FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Filtr nevhodného textového obsahu pro herní server s využitím strojového učení

Mikhail Belov
German Semin
+420 702 052 432
+420 705 923 370
mikhail.belov@tul.cz
german.semin@tul.cz

Filtr nevhodného textového obsahu pro herní server

s využitím strojového učení

Uvod

- Discord je klíčová komunikační platforma pro herní komunity.
- Automatická moderace je často nedostatečná – buď příliš přísná, nebo neefektivní.
- Porušení pravidel Discordu (např. toxický obsah) může vést k trvalému zrušení serveru.
- Zablokování serveru = komerční riziko (ztráta hráčů, výpadek mikrotransakcí).
- Chybí nástroj, který by moderaci řešil spolehlivě a efektivně.

Cíle projektu

- Automaticky rozpoznávat toxické zprávy v prostředí Discordu v reálném čase.
- Minimalizovat manuální práci moderátorů a snížit jejich zátěž.
- Omezit riziko blokace Discord serverů ze strany platformy kvůli porušení pravidel.
- Zvýšit bezpečnost a kvalitu komunikace v herních komunitách (např. Minecraft a DayZ).
- Porovnat různé přístupy k detekci toxicity a najít nejefektivnější řešení pro praktické nasazení.

Sběr a příprava dat

- Zdroje: Discord zprávy (s využitím bota, discord.py) + existující Kaggle dataset
- Anotace zprav: 0 = normální, 1 = toxická (ručně)
- Poměr tříd: ~60% normální / 40% toxické
- Čištění: odstranění systémových zpráv, normalizace textu
- Rozdělení: 80 % trénink, 20 % test
- Celkem: 24 412 zpráv (40 % toxických)

Použité technologie

redis

- **Python – pro zpracování zpráv, trénování modelů a analýzu dat**
- **discord.py – získávání zpráv a integrace moderátorského bota**
- **ruBERT – předtrénovaný jazykový model pro detekci toxicity v ruštině**
- **PyTorch / Transformers – frameworky pro práci s neuronovými sítěmi**
- **Scikit-learn – vyhodnocení modelů a metriky klasifikace**
- **Pandas / NumPy – práce s datovými rámci a předzpracování dat**
- **Groq / LLaMA 3 – nasazení LLM pro srovnání s klasickými přístupy**
- **Redis – dočasné ukládání zpráv a výsledků klasifikace**
- **TypeScript – pro vývoj frontendu**

PyTorch

Přístupy k detekci

- Regulární výrazy (Regex)
- Předtrénovaný ruBERT-toxic
- Fine-tuned (vlastní dataset) ruBERT
- LLM llama-8b-8192 přes Groq API

Přístupy k detekci — výsledky

Tabulka 3.1: Výsledky klasifikace zpráv (Regex)

Třída	Přesnost	Záchyt	F1 skóre	Podpora
0 (normální)	0,73	0,96	0,83	2921
1 (nevhodná)	0,89	0,46	0,61	1962
Přesnost			0,76	4883
Makro průměr	0,81	0,71	0,72	4883
Vážený průměr	0,79	0,76	0,74	4883

Tabulka 3.2: Výsledky klasifikace pomocí předtrénovaného modelu ruBERT-toxic

Třída	Přesnost	Záchyt	F1 skóre	Podpora
0 (normální)	0,97	0,91	0,94	2921
1 (nevhodná)	0,88	0,96	0,92	1962
Přesnost			0,93	4883
Makro průměr	0,92	0,93	0,93	4883
Vážený průměr	0,93	0,93	0,93	4883

Tabulka 3.3: Výsledky klasifikace pomocí vlastního modelu ruBERT

Třída	Přesnost	Záchyt	F1 skóre	Podpora
0 (normální)	0,97	0,96	0,96	2921
1 (nevhodná)	0,93	0,96	0,95	1962
Přesnost			0,93	4883
Makro průměr	0,95	0,96	0,95	4883
Vážený průměr	0,96	0,96	0,96	4883

Tabulka 3.4: Výsledky klasifikace pomocí Groq (llama3-8b-8192)

Třída	Přesnost	Záchyt	F1 skóre	Podpora
0 (normální)	0,91	0,92	0,92	2921
1 (nevhodná)	0,89	0,88	0,89	1962
Přesnost			0,67	4883
Makro průměr	0,90	0,90	0,90	4883
Vážený průměr	0,90	0,90	0,90	4883

