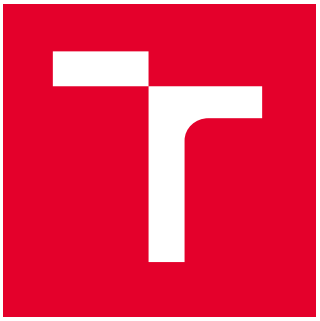


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ

DEPARTMENT OF TELECOMMUNICATIONS

NÁSTROJ PRO SHROMAŽĎOVÁNÍ DAT A INFORMACÍ Z PROSTŘEDÍ INTERNETU

DATA COLLECTING TOOL FROM THE OPEN-SOURCE DATASETS

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. Petr Kříž

AUTHOR

VEDOUCÍ PRÁCE — Ing. Tomáš Gerlich

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Informační bezpečnost**

Ústav telekomunikací

Student: Bc. Petr Kříž ***ID:*** 220894

Ročník: 2 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Nástroj pro shromažďování dat a informací z prostředí Internetu

POKYNY PRO VYPRACOVÁNÍ:

Hlavním cílem diplomové práce je návrh a implementace nástroje pro sběr informací z otevřených zdrojů (sociální síť Instagram). Vlastní aplikace bude schopna shromažďovat data z prostředí Internetu. Na základě získaných výsledků bude možné provést zhodnocení získaných dat k definované vlastnosti (přednost, slabina, příležitost a hrozby). Aplikace bude reflektovat potřeby AČR. V teoretické části nastudujte současný stav problematiky (dostupné metody a nástroje, reflektujte potřeby AČR v ČR i zahraničí), navrhněte postup sběru dat z otevřených zdrojů. Na základě analýzy navrhněte a implementujte nástroj v jazyce Python a jeho funkčnost ověřte.

DOPORUČENÁ LITERATURA:

[1] PASTOR-GALINDO, Javier, et al. The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. IEEE Access, 2020, 8: 10282-10304.

[2] TABATABAEI, Fahimeh; WELLS, Douglas. OSINT in the Context of Cyber-Security. Open Source Intelligence Investigation: From Strategy to Implementation, 2016, 213-231.

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 21.5.2024

Vedoucí práce: Ing. Tomáš Gerlich

doc. Ing. Jan Hajný, Ph.D.
předseda rady studijního programu

ABSTRAKT

Diplomová práce se detailně věnuje oblasti zpravodajských disciplín, přičemž klade důraz na zpravodajství z otevřených zdrojů (OSINT). V rámci práce je podrobně popsán zpravodajský cyklus, včetně rozboru jednotlivých fází tohoto cyklu. Dále teoretická část práce přináší komplexní přehled o sociální síti Instagram z pohledu OSINT. Zahrnuje to popis druhů dat dostupných na této platformě a metod, které jsou používány pro jejich sběr. V praktické části je pak předveden návrh a realizace nástroje pro monitorování otevřených dat na Instagramu. Tento nástroj je navržen tak, aby monitoroval veřejné příspěvky a komentáře na Instagram. Hlavním výstupem této práce je funkční nástroj, který nejen shromažďuje data, ale také je schopen tato data efektivně vizualizovat.

KLÍČOVÁ SLOVA

Zpravodajský cyklus, OSINT, sociální sítě, Instagram, Selenium

ABSTRACT

The thesis examines the field of intelligence disciplines in detail, with an emphasis on open source intelligence (OSINT). The thesis includes a detailed description of the intelligence cycle, including an analysis of the different phases of the cycle. Furthermore, the theoretical part of the thesis provides a overview of the social network Instagram from an OSINT perspective. This includes a description of the types of data available on this platform and the methods that are used to collect it. The practical part then demonstrates the design and implementation of an open data monitoring tool on Instagram. This tool is designed to monitor public posts and comments on Instagram. The main output of this thesis is a functional tool that not only collects data but is also able to visualize this data effectively.

KEYWORDS

The Intelligence cycle, OSINT, social networks, Instagram, Selenium

Vysázeno pomocí balíčku `thesis` verze 4.09; `https://latex.fekt.vut.cz/`

KŘÍŽ, Petr. *Nástroj pro shromažďování dat a informací z prostředí Internetu*. Diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2024. Vedoucí práce: Ing. Gerlich Tomáš,

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Bc. Petr Kříž
VUT ID autora:	220894
Typ práce:	Diplomová práce
Akademický rok:	2023/24
Téma závěrečné práce:	Nástroj pro shromažďování dat a informací z prostředí Internetu

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Tomášovi Gerlichovi, za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Seznam obrázků

Úvod

V digitálním věku jsou sociální sítě neodmyslitelnou součástí každodenního života milionů uživatelů po celém světě. Jednou z dnes nejpopulárnějších sociální sítí je Instagram, který poskytuje uživatelům možnost sdílet fotografie, videa a příběhy. Instagram, díky svému vizuálnímu obsahu a široké uživatelské základně, představuje nejen bohatý zdroj zábavy, ale i informací. Takové informace jsou cenný materiál pro zpravodajství z otevřených zdrojů (OSINT). OSINT využívá veřejně dostupné informace pro získávání důležitých poznatků. V kontextu sociálních sítí lze takové informace použít k různým účelům ať už monitoring a prevence kriminality, cílení reklam, analýza trhů, průzkumy nebo investigativní žurnalistika.

Diplomová práce je zaměřena na tématiku sběru a analýzu dat z otevřených zdrojů, včetně vlastní implementace automatizovaného nástroje umožňující sběr.

Teorie v práci popisuje téma zpravodajství, včetně popisu jednotlivých fází zpravodajského cyklu. Následně jsou rozebrány dostupné nástroje a nejčastější techniky pro sběr dat, ať už automatizované nebo manuální. Součástí je i analýza Instagramu z hlediska jeho jeho fungování, soukromí uživatelů a veřejně dostupných dat, která poskytuje.

Ze získaných poznatků z teoretické části diplomové práce vychází vlastní návrh a implementace nástroje pro sběr, zpracování a analýzu dat z otevřených zdrojů u sociální sítě Instagram. Výsledná aplikace na základě získaných výsledků provádí komplexní zhodnocení dle požadovaných parametrů (přednost, slabina, příležitost a hrozby).

Je však důležité si uvědomit, že informace získané z otevřených zdrojů nemusí být vždy spolehlivé. Lidé na sociálních sítích často šíří nepravdivé nebo zavádějících informací, často se tak děje díky pocitu anonymity. Proto je klíčové tyto informace důkladně ověřovat a kriticky hodnotit jejich pravdivost a relevanci.

1 Zpravodajství

Zpravodajství je klíčový prvek jehož cílem je chránit národní zájmy. Náplní zpravodajské činnosti je zpravodajský cyklus. [1]

Zpravodajství v České republice upravuje zákon zákon č. 153/1994 Sb., o zpravodajských službách České republiky [2], ten definuje následující zpravodajské služby:

- **Bezpečnostní informační služba** – shromažďuje informace o hrozbách proti demokracii, svrchovanosti a územní celistvosti ČR. Činnost zahrnuje sledování cizích zpravodajských služeb, státního tajemství a organizovaného zločinu.
- **Úřad pro zahraniční styky a informace** – úřad shromažďuje informace zahraničního původu, které jsou důležité pro bezpečnost a ochranu zahraničních politických a ekonomických zájmů ČR.
- **Vojenské zpravodajství** – zajišťuje informace důležité pro obranu a bezpečnost ČR, včetně informací o cizích zpravodajských službách v oblasti obrany a hrozbách proti utajovaným skutečnostem v obranném sektoru. [3]

1.1 Disciplíny zpravodajství

Zpravodajství rozlišuje několik základních disciplín. Neexistuje žádný standard, který by definoval přesný počet disciplín, ale komunita se shoduje na existenci pěti základních disciplín zpravodajství. [4] V současnosti se stále více zdůrazňuje význam otevřeného zdrojového zpravodajství. Disciplíny:

- **HUMINT** – nejstarší forma sběru dat, zabývá se shromažďováním informací přímo od lidí. Může probíhat otevřeně, například při výsleších, nebo tajně prostřednictvím špionáže.
- **SIGINT** – disciplína zachytává a využívá signál. Provádí se z různých platforem včetně lodí, letadel a pozemních stanovišť. Platformy nebo zařízení monitorují i vysílání z pozemních zařízení a družic.
- **GEOINT** – kombinuje obrazového zpravodajství a geoprostorových informací, které se získávají analýzou snímků a dat spojených s určitým místem. Využívá se, jak pro vojenské účely, tak i pro sledování geografické změny (eroze, využití zemědělské půdy). Problémem může být kvalita snímků.
- **MASINT** – analyzuje vyzařování zařízení v rámci elektromagnetického spektra. Používá se k odhalování informačních vzorců, které dosud nebyly využity jinými systémy. Patří sem radarové, jaderné a chemické zpravodajství.
- **OSINT** – disciplína čerpající informace z veřejně dostupných zdrojů, které jsou získávány legálními prostředky. Získal na popularitě s nástupem internetu a digitálních médií. Lze jej vytvářet z informací nalezených v široké škále formátů a zdrojů. [4]

2 Zpravodajství z otevřených zdrojů

Zpravodajství z otevřených zdrojů, dále už jen OSINT, je metoda sběru informací z veřejných nebo otevřených zdrojů. Data a informace v OSINTu mohou mít mnoho forem a mohou být získávány z různých zdrojů.

