinni lepiej rozumieć, jakie treści rzeczywiście naruszają regulamin, aby zgłoszenia były bardziej trafne.

Warto przeanalizować, czy zgłoszenia są odrzucane słusznie, czy może wymagane jest dostosowanie kryteriów moderacyjnych.

Dzięki tym działaniom platforma może zwiększyć skuteczność moderacji i jednocześnie zapewnić lepszą przejrzystość zasad dla użytkowników.

Tabela 2: Podział usuniętych zgłoszeń do moderacji według powodów

Powód zgłoszenia	Łączna liczba zgłoszeń	% wśród usuniętych
Spam	7516	27.83%
Wulgaryzmy	3010	11.15%
Nie na temat	3937	14.58%
Obrażanie	6092	22.56%
Podbija wątek	740	2.74%

Dobra osobiste	1060	3.93%
Fałszywe informacje	339	1.26%
Reklama	146	0.54%
Dane osobowe/teleadresowe	268	0.99%
Samobójstwo	17	0.06%
Prawa autorskie	3	0.01%
Nawoływanie do przestępstwa	17	0.06%
Dyskryminacja	27	0.10%
Nazizm i komunizm	32	0.12%
Przemoc	37	0.14%
Narusza moralność	140	0.52%
Podszywanie	469	1.74%
Wielonickowiec	1450	5.37%
Uporczywe łamanie zasad	1503	5.57%
Nienawiść lub zagrożenie życia	35	0.13%
Inne	28	0.10%
Tajemnica przedsiębiorstwa	1	0.00%
Rozporządzenie MAR	42	0.16%
Informacja poufna	6	0.02%
Naruszenie przepisów	15	0.06%
Naruszenie regulaminu forum	62	0.23%
Zgłoszenie uprawnionego	10	0.04%

Podsumowanie:

Spam (27.83%) oraz obrażanie (22.56%) to najczęstsze powody usunięć zgłoszonych treści. **Nie na temat (14.58%) oraz wulgaryzmy (11.15%)** również stanowią istotny odsetek zgłoszeń.

Podszywanie (1.74%) oraz wielonickowiec (5.37%) wskazują na nadużycia w tworzeniu fałszywych kont.

Zgłoszenia dotyczące poważnych naruszeń, takich jak przemoc (0.14%) czy nawoływanie do przestępstwa (0.06%), występują rzadziej, ale wymagają szczególnej uwagi.

Rekomendacja:

Spam pozostaje dominującym problemem – warto poprawić algorytmy automatycznego wykrywania i weryfikacji.

Znaczny odsetek zgłoszeń wynika z publikowania treści niezwiązanych z tematyką. Może być konieczne ulepszenie mechanizmów kategoryzacji treści i powiadamianie użytkowników o zasadach przed publikacją.

Tabela 3: Podział usuniętych, zgłoszonych wpisów według podstawy prawnej

Podstawa usunięcia	Liczba usuniętych wpisów	% Wśród wszystkich usuniętych wpisów
Przepisy prawa	9 505	35.21%
Regulamin platformy	17 497	64.79%
Łącznie	27 002	100.00%

Podsumowanie:

35.21% wpisów zostało usuniętych z powodu naruszenia przepisów prawa, co oznacza, że ich treść łamała regulacje prawne.

64.79% wpisów zostało usuniętych na podstawie warunków korzystania z usług dostawcy (regulaminu), co obejmuje reklamy, wulgaryzmy i inne naruszenia wewnętrznych zasad.

Powiadomienia użytkowników o decyzjach i proces odwoławczy

Zarejestrowani użytkownicy są automatycznie powiadamiani o decyzjach moderacyjnych. W przypadku usunięcia treści, ograniczenia jej widoczności lub innej interwencji, użytkownik otrzymuje informację o podjętej decyzji oraz jej uzasadnienie.

Każdy zarejestrowany użytkownik ma prawo do odwołania się od decyzji poprzez opcję "Zgłoś skargę", co zapewnia mu dostęp do **skutecznego mechanizmu apelacyjnego**.

Niezarejestrowani użytkownicy nie otrzymują powiadomień, ponieważ platforma nie posiada ich danych kontaktowych. **Mimo to, nadal mogą skorzystać z procedury odwoławczej**, korzystając z opcji "Zgłoś skargę" dostępnej na platformie.

Zgodność z DSA

- Powiadamianie zarejestrowanych użytkowników i możliwość odwołania spełnia wymogi **Artykułu 15 oraz Artykułu 24 (1c)**.
- Udostępnienie opcji odwoławczej także dla niezarejestrowanych użytkowników zapewnia zgodność z zasadą **przejrzystości i dostępności mechanizmu skargowego**.

