Vyšší odborná škola
a Střední průmyslová škola elektrotechnická,
Plzeň, Koterovská 85

Ročníková práce s obhajobou

Téma: SMDM - Social Media Data Model

Autor práce: Marek Ruttner
Třída: 3.L
Vedoucí práce: Jiří Švihla
Dne: 30. 4. 2024

Hodnocení:

Vyšší odborná škola a
Střední průmyslová škola elektrotechnická Plzeň,
Koterovská 85

ZADÁNÍ ROČNÍKOVÉ PRÁCE	
Školní rok	2023/ 2024
Studijní obor	78-42-M/01 Technické lyceum
Jméno a příjmení	Marek Ruttner
Třída	3. L
Předmět	Kybernetika
Hodnoceno v předmětu	Kybernetika
Téma	SMDM – Social Media Data Model
Obsah práce	1. Vytvoření aplikace pro shromažďování dat o tom, co ovlivňuje úspěch příspěvků na sociálních sítích. 2. Sběr dat s cílem vytvořit model, který bude předpovídat úspěch nových příspěvků. 3. Kombinace vlastního datového modelu a existujícího jazykového modelu s cílem poradit uživateli, jak upravit specifické části příspěvku a zvýšit jeho úspěšnost. 4. Poskytování konkrétních tipů uživateli, jak optimalizovat obsah příspěvku na základě analýzy datových faktorů. 5. Analýza úspěšnosti datového modelu
Zadávající učitel Příjmení, jméno	Švihla Jiří
Podpis zadávajícího učitele	
Termín odevzdání	30. dubna 2024

V Plzni dne: 30. 11. 2023

Mgr. Vlastimil Volák
ředitel školy

Anotace

Tato ročníková práce se zaměřuje na vývoj aplikace určené k shromažďování a analýze dat ze sociálních sítí X (dříve Twitter) či Threads, s cílem identifikovat faktory, které ovlivňují úspěch příspěvků. Klíčovým cílem je vytvořit datový model, který bude schopen predikovat úspěšnost nových příspěvků na základě analýzy existujících dat. Práce se dále zaměřuje na kombinaci tohoto datového modelu s existujícím jazykovým modelem, což umožní poskytovat uživatelům konkrétní rady, jak upravit a optimalizovat obsah jejich příspěvků pro dosažení lepších výsledků. V práci je také zahrnuta analýza efektivity a úspěšnosti vytvořeného datového modelu, což je klíčové pro ověření jeho praktické použitelnosti a spolehlivosti. Výsledkem této práce bude tedy nástroj, který umožní uživatelům sociálních sítí X a Threads přizpůsobit obsah jejich příspěvku k zvýšení impression a engagement.

Prohlášení

„Prohlašuji, že jsem tuto ročníkovou práci vypracoval samostatně a použil literárních pramenů a informací, které cituji a uvádím v seznamu použité literatury a zdrojů informací.“

V Plzni dne: Podpis:

Obsah

Úvod

V době digitální komunikace se sociální sítě staly klíčovým prostředkem pro sdílení informací, názorů a propagaci různých obsahů. Úspěch příspěvků na těchto platformách, zejména na sociálních sítích X (dříve Twitter) a Threads, je nejen indikátorem popularity, ale také významným faktorem při měření úspěšnosti marketingových a komunikačních strategií firem či osob. Tato ročníková práce se zaměřuje na vývoj a implementaci aplikace určené pro sběr a analýzu dat z těchto sociálních sítí, s cílem identifikovat klíčové faktory, které ovlivňují úspěch a viralitu příspěvků.

Cílem práce je prozkoumat možnosti pro předpovídání počtu interakcí s příspěvkem na základě obsahu příspěvku. K zjištění nejoptimálnějšího postupu, který by bylo možné na tu to problematiku využít, bude práce porovnávat několik způsobů. Od ruční datové analýzy až po využití velkých jazykových modelů (LLM), které se v dnešní době již objevují i na poli datové analýzy.

Kritérii pro porovnávání různých přístupů bude přesnost předpověděné hodnoty, rychlost potřebná k vytvoření predikce a možnost zapojení metody do dalších aplikací a případné využití dat k optimalizaci samotného modelu.

Práce také řeší sběr dat, které budou využity k predikci.

Pořadí	Téma
1.	Politika
2.	Zprávy
3.	Lifestyle
4.	Finance
5.	Kultura

Tabulka 1: Výsledky průzkumu nejsledovanějších témat na X

1 Sběr dat

Pro vytvoření statistiky či modelu, který bude schopný predikce počtu interakcí, je nutné sebrat dostatečné množství dat a tím vytvořit strukturovaná data, která jsou využitelná jak ve statistice, při použití vlastní neuronové sítě, tak pro dotrénováni (fine-tunning) LLM.

1.1 Výběr dat

Pro potřeby sběru dat byl na platformě X vytvořen zcela nový uživatelský účet, který není zatížen předchozí digitální stopou. Účet sleduje pouze účty, které tématicky odpovídají účtům, které označili osoby vyplňující dotazník, za nejvíce sledované a schopné nejvíce udržet uživatele u čtení příspěvků.

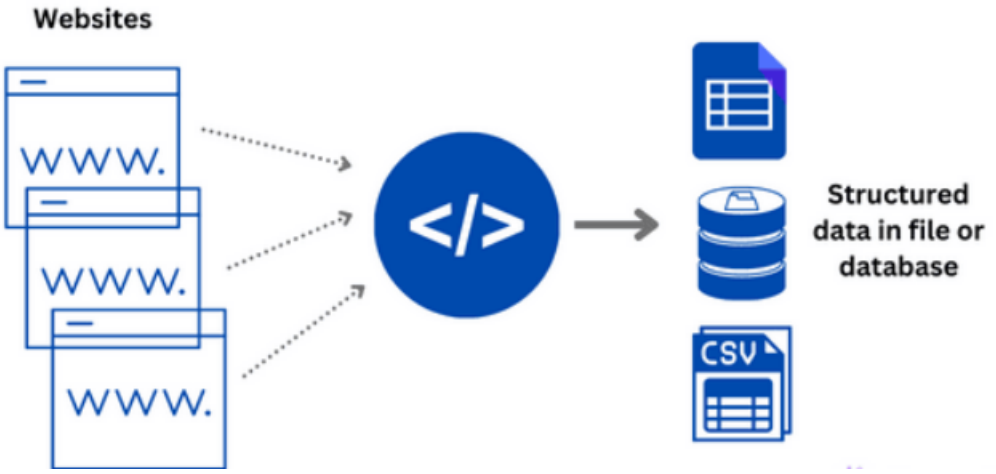
Dotazník vyplnilo celkem 40 osob, přičemž po očištění výsledků průzkumu se jako nejsledovanější témata objevila politika, finance a lifestyle. Výsledky celého průzkumu naleznete v Tabulce 1.

1.2 Sběr dat z X

1.2.1 Teorie

Sběr dat probíhal z webové aplikace platformy X pomocí tzv. data resp. web scraping. Tato metoda umožňuje získat data i z programu, který sám o sobě neumožňuje export dat mimo program.

Web scraping využívá upraveného webového prohlížeče, který se dá spustit pomocí scriptu a automatizovaně se ovládá pomocí předem nastavených parametrů. Ilustrace fungování web scrapingu viz Obrázek 1.



Obrázek 1: Ilustrace web scrapingu

2 Ruční statistika

3 Statistika dle výskytu nejpoužívanějších slov

4 Vektorizace slov a použití lineární regrese

5 Vlastní neuronová síť

6 Analýza pomocí LLM

7 Závěr a zhodnocení