Dříve byla metoda výhradně armádní záležitostí a sloužila pro zajištění národních zájmů. V dnešní době se OSINT stal klíčovou technikou ve státním i soukromém sektoru a je používán pro zajištění bezpečnosti, ale i k útokům. Celý proces je složen z pěti kroků, které se dohromady nazývají zpravodajský cyklus. [5]

2.1 Zpravodajský cyklus

Zpravodajský cyklus je tvořen z pěti po sobě jdoucích kroků:

- **Příprava** – definice potřeb a požadavků.
- **Sběr dat** – sběr dat z relevantních zdrojů.
- **Zpracování** – zpracování a organizování shromážděných dat.
- **Analýza** – poskytování interpretace shromážděným datům.
- **Report** – prezentace výsledků, report, časová osa a doporučení. [5]

2.1.1 Fáze plánování a požadavků

Plánování strategie a definice požadavků je základem každé úspěšné investigace. Na fázi je kladený velký důraz. V opačném případě může dojít k zbytečnému shromažďování informaci, neproduktivní analýze nebo reportování informací, které jsou nepoužitelné a neakceschopné. Požadavky jsou stanoveny zúčastněnými stranami nebo zákazníky a reflektují, co bude obsahem závěrečného reportu. Výstupem fáze je vytvořený plán s rozvrhnutím jednotlivých fází společně. Součástí plány by měly být i odpovědi na následující otázky:

- Kolik potřebujeme analytiků a jaká je potřeba jejich specializace?
- Kolik času bude věnováno jednotlivým fázím?
- Jaké zdroje dat budeme používat?
- Jaké nástroje budou použity?
- Kam a jak budeme data ukládat a kdo k nim bude mít přístup?
- Jakých potencionálních právních problému si je třeba být vědom?
- Co bude obsahem závěrečného reportu a jaký bude mít formát? [6]

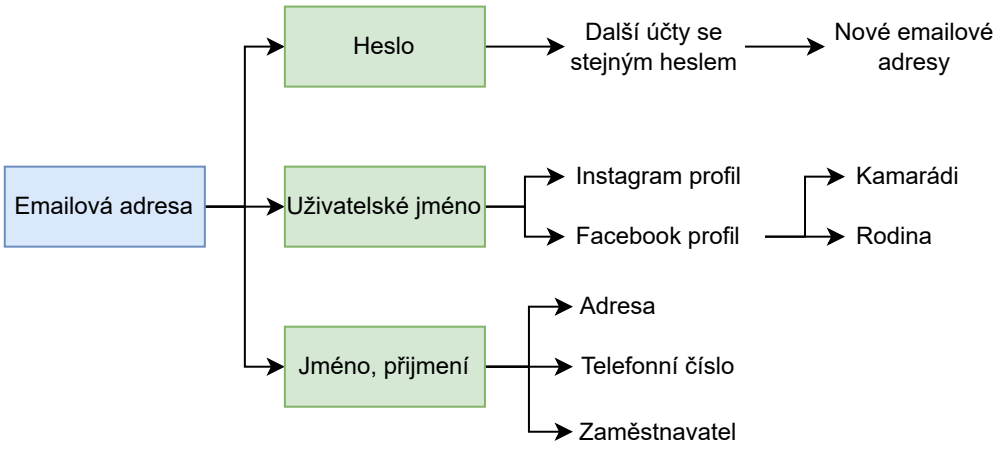
2.1.2 Shromažďování informací

Cílem fáze je shromáždit surové data a informace. Informace se nejčastěji shromažďuji dvěma způsoby, manuálně nebo automaticky. Mezi hlavní zdroje z kterých jsou

data shromažďována patří:

- veřejné záznamy (soudní dokumenty, majetkové, obchodní záznamy),
- zpravodajské zdroje (noviny, časopisy, on-line zpravodajství),
- vyhledávače (Google, Bing, Yandex),
- sociální sítě (Facebook, Instagram, Tiktok, LinkedIn),
- knihovny,
- multimédia,
- webové stránky,
- dark web,
- úniky dat (haveibeenpwned.com). [6]

Součástí sběru dat je i proces zvaný pivot-ování. Jedná se o proces, kdy se setkáme s údaji, které nás přímo vedou k dalším údajům, jež mohou korelovat další zjištění. Pivot-ování může začít úplně obyčejným údajem jako je emailová adresa viz. obr. 2.1, pomocí které je možné nalézt údaje jako jsou jméno, adresa, telefonní číslo. Každý další údaj může vést ke zjištění dalšího údaje. [6]



Obr. 2.1: Pivot-ování

2.1.3 Zpracování

Cílem fáze je zpracovat surová data, vyfiltrovat a přeměnit je v data, která se použijí v dalších fázích cyklu. Příkladem zpracování je dekódování dat, očištění nebo překlady. Součástí zpracování dat je i ověřování pravosti dat. Jedna z hodnotících metod pro ověřování pravosti, kterou využívá Severoatlantickou aliancí je Admiralty Code. [6]

2.1.4 Analýza

Proces analýzy zahrnuje identifikaci vzorců, trendů a vztahů ve zpracovaných datech, s cílem získat jasné a relevantní informace, které pomáhají uživateli pochopit určitou situaci nebo problém. Během analýzy mohou analytici používat různé techniky a nástroje, včetně datové analýzy, vizualizace dat a sémantické analýzy, aby lépe pochopili obsah a kontext informací. Součástí fáze je odpověď na otázky položené v plánovací fázi, hodnocení platnosti, relevance a spolehlivosti shromážděných informací, zvýrazňování inteligentních mezer a nezodpovězených otázek. Na nezodpovězené otázky je nutné se zaměřit v další iteraci zpravodajského cyklu. Během analýzy je důležité rozlišovat mezi podporujícími a vyvracejícími teoriemi, aby se zabránilo potvrzovacímu zkreslení. Analýza by měla také zahrnovat vizualizaci informací pro lepší prezentaci vzorců a trendů. Pro vizualizaci je vhodné použít myšlenkové mapy, tabulky a grafy. [7]

2.1.5 Report

Účelem reportu je informovat čtenáře o zjištěních a poskytnout doporučení pro rozhodování a kontextové znalosti. Zpráva by měla být přizpůsobena spotřebitelům a mělo by se předpokládat, že zpráva bude šířena mimo ně, tzn. že veškeré podrobnosti musí být čtenáři schopni pochopit bez prezentace analytika. Je také vhodné zvážit adresáta reportu, generální ředitel by měl dostat jinou zprávu než soukromý občan. Report by měl obsahovat následující položky:

- **Titulní strana** – musí obsahovat nadpis, zadání, datum vyhotovení a zkrácené shrnutí.
- **Analýza** – musí obsahovat odpověď na všechny definované otázky v prvotní fázi, včetně všech postupů a zjištění během investigace.
- **Shrnutí** – zopakování shrnutí analýzy a zjištění.
- **Doporučení** – čtenář musí obdržet doporučení. Například odborné návrhy nebo postupy vhodné k nápravě.
- **Přílohy** – obrázky a tabulky by měly být součástí reportu. [5]

2.2 Techniky a služby sběru OSINT

Díky internetu existuje stále větší počet technik a služeb, které usnadňuji shromažďované veřejně dostupných dat na internetu. [8] Mezi časté techniky a služby manuálního sběru OSINT patří:

- **Vyhledávače** – vyhledávače Google, Bing nebo Yahoo jsou široce používanými jako základní prostředek pro OSINT. Výsledky těchto vyhledávačů poskytují

relevantní informace, ale výsledků může být někdy nadbytek. Z tohoto důvodu je dobré umět specifikovat požadavky. Specifikace požadavků je u Googlu možná díky Google Dorking, které je popsáno v následující sekci. Je vhodné se také zaměřit na vyhledávače jako Yandex nebo Baidu, které jsou populární ve východní kultuře.

- **Sociální sítě** – sociální sítě obsahují velké množství osobních informací. Facebook, YouTube, Twitter, Instagram a LinkedIn jsou příklady globálně používaných sociálních sítí, které poskytují údaje jako zaměstnání, vzdělání, místa návštěv, záliby a vztahy, které mohou pomoci identifikovat cílové osoby nebo organizace.
- **Techniky sběru** – emailových adres, uživatelského jména, skutečného jména, lokalit, IP adres a doménových jmen. [8]

Pro každou zmíněnou techniku existuje několik služeb, které sběr uživateli zjednoduší. Aktualizovaný seznam služeb poskytuje OSINT framework.

2.2.1 OSINT framework

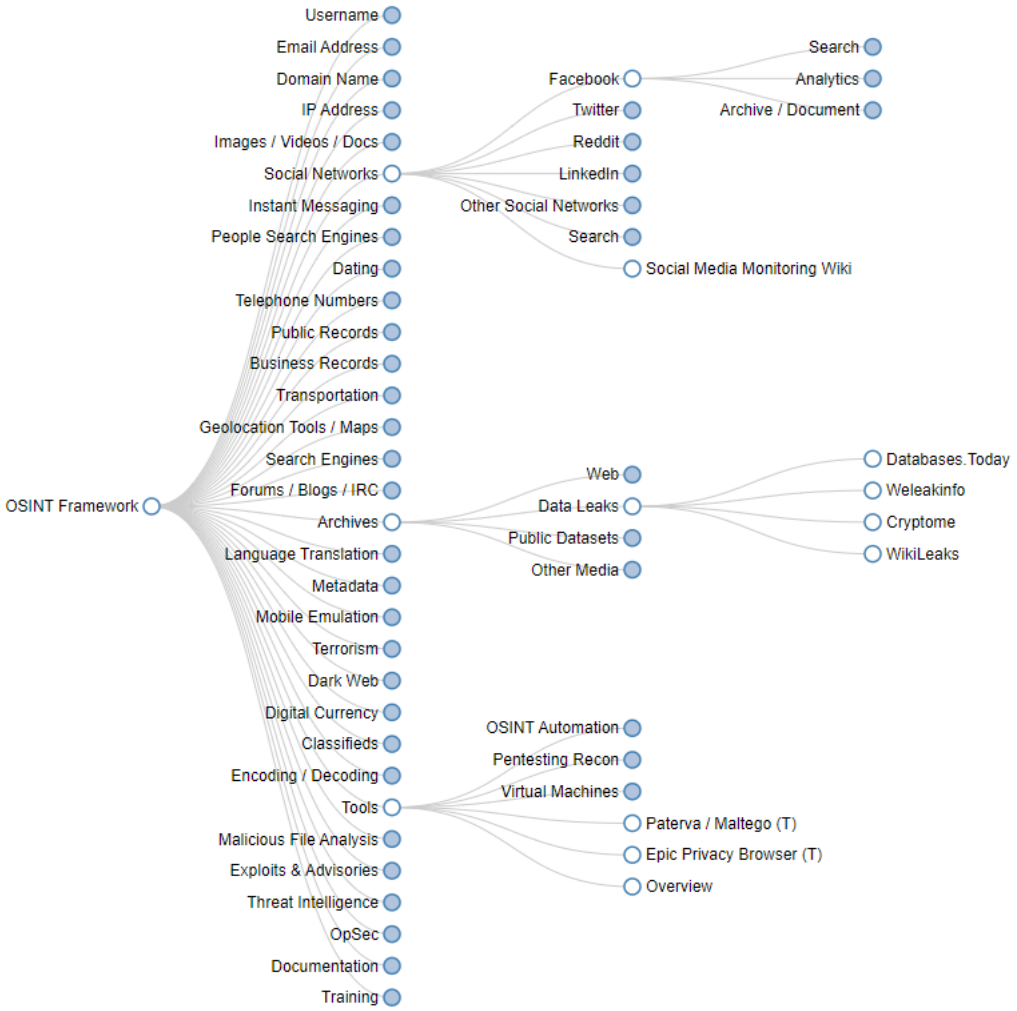
OSINT framework usnadňuje hledání bezplatných služeb a zdrojů pomocí kterých lze hledat informace o cíli. Webová platforma OSINT framework má stromovou podobu s 32 kategoriemi. Každá kategorie je rozdělena do podkategorii podle nástrojů, které lze pro danou kategorii použít. Možnou nevýhodou může být zaměření frameworku pro americký trh.