Skargi na decyzje moderacyjne

Tabela 4: Podsumowanie skarg na decyzje

Typ skargi	Liczba skarg	% Skarg
Uznane	27	15.88%
Odrzucone	143	84.12%
łącznie	170	100.00%

Podsumowanie:

- **Tylko 15.88% skarg zostało uznanych**, co sugeruje, że większość decyzji moderacyjnych była poprawna.
- **Odrzucono 84.12% skarg**, co może oznaczać, że choć użytkownicy kwestionują decyzje moderacji, to większość tych decyzji była zgodna z regulaminem.

Rekomendacja:

- **Zwiększenie transparentności procesu odwoławczego**, aby użytkownicy lepiej rozumieli powody decyzji.
- **Lepsza komunikacja uzasadnień odrzuconych skarg**, aby zmniejszyć liczbę bezzasadnych apelacji.

Tabela 5: Podział skarg na decyzje według powodów

Powód skargi	Uznane skargi	Odrzucone skargi	Łączna liczba skarg	% Uznanych	% Odrzuconych
Usunięcie wpisu	23	104	127	18.11%	81.89%
Odrzucenie prośby o usunięcie wpisu	3	13	16	18.75%	81.25%
Tymczasowe zawieszenie usługi pisania	1	8	9	11.11%	88.89%
Bezterminowe zawieszenie usługi pisania	0	18	18	0.00%	100.00%

Podsumowanie:

- **Najwięcej skarg dotyczyło usunięcia wpisu (127 przypadków), z czego tylko 18.11% zostało uznanych.**
- **Skargi na odrzucenie prośby o usunięcie wpisu (16 przypadków) miały podobny odsetek uznań – 18.75%.**
- **Skargi na zawieszenie konta miały najmniejszą szansę powodzenia, a w przypadku bezterminowego zawieszenia konta żadna skarga nie została uznana.**

Rekomendacja:

- **Dokładna analiza skarg na zawieszenie konta, ponieważ 100% takich odwołań zostało odrzuconych.**
- **Lepsza informacja dla użytkowników na temat powodów usunięcia wpisów, aby ograniczyć liczbę skarg.**

Automatyczna moderacja treści

Platforma korzysta z **zaawansowanego algorytmu sztucznej inteligencji**, który **regularnie podlega udoskonaleniom**, aby jak najefektywniej wykrywać **nielegalne treści oraz naruszenia regulaminu**.

W ramach moderowania treści wykorzystywane są zaawansowane algorytmy sztucznej inteligencji, oparte o model językowy ChatGPT. Algorytm ten wykorzystuje metody głębokiego uczenia (deep learning), a dokładniej architekturę transformera, która umożliwia rozpoznawanie kontekstu oraz semantyczne analizowanie treści. Prompty używane do interakcji z modelem generowane są w oparciu o regulamin serwisu Bankier.pl oraz aktualne przepisy prawa, co pozwala na precyzyjne wychwytywanie naruszeń.

Skuteczność algorytmów regularnie jest analizowana na podstawie bieżących obserwacji rezultatów moderacji. Metody doskonalenia obejmują cykliczną optymalizację promptów oraz wprowadzanie korekt do ustawień systemu, co przekłada się na wyższą precyzję, szybszą reakcję na naruszenia oraz zmniejszenie liczby fałszywych trafień (false positives). Proces ten zapewnia ciągłe zwiększanie efektywności moderowania treści zgodnie z wymogami Digital Services Act.

Algorytm jest dostosowany do wykrywania **różnych kategorii naruszeń**, takich jak:

- Spam
- Wulgaryzmy
- Obrażanie
- Podbijania wątków
- Wielonickowców

System ten działa **proaktywnie**, wykrywając naruszenia nawet przed ich zgłoszeniem przez użytkowników. **Dzięki regularnym aktualizacjom** algorytm dostosowuje się do **nowych form nadużyć**, co zwiększa jego skuteczność.

Tabela 6: Usunięte wpisy przez automatyczny system

Powód usunięcia	Usunięte automatycznie	% Usuniętych automatycznie
Spam	76361	27.32%
Wulgaryzmy	24208	8.67%
Nie na temat	21811	7.81%
Obrażanie	68463	24.51%
Podbija wątek	5480	1.96%
Przemoc	1818	0.65%
Narusza moralność	187	0.07%
Wielonickowiec	81095	29.02%

Podsumowanie:

- **Najwięcej wpisów usunięto z powodu wielonickowców (29.02%),** co wskazuje na częste nadużycia w tym zakresie.
- **Spam (27.32%) i obrażanie (24.51%)** to kolejne główne powody interwencji automatycznych systemów.
- **Pozostałe powody, takie jak przemoc czy naruszenie moralności, stanowią niewielką część usuniętych treści.**
- Łącznie w ramach automatycznej moderacji treści usunięto: **279 423 wpisów**

Rekomendacja:

- **Optymalizacja systemu wykrywania wielonickowców,** aby ograniczyć nadużycia.
- **Poprawa algorytmów wykrywających obraźliwe treści,** aby zmniejszyć liczbę przypadków wymagających interwencji ręcznej.