Přístupy k detekci — analýza

- Regex: nízká úspěšnost (F1 = 0,61), nerozpozná skrytá nebo kreativní vulgarismy
- ruBERT-toxic: vysoká přesnost a recall (F1 = 0,92), hůře chápe herní slang a kontext
- ruBERT (vlastní): nejvyšší výkon (F1 = 0,95), vyžaduje anotaci a trénink modelu
- llama3: rozpozná i složité a kontextové případy, kategorizuje důvod toxicity, nižší přesnost než ruBERT, závislost na API

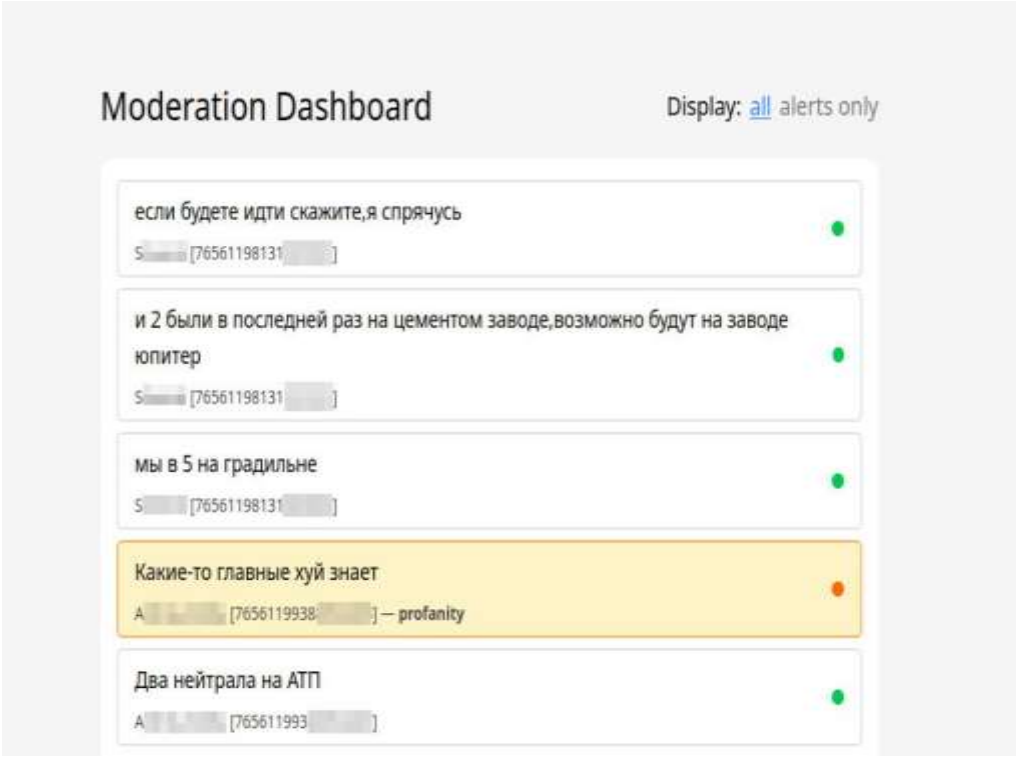
Filtr nevhodného textového obsahu pro herní server

s využitím strojového učení

Reálné nasazení

Slouží jako hlavní nástroj pro moderátory.

- Zobrazuje nové zprávy spolu s automatickou klasifikací („normální“ / „toxická“).
- Moderátor může ručně upravit hodnocení zprávy.
- Dashboard je napojený na Redis pro ukládání dat.
- Cílem je rychlá a přehledná práce bez nutnosti číst celý log.



Zaver

Plány do budoucna

- Vylepšení dashboardu — přehlednější zobrazení, filtrování podle typu toxicity, rychlejší reakce na nové zprávy.
- Pokračování ve sběru a anotaci specializovaných datasetů z RP komunity (např. DayZ, GTA RP).
- Testování filtrů v reálném provozu a jejich ladění na základě zpětné vazby od moderátorů.
- Rozšíření podpory o další jazyky nebo jazykové varianty (např. slang, transliterace).
- Nasazení pokročilého modelu přímo v rámci Discord bota pro automatické označování zpráv v reálném čase.
- Github

FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Děkujeme za pozornost!

Mikhail Belov

German Semin

+420 702 052 432

+420 705 923 370

mikhail.belov@tul.cz

german.semin@tul.cz