Všechny nástroje či zdroje se liší svoji přístupností či vlastnostmi. OSINT framework z tohoto důvodu obsahuje ukazatele, pomocí kterých lze jednotlivé nástroje rozlišit. Ukazatele lze vidět na obr. 2.2 a znamenají následující:

- (T) – nástroj musí být nainstalován a spuštěn lokálně.
- (D) – další informace získáte na Google Dorking: Google Hacking.
- (R) – nástroj vyžaduje registraci.
- (M) – označuje adresu URL, která obsahuje hledaný výraz. Samotná adresa URL musí být upravena ručně. [9]

2.2.2 Google Dorking

Google Dork je technika vyhledávání za použití specifických vyhledávacích příkazů (včetně speciálních parametrů a vyhledávacích operátorů) k pokročilým vyhledávacím dotazům. Použití těchto příkazů slouží k odhalení skrytých informací ve službě Google, které jsou jinak běžným vyhledáváním nedostupné. Google dorking může odhalit citlivé nebo soukromé informace o webových stránkách a společnostech, organizacích a jednotlivcích, kteří je vlastní a provozují. [10] Seznam vyhledávacích operátorů a příklady použití jsou obsaženy v tabulce 2.1.



Obr. 2.2: OSINT framework

2.3 OSINT nástroje

Použití manuálních služeb a technik (vyhledávače, procházení sociálních sítí) může být pro základní vyhledávání dostačující, ale pro koncového uživatele značně nepraktické. Hlavní nevýhodou manuálních technik je velký počet služeb, který je nutné kombinovat, což může být z praktického hlediska o zdlouhavý proces. Z těchto důvodů existují automatické nástroje, které provádí jednotlivé kroky automaticky. V případě OSINTu se nejčastěji skloňují následující nástroje:

- **FOCA** – nástroj analyzuje metadata dokumentů a souborů zveřejněných online pro odhalení interních informací organizace.
- **Maltego** – interaktivní nástroj pro vizualizaci grafů, který umožňuje sledovat vztahy mezi lidmi, skupinami a organizacemi na internetu.
- **Metagoofil** – nástroj pro získávání informací, který vyhledává metadata z veřejně dostupných dokumentů (např. PDF, DOC, XLS).

Tab. 2.1: Google Dorking

Dork	Popisek	Příklad
allintext	vyhledává výskyt všech zadaných slov	allintext:"slovo"
intext	vyhledává výraz v obsahu stránky	intext:"slovo"
inurl	vyhledává výraz v adrese URL stránky	inurl:"slovo"
intitle	vyhledává výraz v názvu stránky	intitle:"slovo"
site	hledá na specifické stránce	site:"www.instagram.com"
filetype	hledá specifický formát dokumentu	filetype:"pdf"
link	vyhledává stránky s odkazem	link:"slovo"
related	vyhledává podobné stránky	related:"www.instagram.com"
cache	vyhledávání verze stránky v mezi paměti	cache:"www.instagram.com"

- **Recon-NG** – nástroj sleduje internetové zdroje a databáze k identifikaci informací o osobách, doménách a síťových zařízeních.
- **Shodan** – nástroj pro vyhledávání veřejné informace o připojených uzlech k internetu, včetně zařízení IoT, serverů a kamer, a je využíván hlavně pro zabezpečení sítí.
- **The Harvester** – nástroj pro získávání e-mailových adres, jmen a subdomén.
- **IntelTechniques** – nástroj pro pokročilé vyhledávání informací, zaměřující se na osobní údaje a související sítě. [8]

V další sekci jsou detailněji popsány nástroje, které se používají pro OSINT sociálních sítích.

2.3.1 Maltego

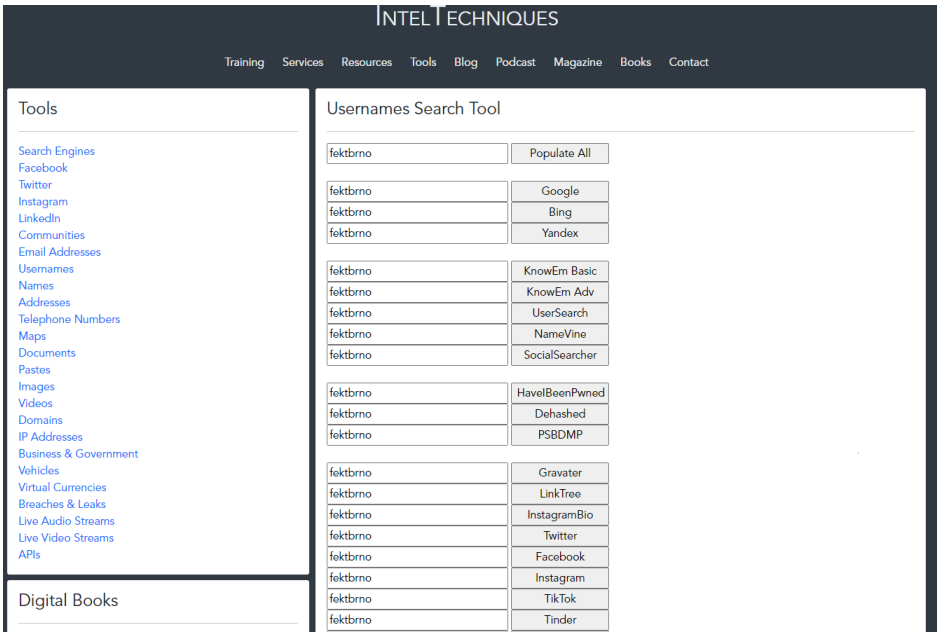
Maltego je univerzální nástroj pro vytěžování dat a shromažďování informací v reálném čase. Poskytuje strukturované a snadno čitelné grafy, díky kterým lze snadno identifikovat vzory a mezi uvedenými informace. [11] Jedná se o placený nástroj, existuje však i komunitní verze je zdarma a určena pro nekomerční účely, je však limitována omezeným počet výsledků. Maltego je rozšířeno o nástroje třetích stran. Nástroje třetích stran, které lze použít pro investigaci na sociálních sítí. [12]

- **ShadowDragon SocialNe**t – umožňuje mapovat spojení z více než 150 sociálních sítí. Nástroj umožňuje identifikovat on-line identity a sledovat vazby mezi těmito identitami. Z nalezených sociálních sítí lze automaticky shromažďovat data z profilů, příspěvků a komentářů na sociálních sítích do jednoho grafu, kde lze provádět vyhledávání v textech a zjišťovat vzájemné provázání. V praxi to umožňuje rychle zúžit počáteční vyšetřování na několik jednotlivců používajících různé aliasy spojené s podezřelými osobami. [13]

- **Social Links Professional** – Nástroj poskytuje přístup k velkému objemům dat a rozsáhlým přehledům z více než 500 otevřených zdrojů v sociálních médiích, blockchainu a temném webu, čímž vizualizuje komplexní obraz pro rozhodování a vyšetřování založená na datech. [14]

2.3.2 IntelTechniques

IntelTechniques je webová stránka, kterou vytvořil MichaelBazzel, která nabízí stovky on-line vyhledávacích nástrojů. [15] Existuje několik modulů rozdělených podle cílových údajů, které umožňují vyhledávání mimo jiné podle e-mailu, profilu sociální sítě, skutečného jména a uživatelského jména (viz. obr. 2.3), aby koncovému uživateli předložily shromážděné veřejné informace. [16]



Obr. 2.3: IntelTechniques

2.3.3 Sherlock

Je obecně známo, že mnoho uživatelů používá na sociálních sítích stejná nebo obdobná uživatelská jména. Sherlock je nástroj, který usnadňuje vyhledávání uživatelských jmen. [17]. Jedná se o nástroj s otevřeným zdrojovým kódem a příkazovým řádkem vyvinutý v jazyce Python, který vyhledává konkrétní uživatelské jméno na nejpopulárnějších sociálních sítích. Výstupem jsou odkazy na profily uživatelů. Sherlock se ovládá příkazového terminálu systému Windows, Linux nebo MacOS. Největší výhodou nástroje je, že pomocí jednoho příkazu Sherlock projde přes 400 sociálních sítí. [18]

2.3.4 Web scraping

Web scraping je technika používaná k získávání velkého množství dat z webových stránek a jejich následnému ukládání do místního počítače ve formě XML, Excelu, CSV atd. [19] Na základě zadaných požadavků dokáží ve zlomku času získat data z libovolné webové stránky. Web scraping nástroje fungují ve čtyřech krocích:

- Odeslání požadavku na cílovou stránku,
- získání odpovědi z cílové stránky,
- extrakce odpovědi,
- stažení dat. [20]

Web scraping v kontextu OSINTu pro sociální sítě znamená využívání automatizovaných nástrojů pro extrakci veřejně dostupných informací z různých sociálních médií, jako jsou Facebook, Twitter, Instagram a další. Výzvou může být rychlá změna API a pravidel sociálních sítí.