Tabela 7: Przywrócone wpisy po automatycznej moderacji

Kategoria	Liczba wpisów	% Wśród usuniętych wpisów
Usunięte przez automatyczny system	279 423	100.00%
Przywrócone przez moderatorów	34 361	12.30%
Łącznie	279 423	100.00%

Podsumowanie:

12.30% wpisów usuniętych przez automatyczny system zostało przywróconych po ręcznej weryfikacji moderatorów, co oznacza, że system mógł błędnie sklasyfikować część treści jako naruszające zasady.

Pozostałe 87.70% wpisów zostało uznane za prawidłowo usunięte, co wskazuje na wysoką skuteczność systemu automatycznej moderacji.

Rekomendacje:

Dalsze doskonalenie algorytmów automatycznej moderacji, aby zminimalizować przypadki błędnego usunięcia treści.

Dokładniejsza analiza przywracanych wpisów, aby określić najczęstsze przyczyny błędnej klasyfikacji i dostosować mechanizmy wykrywania naruszeń.

Ręczne wykrywanie naruszeń

Tabela 8: Usunięte wpisy przez moderatorów w ramach działań własnych

Powód usunięcia	Usunięte przez moderatorów	% Usuniętych przez moderatorów
Spam	108565	67.57%
Wulgaryzmy	8150	5.07%
Nie na temat	6432	4.00%
Obrażanie	29635	18.45%
Podbija wątek	2326	1.45%
Dobra osobiste	1702	1.06%
Fałszywe informacje	126	0.08%
Reklama	284	0.18%
Dane osobowe/teleadresowe	226	0.14%
Samobójstwo	4	0.00%
Prawa autorskie	5	0.00%
Nawoływanie do przestępstwa	12	0.01%
Dyskryminacja	27	0.02%
Nazizm i komunizm	27	0.02%
Przemoc	66	0.04%
Narusza moralność	313	0.19%
Podszywanie	499	0.31%
Wielonickowiec	175	0.11%
Uporczywe łamanie zasad	1709	1.06%
Nienawiść lub zagrożenie życia	57	0.04%
Inne	270	0.17%
Tajemnica przedsiębiorstwa	1	0.00%
Rozporządzenie MAR	18	0.01%
Informacja poufna	1	0.00%
Naruszenie przepisów	4	0.00%
Naruszenie regulaminu forum	23	0.01%
Zgłoszenie uprawnionego	8	0.01%

Podsumowanie

- Najwięcej wpisów usunięto z powodu **spamu (67.57%)** oraz **obrażania (18.45%)**.
- Wpisy dotyczące **wulgaryzmów i treści nie na temat** stanowią łącznie około **9% usuniętych wpisów**.
- Inne powody, takie jak **dane osobowe, podszywanie i wielonickowiec**, mają stosunkowo niski udział w całkowitej liczbie usuniętych treści.
- Łącznie w ramach działań własnych moderatorów usunięto: **160 665 wpisów**

Rekomendacje:

- **Weryfikacja procesu moderacji w zakresie spamu**, który dominuje w ręcznych interwencjach.
- **Skrócenie czasu reakcji moderatorów na obrażające treści**, które stanowią znaczną część usuniętych wpisów.
- **Lepsze algorytmy do rozpoznawania wulgaryzmów i treści obraźliwych**, aby zmniejszyć liczbę koniecznych interwencji moderatorów.

Tabela 9: Podział usuniętych wpisów w ramach działań własnych platformy według podstawy prawnej

Podstawa usunięcia	Liczba usuniętych wpisów	% Wśród wszystkich usuniętych wpisów
Przepisy prawa	189 488	43.05%
Regulamin platformy	250 600	56.95%
łącznie	440 088	100.00%

Podsumowanie:

43.05% wpisów (189 488) zostało usuniętych z powodu naruszenia przepisów prawa, co oznacza, że ich treść łamała regulacje prawne.

56.95% wpisów (250 600) zostało usuniętych na podstawie warunków korzystania z usług dostawcy (regulaminu), co obejmuje naruszenia wewnętrznych zasad, takich jak reklamy, wulgaryzmy i inne naruszenia wewnętrznych zasad.

Tabela obejmuje wpisy usunięte zarówno przez automatyczny system moderacji (279 423 wpisów), jak i przez działania własne moderatorów (160 665 wpisów).

łącznie usunięto 440 088 wpisów, co odzwierciedla skalę działań moderacyjnych i dbałość platformy o przestrzeganie regulaminu oraz zgodność z obowiązującym prawem.