2.3.5 Web crawler

Web crawler navštěvuje webové stránky a čte je za účelem vytvoření záznamů pro index vyhledávače. Analyzují se různé webové stránky a sledují se odkazy na nich, což umožňuje provádět hloubkové vyhledávání pro získání dalších informací. Známé vyhledávače, jako jsou Google, Yahoo a Bing, provádějí procházení webu a tyto informace používají k indexování webových stránek. [21] Cílem crawleru je tedy porozumět obsahu webových stránek. Uživatelé tak mohou podle potřeby získávat informace o jedné nebo více stránkách. [22]

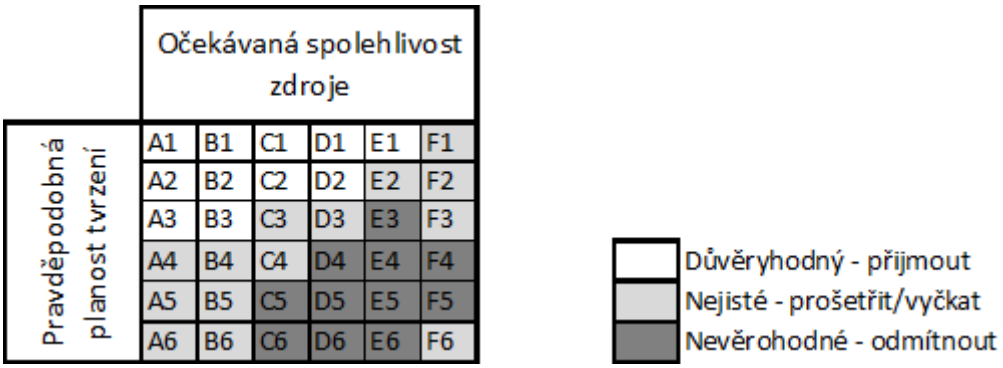
2.4 OSINT limitace

Veřejné informace dostupné na internetu jsou ze své podstaty masivně neuspořádané. To znamená, že shromážděné údaje jsou natolik různorodá, že je obtížné je klasifikovat, propojit a prozkoumat s cílem získat relevantní vztahy a znalosti. Sociální sítě a komunikační média jsou zaplaveny subjektivními názory, falešnými zprávami a falešnými informacemi. [23] OSINT činnosti by se měly vždy zabývat spolehlivými a důvěryhodnými informacemi, aby se zajistilo pozitivních a přesvědčivých výsledky [24]. Jednou z používaných technik Severoatlantické aliance pro ověřovaní pravosti informací a zdrojů je Admiralty Code.

OSINT představuje i etický problém. Odhalené výsledky by měly respektovat soukromí uživatelů a neodhalovat intimní, osobní nebo politické záležitosti [25], přičemž je třeba vzít v úvahu s přihlédnutím k platným souvisejícím předpisům, například GDPR.

2.4.1 Admiralty Code

Součástí technik pro OSINT jsou i hodnotící techniky pro ověřování pravosti informací a zdrojů. Admiralty Code je jedna z používaných technik Severoatlantické aliance a obsahuje techniky pro hodnocení důvěryhodnosti důkazů. Hodnotí jejich spolehlivost (A-F) a platnost (1-6). Spolehlivost zdroje, jako je osoba nebo publikace, závisí na jeho pověsti a faktorech, jako jsou motivace a kompetence. Platnost tvrzení se posuzuje na základě jeho konzistence s dalšími důkazy a existujícími teoriemi. Důkazy jsou klasifikovány podle diagonální škály, přičemž A1 je vysoce důvěryhodné (renomovaný zdroj, ověřené tvrzení) a E5 je nedůvěryhodné (pochybný zdroj, nekonzistentní tvrzení). Tvrzení mimo diagonálu, jako E2 nebo B5, vyžadují pečlivé zvážení motivu zdroje nebo další ověření. Vyšetřovatel pomocí kódu rozhoduje, které důkazy jsou důvěryhodné, nejisté nebo nedůvěryhodné, přičemž světle šedá barva označuje nejistotu a potřebu dalšího šetření. Tento systém pomáhá dosahovat informovaných soudů, zatímco se snaží vyhnout chybám, i když příliš opatrné přístupy mohou bránit rozhodování. [26]



Obr. 2.4: Admiralty Code

3 Analýza sociálních sítí

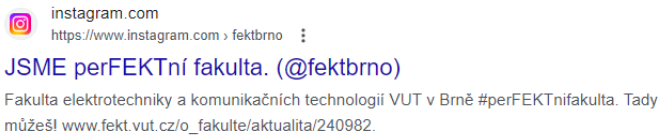
Sociální sítě dnes umožňují volný a snadný přístup k osobním informacím. Analýza sociálních médií vychází z disciplíny OSINT a je označována SOCMINT. Obecně se zabývá uživatelskými profily jednotlivců nebo skupin za účelem získání informací o tom, co dělají, s kým a jak komunikují. Informace ze sociálních sítí lze získat pasivně, kdy nedochází ke komunikaci s jednotlivci a riziko prozrazení je nízké nebo aktivně, kdy je vytvořeno spojení s cílem např. přidávání do přátel či sledování, zde může být riziko prozrazení vysoké a je nutné zvážit následky. [27]

3.1 Instagram

Instagram je sociální sít, která umožňuje uživatelům zveřejňovat obrázky a krátká videa, a to buď veřejně, nebo soukromě pro sledující. Mezi hlavní funkce patří zveřejňování příspěvků ve stálém kanálu uživatele nebo v příbězích (obsah, který trvá 24 hodin), živé vysílání, reels, zasílání soukromých zpráv, sledování jiných účtů a procházení podle tématu, hashtagu nebo lokací. Uživatelé mohou v aplikaci také upravovat fotografie nebo videa pomocí textu, gifů, ikon a filtrů. [28]

3.2 Vyhledávání na Instagramu

Pro vyhledávání na Instagramu existují dvě možnosti. Lze vyhledávat pomocí vyhledávacího panelu Instagramu nebo pomocí Google Dorks. Vyhledávací panel Instagramu je užitečný pro vyhledávání profilů hashtagů nebo lokací, ale není tolik efektivní pro vyhledávání slov ve větách či biografii profilu. Při použití Google Dorking z obr. 3.1 je výsledkem vyhledávání profil uživatele @fektbrno, mezi další výsledky hledání patří příspěvky uživatele.



Obr. 3.1: site:instagram.com fektbrno

Pokud je však použito vyhledávání obr. 3.2 výsledkem vyhledávání jsou příspěvky, kde byl cíl vyhledávání zmíněn (popisky kde cíl označil jiný uživatel). [5] Další možností je pokročilejší dotaz, který z výsledků vyřadí stránku uživatele @fektbrno:

```
site:instagram.com "@fektbrno" inurl:instagram.com/fektbrno/
```

Obr. 3.2: site:instagram.com @fektbrno

3.3 Soukromí na Instagramu

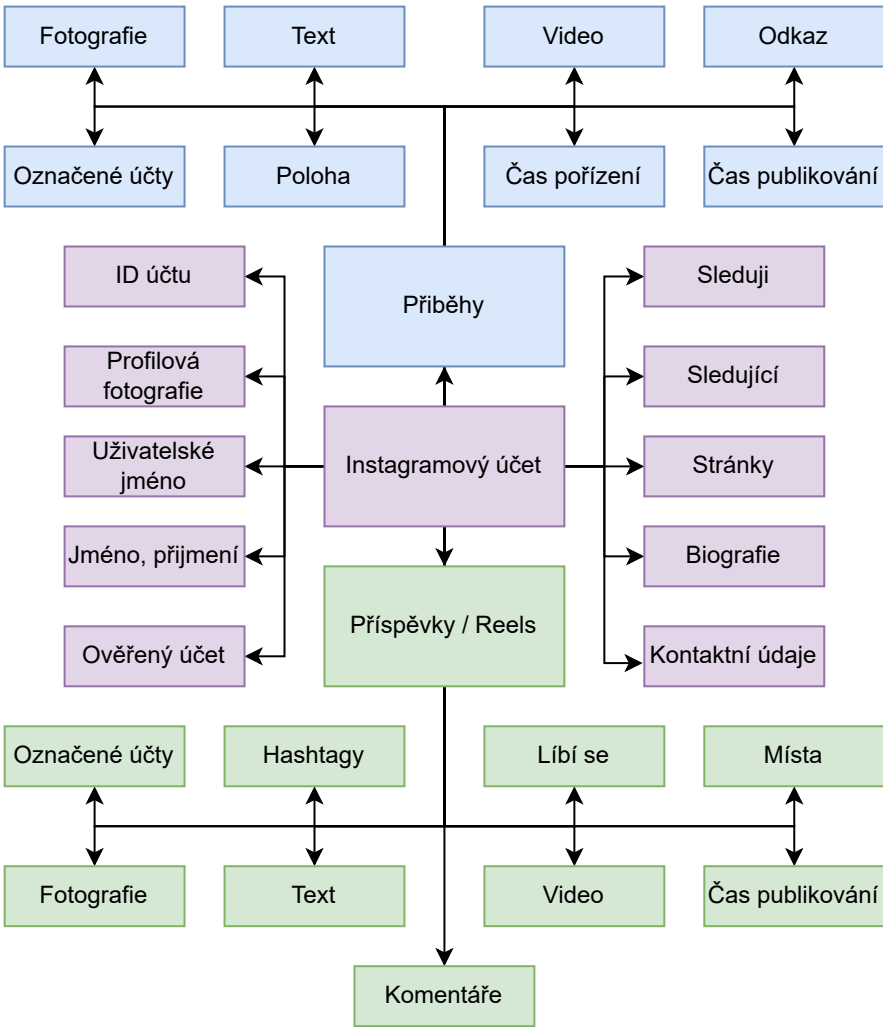
Instagram dělí profily na soukromé a veřejné. K veřejným profilům lze přistupovat bez nutnosti získat povolení o sledování od majitele účtu. Z veřejného profilu lze následně získávat všechna data, která popisuje obr. 3.3. [29]

Pokud je uživatelský profil označen jako soukromý lze bez zaslání žádosti o sledování získat informace pouze o datu založení uživatelského účtu, profilovou fotografii, uživatelské jméno, ID účtu a biografii.

S rozvojem technologií je navíc možné z veřejných profilů pomocí nástrojů a skriptů získávat veřejně dostupná data. Například pomocí jazyka Python lze automatizovat sběr veřejně dostupných dat. Data a informace lze následně analyzovat pro různé účely, jako je pochopení chování spotřebitelů nebo sledování popularity určitých témat či hashtagů.

Při shromažďování dat z veřejných Instagramových profilů je důležité poznamenat, že vlastníci těchto profilů nejsou informováni o takovém přístupu. To znamená, že pokud někdo prochází příspěvky, komentáře nebo veřejně dostupné informace na veřejném profilu, majitel profilu nedostane žádné oznámení nebo indikaci, že k jeho profilu byl přistoupeno. Na druhé straně, pokud někdo přistupuje k Instagram příběhům na profilu, vlastník profilu má možnost vidět seznam uživatelů, kteří tyto příběhy prohlíželi. Tato funkce poskytuje majitelům profilů přehled o tom, kdo se zajímá o jejich krátkodobě zveřejněný obsah.

Vlastníci profilů mohou však získat určité informace o interakcích na svém profilu, jako jsou statistiky o počtu lajků, komentářích a sdílení, ale tyto informace nespecifikují, kdo konkrétně jejich profil navštívil. Pro uživatele, kteří mají obchodní nebo tvůrčí účet, Instagram poskytuje detailnější analytické údaje, ale ty se zaměřují spíše na celkové metriky, jako je dosah a zapojení uživatelů, nikoliv na konkrétní identifikaci návštěvníků profilu.



Obr. 3.3: Volně dostupná data z Instagramu

3.4 Algoritmus

Instagram obsahuje velké množství obsahu a pro člověka je nemožné ho vidět všechen. Z tohoto důvodu sociální síť obsahuje několik algoritmů, klasifikátorů a procesů, přičemž jejich hlavním účelem je personalizovat obsah. To znamená ukázat

uživatelům obsah, o který se nejvíce zajímají. Z pohledu algoritmu lze Instagram rozdělit na tři sekce úvodní strana a příběhy, objevovat a reels.

Algoritmus Instagramu pro úvodní stránku a příběhy se zaměřuje na zobrazování příspěvků od sledovaných osob. Využívá množství signálů, jako je popularita příspěvku, čas jeho zveřejnění, interakce s autorem příspěvku a uživatelova aktivita. Algoritmus dělá předpovědi o tom, jak pravděpodobné je, že uživatel s příspěvkem interaguje. Interakcí se myslí komentováním, označení líbí se nebo ukládání.

Sekce objevovat pomáhá objevovat nový obsah na základě minulých interakcí uživatele a zájmů podobných uživatelů. Základem jsou signály jako popularita příspěvků a historie interakcí s osobou, která příspěvek zveřejnila.

Reels je zaměřen na zábavu a vybírá obsah od účtů, které uživatel nesleduje. Důležité jsou signály jako uživatelova aktivita a historie interakcí s autorem videa. Algoritmus se zaměřuje na to, jak pravděpodobné je, že uživatel sleduje video do konce, že se mu líbí, nebo že jej shledá zábavným. Pokud chce uživatel ovlivnit zobrazovaný obsah, je nutné sledovat požadovaný obsah až dokonce, přidávat ho do oblíbených, komentovat, ukládat. Pokud však o obsah nemá zájem, je nutné ho označit pomocí tlačítka nemám zájem. [30]

3.5 Vytvoření účtů a jejich správa

Sběry dat ze sociálních sítí mohou být v rozporu s pravidly a podmínkami použití této platformy. Instagram má striktní politiku proti používání automatizovaných nástrojů pro získávání dat, která je zaměřena na ochranu soukromí uživatelů a integritu platformy. Proto je důležité aby takový profil nezasahoval do soukromí uživatelů nebo nebyl používán pro spamování nebo škodlivé aktivity.

Instagram aktivně monitoruje a blokuje účty, které se domnívá, že porušují jeho pravidla. To zahrnuje účty, které jsou identifikovány jako účty, které provádějí podezřelé aktivity, jako je masové sledování, masové odebírání nebo automatizované komentování. [31]

4 Sentimentová analýza

Téměř 80 % internetových dat tvoří text. Jedná se o proces zpracování přirozeného jazyka, jehož cílem je analyzovat názory, nálady, emoce nebo postoje osob v textových datech. Z pohledu sociálních sítí se používá k rozpoznávání veřejného mínění a trendů. Analýza sentimentu může pomoci identifikovat, jak veřejnost reaguje na různé události, zprávy nebo značky. [32]

Z pohledu bezpečnostních složek lze využívat k monitorování a vyhodnocování veřejného mínění v určitých regionech nebo zemích, což může být důležité pro plánování misí nebo operací. Analýza sentimentu může pomoci identifikovat potenciální bezpečnostní hrozby, jako jsou šíření extremistických názorů nebo plánování teroristických činů. Bezpečnostní složky mohou sledovat sentiment svých vlastních zaměstnanců na sociálních sítích, aby lépe porozuměla jejich náladám a potřebám. [33]

4.1 Postup analýzy sentimentu

Postup pro všechny metody přístupu je stejný:

- **Sběr dat** – výsledkem sběru dat by měl být seznam uživatelů a jejich názory.
- **Zvolení přístupu** – volba metody přístupu určení sentimentu (nebo jejich kombinaci) podle toho, která odpovídá nejvíce zájmu uživatele nebo jeho možnostem (např. časovým, finančním atd.).
- **Předzpracování dat** – normalizace a očištění komentářů. Procesem normalizace se myslí sjednocení všech komentářů do jednotného formátu. Proces čištění zahrnuje odstranění emotikon, opakujících se slov, odkazů a speciálních znaků.
- **Klasifikace** – provádí se do negativní, pozitivní nebo neutrální kategorie. Klasifikaci lze provést použitím slovníkového přístupu, strojového nebo hloubkového učení. V případě strojového učení se jako trénovací množina pro vytvoření klasifikátoru používá jedna třetina databáze. Přesnost klasifikátoru do značné míry závisí na trénovací množině.
- **Výpočet sentimentu** – učení sentimentu, pro každý přístup se stanovuje jiným postupem. [34]

4.2 Metody přístupů

Existuje několik metod přístupů, kterými lze analýzu provést. Mezi používané metody patří slovníkový přístup, strojové učení nebo hluboké učení. Podle různých

klasifikačních metod se výsledky analýzy sentimentu dělí na pozitivní, negativní a neutrální. [35]

Metody slovníkového přístupu a strojového učení se snadněji implementují. Metody hlubokého učení i kombinace metod hlubokého a strojového učení jsou však nadřazené a mají komplexní povahu. Ve většině případů poskytují lepší výsledky. V některých případech však rozdíl v přesnosti obou technik není příliš vysoký, v takových případech metoda hlubokého učení pouze zvyšuje složitost řešení.

4.2.1 Slovníkový přístup

Jedna z často metod používaných při analýze sentimentu je použití slovníku. Tento přístup efektivní, ale často postrádá zohlednění studovaného kontextu. Přesnost výsledku je omezena na velikost, rozsah a jazyk slovníku. S rostoucí velikostí slovníku se přístup stává nesprávnějším a časově náročnějším.

Důležitým krokem je výběr slovníku. Existuje několik druhů a uživatel by měl zvolit ten, který nejvíce odpovídá jeho zájmu. Cílem je určit sentiment, ten lze vypočítat pomocí porovnání výskytů pozitivních a negativních slov. Přesnost slovníku se vypočítá porovnáním výsledků detekce sentimentů ze systému se skutečnou sentimentální hodnotou předem definovanou lidmi (autory). [36]

4.2.2 Strojové učení

Algoritmy strojového učení se dělí na strojové učení s kontrolou a strojové učení bez kontroly. V oblasti analýzy sentimentu hraje klíčovou roli klasifikace. V kroku klasifikace k trénování a vytvoření klasifikátoru se používá předem klasifikovaný vzorek databáze, nazývaný trénovací množina. Tento klasifikátor pak po naučení vzoru označí dříve neoznačená data. Přesnost každého klasifikátoru je však také velmi závislá na datech použitých k jeho trénování. [37]

- **Klasifikátor Naive Bayes** – klasifikátor Naive Bayes je považován za velmi jednoduchý a snadno implementovatelný. Nejedná se o jediný algoritmus, ale o soubor různých klasifikačních algoritmů založených na Bayesově teorému, který se používá k popisu pravděpodobnosti události. Klasifikátor využívá všechny různé příznaky přítomné ve vektoru příznaků a analyzuje je odlišně (nezávisle). Klasifikátor je schopen naučit se vzor zkoumáním předem roztříděné množiny dokumentů. Tento model uvádí, že v přítomnosti dvou událostí `p1` a `p2`, lze vypočítat podmíněnou pravděpodobnost výskytu události `p1`, pokud již došlo k události `p2`. Vstup pro trénování tvoří předzpracovaná data spolu s extrahovanými příznaky. Po dokončení trénování se proces klasifikace provede na testovacích datech a poté v závislosti na výsledcích na nových datech. Při této klasifikační metodě se stanoví polarita sentimentů dat. [37]

- **Metoda podpůrných vektorů** – Metoda podpůrných vektorů (SVM) analyzuje data, definuje rozhodovací hranice a k výpočtu používá jádra, která se provádějí ve vstupním prostoru. Vstupními daty jsou dvě sady vektorů, každá o velikosti `m`. Poté každý údaj reprezentovaný jako vektor je zařazen do určité třídy. Úkolem najít hranici mezi dvěma třídami, která je vzdálená od jakéhokoli dokumentu. Vzdálenost definuje rozpětí klasifikátoru a maximalizace rozpětí snižuje nerozhodnost rozhodnutí. [38]

4.2.3 Hluboké učení

Jedná se o složitější a podrobnější dílčí část strojového učení. Skládá se z různých typů neuronových sítí a neuronových přesvědčení, které fungují na principu neuronů v mozku. Datové soubory, které se pro tyto techniky obvykle pořizují, mají obrovskou velikost. Hluboké učení se používá k vytváření hlubokých neuronových sítí, které řeší složité binární otázky, rozhodují nebo vracejí číselné odpovědi s vysokou přesností. V případě sentimentové analýzy mezi používané typy neuronových sítí patří konvoluční a pravděpodobnostní neuronové sítě. [39]

5 Praktická část

V rámci praktické části byl vytvořen monitorovací nástroj pro platformu Instagram. Vytvoření nástroje v praktické části je rozděleno na definici požadavků, realizaci samotného nástroje a jeho testování.

5.1 Definice požadavků

Hlavním stanoveným požadavkem v rámci diplomové práce bylo vytvoření nástroje pro sběr veřejně dostupných informací ze sociální sítě Instagram. Nástroj by měl být naprogramován v jazyce Python a měl by být schopen veřejně dostupná data shromažďovat z vybraných uživatelských profilů, lokací, hashtagů a příspěvků na Instagramu.

Nástroj by měl sloužit pouze pro monitorování skupin, nikoliv jednotlivců, což by mohlo vést k porušení etických principů.