Zgłoszenia do moderacji od uprawnionych organów i zaufanych podmiotów

W analizowanym okresie **nie odnotowano żadnych zgłoszeń do moderacji treści** od:

- **Uprawnionych organów państwowych** – nie wpłynęły żadne nakazy dotyczące usunięcia lub ograniczenia dostępu do treści zgodnie z regulacjami DSA.
- **Zaufanych podmiotów sygnalizujących** – nie wpłynęły żadne zgłoszenia, które zostałyby oznaczone jako priorytetowe na podstawie Artykułu 22 DSA.

Brak takich zgłoszeń oznacza, że w analizowanym okresie **moderacja treści opierała się wyłącznie na zgłoszeniach użytkowników, działaniach automatycznych systemów moderacyjnych oraz interwencjach zespołu moderatorów.**

Podział moderacji na proaktywną i reaktywną

Proaktywna moderacja:

- **Treści usunięte przez moderatorów oraz usunięte przez automatyczny system moderacji** są wynikiem **proaktywnej moderacji**, czyli działań podjętych z własnej inicjatywy platformy, bez konieczności zgłoszenia przez użytkowników.
-

Reaktywna moderacja:

- **Zgłoszenia użytkowników do moderacji** to moderacja **reaktywna**, czyli decyzje podjęte w odpowiedzi na zgłoszenia użytkowników. Wpisy analizowane w tym procesie mogą zostać usunięte przez moderatorów po przeprowadzeniu oceny naruszenia zasad.

Zgodność z DSA

- Raport jasno rozróżnia działania **proaktywne i reaktywne**, zgodnie z wymogami **Artykułu 24 (1a) i (1b)**.
- Działania moderacyjne są **przejrzyste**, a użytkownicy mają dostęp do systemu skarg i odwołań.

Czas reakcji na zgłoszenia

Czas reakcji na zgłoszenia jest kluczowym wskaźnikiem efektywności reakcji zespołu moderacji na zgłoszenia zewnętrzne. W analizowanym okresie **mediana** czasu reakcji dla poszczególnych kategorii zgłoszeń wynosiły:

- **Zgłoszenia dotyczące naruszenia przepisów prawa – 34 minuty**
- **Zgłoszenia dotyczące naruszenia regulaminu serwisu – 32 minuty**
- **Zgłoszenia skarg na decyzje moderacyjne – 4 dni, 12 godzin i 36 minut**

Dane te wskazują, że zgłoszenia związane z naruszeniem przepisów prawa oraz regulaminu platformy są rozpatrywane w relatywnie krótkim czasie, co pozwala na szybkie podejmowanie działań w przypadku treści wymagających interwencji. Proces rozpatrywania skarg na decyzje moderacyjne jest bardziej złożony i wymaga dodatkowej weryfikacji, co przekłada się na dłuższy czas reakcji.

Platforma nieustannie pracuje nad optymalizacją procesów moderacji, aby zapewnić **jeszcze szybsze i bardziej efektywne przetwarzanie zgłoszeń, jednocześnie zachowując wysoką jakość i rzetelność podejmowanych decyzji.**

Podsumowanie

- **Automatyczna moderacja skutecznie eliminuje większość spamu, obrażania i wielonickowców.**
- **Ręczna moderacja często podtrzymuje swoje decyzje, co widać po niskim odsetku uznanych skarg.**
- **Czas reakcji jest bardzo szybki w przypadku zgłoszeń naruszeń, ale dłuższy w skargach na decyzje.**
- **Platforma korzysta z zaawansowanej sztucznej inteligencji do moderacji treści, co spełnia wymogi DSA w zakresie skutecznej i dynamicznej ochrony przed nielegalnymi treściami.**
- **Użytkownicy zarejestrowani są powiadamiani o decyzjach i mogą składać odwołania poprzez system skarg.**
- **Niezarejestrowani użytkownicy także mają dostęp do odwołania się, mimo braku powiadomień.**
- **Moderacja jest podzielona na proaktywną (inicjowaną przez platformę) i reaktywną (opartą na zgłoszeniach użytkowników), co zapewnia pełną zgodność z regulacją DSA.**

Rekomendacje ogólne:

- **Poprawa wykrywania wielonickowców i spamu** – są to najczęstsze naruszenia.
- **Lepsza komunikacja odrzuconych zgłoszeń i skarg**, aby zmniejszyć liczbę nieuzasadnionych odwołań.
- **Szybsza moderacja ręczna dla skarg na decyzje**, aby użytkownicy szybciej otrzymywali odpowiedzi.

Raport zgodny z regulacją DSA, Artykuł 15 oraz Artykuł 24 (1a, 1b).