Mezi data, která, které jsou z hlediska bezpečnosti zajímavé patří popisky příspěvků a komentáře uživatelů. Cílem je, aby nástroj dokázal tyto data automaticky sbírat, zpracovat, analyzovat a vytvořit report = zpravodajský cyklus.

U sběru popisků nástroj bude schopný detekovat použití specifických klíčových slov v příspěvku. Následně by měl nástroj být schopen vytvořit výstup z kterého bude možné vytvořit vhodnou vizualizaci dat. Vhodnou vizualizací je myšlěno vytvoření dashboardu časové osy použití počtu daného klíčového slova v čase.

Ze sesbíraných komentářů by měl být nástroj schopen určit pomocí sentimentitové analýzy sentiment daného příspěvku a vytvořit anonymizovaný výstup, z kterého bude možné vytvořit obdobný dashboard jako v předchozím případě, tzn. časová osa sentimentu v čase.

5.2 Použití sesbíraných dat

Nástroj poskytuje dvě hlavní funkcionality:

- **Vytvoření časové osy použití klíčových slov v popisku** – časová osa slouží k monitorování bezpečnostních hrozeb, predikci a prevenci konfliktů. V případě monitorovaní bezpečnostních hrozeb může časová osa pomoci identifikovat vzory v aktivitě nepřátelských skupin. Sledováním počtu klíčových slov lze také signalizovat narůstající napětí, což může vést i k predikci možných konfliktů. Osa také může odhalit dezinformační kampaně nebo šíření falešných informací, které by mohly ohrozit národní bezpečnost.

- **Sběr komentářů s navazující analýzou sentimentu** – funkci lze použít k monitoringu veřejného mínění, kdy osa sentimentu v čase může identifikovat změny v náladách a postojích obyvatelstva v reakci na konkrétní události či oznámení. Dalším možným použitím je analýza dopadu informačních kampaní. Pomocí analýzy lze zjistit jak veřejnost přijímá rozhodnutí a to využít k predikci a prevenci nepokojů nebo k pochopení účinku komunikační strategie. Současně analýza sentimentu může pomoci odhalovat a analyzovat šíření falešných informací, což umožňuje rychlé reakce na dezinformační kampaně, které by mohly ohrozit národní bezpečnost.

5.3 Selenium vs Instagram Grapg API

Během fáze definice požadavků proběhla úvaha, jakým způsobem ke sběru dat přistupovat. Bylo nutné rozhodnout, zda použit oficiální Instagram Graph API nebo alternativní přístup, jakým je například nástroj pro automatizaci Selenium.

Instagram Graph API je oficiální rozhraní poskytované platformou Instagram pro vývojáře. Rozhraní nabízí řadu funkcí pro přístup k datům na Instagramu. Rozhraní je však primárně zaměřeno na potřeby obchodních účtů, tvůrců obsahu a jejich interakce s uživateli. Instagram Graph API umožňuje:

- Získávání a správu zveřejněných fotografií, videí a příběhů,
- získávání základních údajů o ostatních účtech Instagram Business a Creator (jedná se o metadata počtu sledujících účtu a médií, základní metrik o médiích, mezi které patří počet komentářů a počet líbí se),
- moderování komentářů a odpovědí na ně,
- měření interakce s médii a profilem,
- zjišťování médií označených hashtagy a @zmínek. [31]

5.3.1 Rozhodnutí

Pro specifické požadavky, které byly definovány použití Instagram Graph API představuje určitá omezení. Klíčovým nedostatkem je, že neumožňuje přímý přístup k všem potřebným datům. Například, získávání popisků příspěvků u jiných uživatelů a získávání komentářů u definovaných příspěvků, není přes rozhraní dostupná. Další nevýhodou je, že Instagram Graph API, klade omezení na množství dat, která lze získat, a na frekvenci požadavků, což je nevyhovující.

S ohledem na uvedená omezení Instagram Graph API bylo rozhodnutu o použití knihovny Selenium. Použití Selenia má několik výhod:

- **Přístup k datům** – Selenium umožňuje získávat data, která nejsou přes API přístupná.

- **Flexibilita ve zpracování dat** – můžeme přizpůsobit proces shromažďování dat specifickým potřebám, což zahrnuje vytváření vlastních filtrů a analytických nástrojů.
- **Obcházení omezení API** – využitím Selenia se lze vyhnout kvótám a omezením frekvence požadavků, které jsou běžné u oficiálních API.

5.4 Příprava vývojového prostředí

Příprava vývojového prostředí se skládá z instalace používaných knihoven a konfigurace knihovny Selenium.

5.4.1 Použité knihovny

Všechny použité knihovny, jsou uvedeny v přiloženém souboru `requirements.txt`. Pro instalaci všech potřebných knihoven lze použít správce balíčků programovacího jazyka `pip`. Knihovny lze následně nainstalovat pomocí příkazu:

```
pip install -r .\requirements.txt
```

Použité knihovny:

- **Pandas** – používá se pro manipulaci s daty a analýzu. Poskytuje datové struktury a funkce pro snadné načítání, čištění, prozkoumávání a transformaci dat.
- **Plotly** – používá se pro vytváření interaktivních grafů a vizualizací v Pythonu. Umožňuje tvorbu složitých grafických reprezentací dat, které jsou interaktivní a mohou být vkládány do webových stránek.
- **Tkinter** – knihovna pro vytvoření grafického uživatelského rozhraní (GUI).
- **Selenium** – knihovna pro automatizaci webového prohlížeče.
- **Textblob, AFINN a VADER** – knihovny pro analýzu sentimentu.

Selenium podporuje automatizaci všech hlavních prohlížečů na trhu pomocí nástroje WebDriver. WebDriver je rozhraní API a protokol, který definuje jazykově neutrální rozhraní pro ovládání chování webových prohlížečů. Každý prohlížeč je podporován specifickou implementací WebDriveru, která se nazývá ovladač. Ovladač je komponenta odpovědná za delegování prohlížeče a zajišťuje s ním komunikaci. Jednou z výhod ovladače je jeho rychlost, jelikož přímo komunikuje s prohlížečem a používá jeho vlastní engine pro ovládání. Ovladač je realističtější v interakci s prvky stránky, například neumožňuje zadávat hodnoty do zakázaných textových polí, což odpovídá skutečnému uživatelskému chování. [40].

5.4.2 Konfigurace knihovny Selenium

Pro používání je nutné stáhnout Selenium ovladač. Stažení lze provést z oficiálních stránek projektu Selenium. Pro správnou funkci Selenium WebDriveru s Google Chrome je důležité, aby verze ChromeDriveru odpovídala verzi prohlížeče Chrome, která je v systému nainstalována. Zjištění verze Google Chrome:

- Otevření Google Chrome.
- Tři tečky v pravém horním rohu prohlížeče -> otevření menu.
- Nápověda -> O prohlížeči Google Chrome
- Zde je uvedena aktuální verze Google Chrome.

Selenium ovladač je dostupný na následující adrese.[1]. Po stažení ovladače je potřeba vložit do adresáře s projektem. V případě uložení na jiné místo je potřeba nastavit cestu v kódu nastavit cestu k ovladači. Následně lze ovladač použít k vytváření automatizovaných úloh. Cesta k ovladači se nastavuje ve třídě `scrapeData.py` ve funkci `init_driver()`. Nastavení ChromeDriveru lze vidět ve výpise 5.1

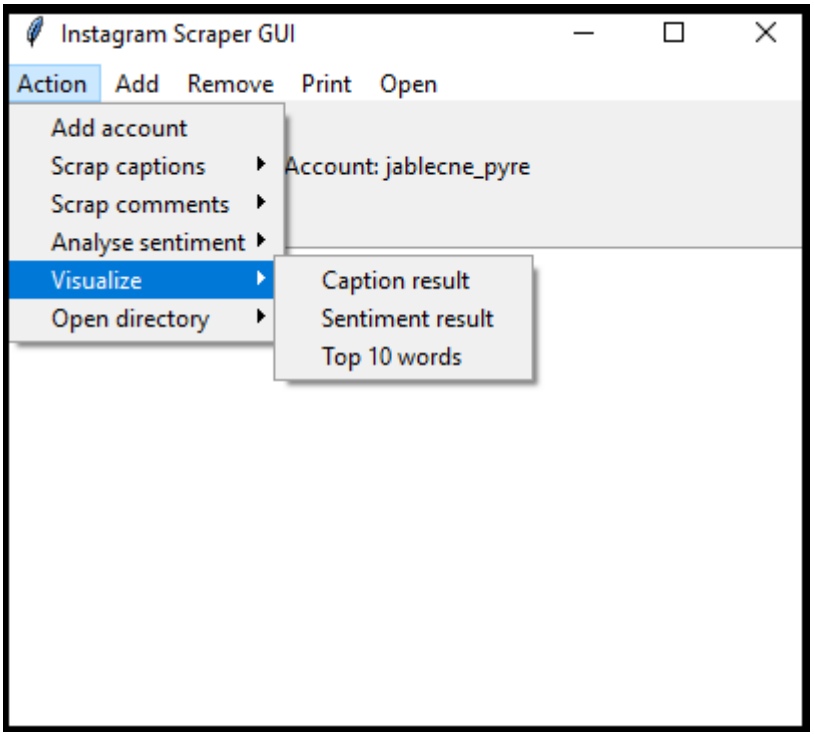
```
s = Service('chromedriver.exe')                                    1
driver = webdriver.Chrome(service=s)                               2
```

Výpis 5.1: Implementace ChromeDriveru

[1] https://googlechromelabs.github.io/chrome-for-testing/#stable

6 Úvodní menu

Úvodní menu nástroje (GUI) je koncipováno, aby bylo intuitivní a přístupné i pro uživatele, kteří nejsou technicky zdatní. Úvodní menu, je implementováno pomocí knihovny Tkinter a poskytuje strukturovaný přístup ke všem funkcionalitám. Podobu úvodního menu je možné vidět na obr. 6.1. Úvodní menu je v projektu tvořeno třídou `menuManager` a celou hierarchii projektu je možné vidět na obr. 6.2.



Obr. 6.1: Grafické rozhraní nástroje

6.1 Akce

Primární panel úvodního menu označený jako `Actions` (Akce), obsahuje klíčové operace nástroje mezi které patří následující možnosti:

- **Add account (přidání účtu)** – umožňuje uživatelům přidat Instagramový účet, který je použit pro interakci s platformou. Po zvolení je otevřeno vyskakovací okno v kterém uživatel údaje vyplní. Po vyplnění proběhne ověření, zda uživatelský účet existuje. Informace o aktuálně používaném účtu jsou zobrazeny v horní části GUI, což zvyšuje přehlednost o aktuálně používaném účtu.

- **Scrap captions (sběr popisků)** – funkce sbírá popisky a porovnává je vůči klíčovým slovům definovaných uživatelem. Popisky nástroj umožňuje sbírat z příspěvků, profilů, lokací nebo hashtagů. Výstupem funkce je `.csv` soubor v adresáři `keywords_summary`. Soubor obsahuje adresu profilu, hledané klíčové slovo a podrobné informace o daném příspěvku. Odborný popis vytvoření časové osy je popsán v sekci sběr dat.
- **Scrap comments (sběr komentářů)** – funkce sbírá komentáře z příspěvků, profilů, lokací či hashtagů. Výstupem je `.txt` soubor v adresáři `\comments` obsahující anonymizované uživatele a jejich komentáře. Výstupní soubor slouží jako vstup pro analýzu sentimentu. Podrobný popis funkcionality je popsán v sekci sběr dat.
- **Analyse sentiment (analýza sentimentu)** – funkce provádí analýzu sentimentu nad souborem vytvořeným v části sběr komentářů. Výstupem je `.txt` soubor v adresáři `\sentiment_analysis`, kde je ke každému komentáři přiřazen sentiment. Z tohoto souboru je vytvořena časová osa sentimentu v čase.
- **Visualize (vizualizace)** – funkce provádí vizualizaci výsledků. Nástroj se dotáže uživatele, jaký výsledek chce vizualizovat. Vizualizace může proběhnout pouze nad výstupním souborem ze sběru popisků nebo z analýzy sentimentu. Po výběru souboru se otvírá webový prohlížeč, v kterém je patřičný graf.
- **Open directory (otevřít)** – všechny výsledky z předchozích části jsou ukládány do specifických adresářů. Pomocí této operace je otevřen uživatelem definovaný adresář. Funkce slouží pro jednodušší správu a vyhledávání adresářů a jejich souborů.

6.2 Další funkcionality

Uživatelské rozhraní nabízí i možnosti pro manipulaci s daty, které slouží jako vstup pro funkce sběru dat. Mezi vstupní data patří uživatelské účty, klíčová slova, lokace, hashtagy, příspěvky. Uživatelské rozhraní obsahuje panely:

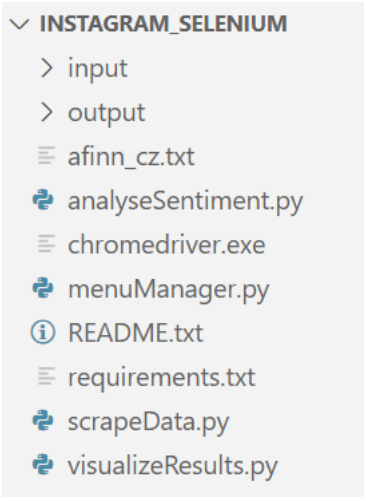
- `Add` (přidat),
- `Remove` (odebrat),
- `Print` (vypiš).

Panely `Add` a `Remove` slouží pro přidání či odebírání vstupních dat do souborů, které slouží jako vstup pro sběr dat. Panel `Print` slouží pro vypsání obsahu vstupního souboru do uživatelského rozhraní.

Uživatel nejprve vybere panel s akcí kterou chce provést. Následně vybere element s kterým manipuluje (profil, lokace, hashtag či příspěvek) a poté provede výběr, zda se jedná vstupní element k vytvoření souboru s popisky či sběr komentářů.

Uživatel může přidat více elementů zároveň. Pokud je element přidán pomocí úvodního menu je vše automatické. Další možností je přidat element manuálně do souboru, zde je nutné každý další element vložit na nový řádek.

V případě, že uživatel nadefinuje více klíčových slov, tak se budou v popisku příspěvku hledat všechny slova definována uživatelem. Pokud uživatel definuje více profilů, lokací, hashtagů nebo příspěvků bude porovnání vůči klíčovým slovům nebo sběr komentářů proveden nad všemi uvedenými.



Obr. 6.2: Hierarchie

7 Sběr a zpracování dat

Kapitola popisuje vytvoření a funkčnost jádra nástroje pro sběr veřejně dostupných dat. Sběr dat z Instagramu lze provádět jak přihlášený i nepřihlášený uživatel. Při sběru jako přihlášený uživatel, má uživatel přístup ke všem veřejně dostupným datům a není ze strany Instagramu omezován.

Určitá veřejná data však lze z Instagramu získat i bez přihlášení, má to však svá omezení. Nepřihlášeny uživatel musí znát přesné URL adresy příspěvků a profilů, které chce zobrazit. Bez přihlášení je nutné přistupovat na profil či příspěvek přímo přes daný odkaz. Další limitací je, že uživatel, který přistoupil na profil pomocí odkazu vidí náhledy příspěvků, ale nemůže si zobrazit jejich detail. Detail lze zobrazit opět pomocí přistoupení k příspěvku pomocí jeho přesného URL, zde je uživatel opět limitován, protože mu je znemožněno zobrazit komentáře. Uživatel je tedy opakovaně vyzýván k přihlášení, výzvu může odmítnout, ale pouze dočasně.

Z výše popsaných důvodu vytvořený nástroj provádí všechny akce jako přihlášený uživatel. Pokud je primárním cílem sběr data je vhodné mít připraveno více záložních účtů a ty pravidelně rotovat. Důležité je držet účty aktivní a provádět na nich i jinou aktivitu než je sběr dat, tak aby nedošlo k omezování či případnému zablokování účtu ze strany Instagramu.

7.1 Jádro programu

Jádro programu pracuje pomocí knihovny Selenium, ta umožňuje lokalizovat elementy na webu různými metodami, včetně použití XPath nebo CSS selektorů. Identifikátory nebo atributy elementů na se mohou změnit v důsledku aktualizací obsahu stránky. V takovém případě je nutné provést aktualizaci identifikátorů v kódu nástroje, aby nástroj nadále fungoval správně. Podrobnosti jaké prvky, kde je nalézt a jak je upravit, jsou uvedeny v přiloženém souboru README.

Pro obě možnosti sběru dat (scrap captions nebo scrap comments) má nástroj obdobné kroky, kde hlavní rozdílem je zde typ sbíraných dat (popisek vs. příspěvek). Kroky lze vidět na obr. 7.1. Popis postupných kroků jádra programu pro sběr dat:

7.1.1 Přihlášení do Instagramu

Pro přihlášení pomocí webového ovladače do Instagramu slouží funkce `login()` ve třídě `scrapeData.py`. Funkce nejprve otevře Instagram na úvodní adrese pro přihlášení. Následně najde XPATH element pro přijmutí souborů cookies, které potvrdí. Nástroj poté vyplní přihlašovací údaje včetně hesla a odešle je.

Uživatel je v případě úspěšného přihlášení informován v úvodním menu. Funkce používá `time.sleep` a `random.randrange` k simulaci lidského chování a zabránění detekce jako automatizovaný skript.

7.1.2 Načtení příspěvků cíle

Po úspěšném přihlášení načte nástroj cíl, který uživatel definoval v grafickém rozhraní. Cílem může být odkaz na profil, lokace nebo hashtag. Cíl si nástroj sám odvodí z uživatelem zadaného URL. URL obsahující:

- explore/tags = hashtagy
- explore/locations = lokace
- /p/ = příspěvek
- / = odkaz na profil

Po načtení cílového URL nástroj získá aktuální výšku stránky a poté postupně posouvá stránku dolů. Automatické posouvání se provádí z důvodu, aby nástroj načetl všechny URL odkazy na příspěvky, které cíl obsahuje. Proces posouvání se opakuje, dokud se výška stránky nezastaví na konstantní hodnotě, což signalizuje, že již nejsou další příspěvky k načtení.

Během popsaného procesu se shromažďují URL na jednotlivé příspěvky cíle. URL příspěvků jsou vyhledávány ovladačem pomocí XPATH výrazu. XPATH výraz, který nachází všechna adresy příspěvků:

```
xpath = "//a[contains(@href, '/p/') or contains(@href, '/reel/')]"
```

V případě sběru dat ze specifického příspěvku je otevřen definovaný příspěvek.

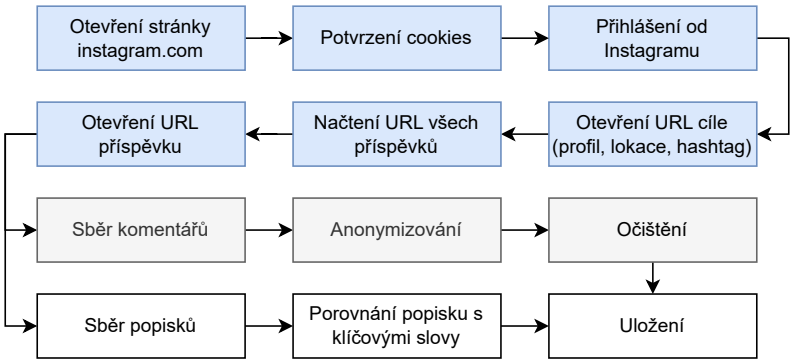
7.1.3 Procházení příspěvků profilu

Po shromáždění adres všech příspěvků jsou jednotlivé odkazy postupně otvírány. V případě sběru popisků po otevření adresy příspěvku nástroj načte popisek příspěvku pomocí XPATH výrazu. XPATH výraz, který obsahuje popisek příspěvku:

```
xpath = "//meta[@property='og:title']")
```

V případě nalezení alespoň jednoho z klíčových slov, které uživatel definoval v grafickém rozhraní, je získaný popisek uložen do výstupního souboru. Do souboru je společně s popiskem uložen také odkaz na příspěvek, klíčové slovo a datum vytvoření příspěvku.

V případě sběru komentářů po načtení adresy s příspěvkem nástroj nejprve načte všechny komentáře. Toho se dosáhne obdobně jako při předchozím kroku. Načte se výška okna v kterém jsou komentáře a nástroj postupně skroluje dolů, dokud se nenačtou všechny komentáře. Jakmile nejsou další komentáře k načtení tak jsou komentáře uloženy a anonymizovány. V výsledném souboru je anonymizovaný uživatel a jeho komentář.



Obr. 7.1: Kroky sběru volně dostupných dat

Názvy výstupních souborů jsou unikátní a jsou generovaný na základě cíle a URL. Jména souboru generuje funkce `makeOutPutFileName`. Například sběr komentářů na profilu @fektbrno vytvoří výsledný název souboru `comments_profile_fektbrno`. V případě sběru popisků na stránce @fektbrno je výsledný soubor pojmenován `descriptions_profile_fektbrno`.

7.2 Přiřazení sentimentu

Přiřazení sentimentu je implementováno v třídě analyseSentiment.py. V úvodním menu je uživateli zpřístupněno pomocí Action -> Analyse Sentiment. Uživateli je umožněno zvolit mezi čtyřmi druhy slovníku.

Slovník AFINN_CZ využívá databáze slov s přiřazenou polaritou a je určen pro přiřazení sentimentu komentářů v českém jazyce. TextBlob, Vader a AFINN jsou určeny pro sentiment anglického jazyka a poskytují jednoduché API rozhraní pro jejich použití. Každý slovník je vhodný pro jiný kontext a je tedy pouze na uživateli jaký slovník zvolí.

Po výběru slovníku je uživateli otevřena adresář s komentáři a provede výběr souboru, kterému chce sentiment přiřadit. Po otevření souboru jsou komentáře zpracovány tak, aby k nim bylo možné přiřadit co nejvíce odpovídající sentiment.

Po zpracování je komentářům přiřazen sentimentu a soubor s výsledným sentimentem je uložen do adresáře `\sentiment_analysis`.

Kroky zpracování komentářů:

- Odstranění čísel,
- odstranění nepotřebných znaků (’“”...),
- odstranění emotikon a dalších symbolů,
- odstranění zkrácených forem (isn’t -> is not),
- odstranění opakujících se písmen (supeeeeer -> super),
- odstranění většího počtu mezer.

7.2.1 TextBlob

Poskytuje jednoduché rozhraní API pro běžné úlohy zpracování přirozeného jazyka (NLP). Analýza sentimentu vrací polaritu a subjektivitu. Polarita leží v rozmezí $[-1, 1]$, přičemž -1 označuje negativní sentiment a +1 pozitivní sentiment. Subjektivita leží v rozsahu $[0, 1]$. Subjektivní věty se obecně vztahují k názoru, emocím nebo úsudku. Textblob nebere během určování sentimentu v úvahu slova, která nezná. Knihovna je v nástroji použita tak, že vrací pouze sentiment. [41] Implementaci funkce slovníku lze vidět ve výpisu 7.1.

```
def get_textblob_score(comment):
analysis = TextBlob(comment)
return analysis.sentiment.polarity
```

Výpis 7.1: Příklad implementace slovníku TextBlob

7.2.2 Vader

Nástroj vrací pravděpodobnost pomocí slovníku. Pravděpodobnost popisuje jaká je šance, že význam věty je pozitivní, negativní nebo neutrální. Dalším výstupem může být také složení těchto třech pravděpodobnosti (compound). Nástroj Vader je optimalizován pro data ze sociálních médií a může přinést dobré výsledky, pokud se použije s daty z Twitteru, Facebooku atd. Vader v nástroji vrací složenou hodnotu všech pravděpodobností. [42] Implementaci funkce slovníku lze vidět ve výpisu 7.2. Příklad – „This car is bueatiful“:

```
{’compound’: 0.6588, ’neg’: 0.0., ’neu’; 0.406, ’pos’: 0.594}
```

```
def get_vader_score(comment):
analyzer = SentimentIntensityAnalyzer()
return analyzer.polarity_scores(comment)[’compound’]
```

Výpis 7.2: Příklad implementace slovníku Vader

7.2.3 AFINN

AFINN je nejpoužívanější nástroj pro analýzu sentimentu. Slova jsou hodnocena v rozmezí $(5, 5)$. Neutrální věty a fráze jsou hodnocena nulou (0), nejvíce pozitivní (+5), nejvíce pozitivní (−5). Nástroj podporuje více jazyků (dánština, finština, švédština a turečtina). [43] AFINN vypočítá složené skóre pro každou větu sečtením vah jednotlivých slov sentimentu. [44] Příklad: „i hate cars“ = (−3.0). Implementaci funkce slovníku lze vidět ve výpisu 7.3.

```
def get_afinn_score(comment):
afinn = Afinn()
return afinn.score(comment)
```

Výpis 7.3: Příklad implementace slovníku AFINN

7.2.4 AFINN_CZ

Jedná se o jedinou volně dostupnou českou databázi slov určených k analýze sentimentu. Obsahuje téměř 20 000 slov. Slova oproti anglické verze neleží v rozsahu $[-5, +5]$, ale mají přiřazené hodnoty (−1 pro negativní sentiment a (+1) pro pozitivní sentiment. [46] Ukázka slovníku je viditelná na obr. 7.2. Databáze je v nástroji implementována pomocí dvou funkcí.

Funkce `load_sentiment_dictionary` načítá slovník sentimentů z CSV souboru, kde klíče jsou česká slova a hodnoty jsou jejich polarity.

Funkce `calculate_sentiment_cz` prochází slova ve větě. Pokud je ve větě nalezeno slovo ze slovníku sentimentů, jeho polarita je přidána do seznamu polarit. Následně se vypočítá průměrná polarita. V případě, že je seznam polarit prázdný, funkce vrací None. Pokud seznam polarit není prázdný polarita se vypočítá následovně: `average_polarity = sum(polarities) / len(polarities)`

V případě, že věta obsahuje 10 slov, z toho jsou čtyři slova ve větě označena ve slovníku pozitivně (+1) a jedno slova je označeno negativně (−1), výsledný sentiment je roven 0,5.

```
id,word_cz,polarity
84,nastavitelný,1
266,kasovní,1
280,zbrusu,1
289,vánek,1
380,doplňkový,1
389,stručný,1
400,důsledně,1
```

Obr. 7.2: Ukázka AFINN_CZ slovníku

7.3 Vytvoření dashboardu pro vizualizaci dat

Cílem vizualizace dat je vytvořit graf časové osy pro výsledný sentiment a počet použití klíčových slov v popiscích příspěvků. Vizualizace je tvořena souborem `vizualizeResult.py` a je vyvinuta s použitím knihoven Plotly a Dash. Tyto knihovny umožňují efektivní manipulaci s daty a interaktivní vizualizaci výsledků.

Dashboard je přístupný prostřednictvím webového prohlížeče a je navržen tak, aby byl intuitivní i pro uživatele bez technických znalosti. Uživateli je zobrazen dashboard obsahující interaktivní graf časové osy. V případě najetí kurzory na bod grafu se zobrazí metadata o příspěvků. Pod grafem je tabulka, která zobrazuje všechny hodnoty ze vstupního souboru.

7.3.1 Spuštění a přístup k dashboardu

Spuštění dashboardu probíhá v úvodním v menu `Action - Visualize`. Po výběru cíle, který má být zobrazen nástroj automaticky otevírá webový prohlížeč na adrese `http://127.0.0.1:8050/`, kde je dashboard dostupný. Tento přístup umožňuje uživatelům ihned interagovat s analýzou dat bez nutnosti manuálního zásahu.

7.3.2 Analýza klíčových slov v příspěvcích

V případě výběru `Keywords result` v úvodním menu, funkce `run_dashboard_keyword` pracuje s vybraným souborem z adresáře `\kewords_summary`, který obsahuje data o příspěvcích obsahující uživatelem definovaná klíčová slova. Data jsou podle specifického klíčového slova filtrována a následně agregována na měsíční úroveň. Aplikace vytváří graf počtu zmínek klíčového slova v čase a tabulku s detailními údaji o jednotlivých příspěvcích, včetně počtu komentářů a odkazů na příspěvky. Výsledný dashboard je viditelný na obr. 8.9.

7.3.3 Analýza sentimentu

V případě výběru Visualize -> Sentiment result, funkce `run_dashboard_sentiment` čte data ze vstupního `.txt` souboru, který obsahuje komentáře a sentimenty anoanymizovaných uživatelů. Data jsou převedena do měsíčního průměru sentimentu, což umožňuje sledovat změny nálady uživatelů v čase. Dashboard zobrazuje interaktivní graf měsíčního průměru sentimentu a tabulku s detaily o jednotlivých komentářích, datu, skóre sentimentu. Výsledný dashboard je viditelný na obr. 6.1.

7.3.4 Analýza komentářů

V případě výběru Visualize -> Top 10 words, funkce `run_dashboard_topten` pracuje s již očištěnými komentáři v adresáři `\sentiment_analysis`. Data jsou zpracována za účelem identifikace nejčastěji používaných slov v komentářích. Komentáře jsou rozdělena na slova, přičemž se do výsledku nepočítají tzv. „stop words“. „Stop words“ obsahují předložky, částice, spojky atd. [45]. Slova se porovnávají vůči seznamu, který je dostupný na adrese.[1]. Následně jsou slova seřazena podle frekvence výskytu a deset nejčastějších slov je zobrazeno v grafu. Výsledný dashboard je automaticky otevřen v webovém prohlížeči. Výsledný dashboard je viditelný na obr. 8.6.

[1] https://countwordsfree.com/stopwords/czech

8 Testování

Kapitol popisuje fázi testování. Cílem testování je ověření, že nástroj splňuje předem definované požadavky a je schopen efektivně analyzovat a prezentovat data.

Před prvním použitím je nutné nainstalovat potřebné balíčky a nainstalovat Vader slovník. Instalace se provádí pomocí následujících příkazů:

```
pip install -r .\requirements.txt
python -m nltk.downloader vader_lexicon
```

Uživatelské rozhraní lze spustit pomocí příkazu:

```
python .\menuManager.py
```

Před prvním použitím nástroje je nutné přidat uživatelský účet pod kterým bude sběr veřejně dostupných dat provádět. V uživatelském rozhraní je možné přidat účet pomocí `Actions -> Add account`. Další možností je přidat účet manuálně do souboru `\account.txt`.

Dále je nutné definovat v uživatelském rozhraní cíl sběru dat (profil, lokace, hashtag nebo příspěvek). V případě analýzy popisků vůči klíčovým slovům je nutné definovat i klíčová slova. V uživatelském rozhraní je přidání možné za použití `Add` a následně vybrat cíl.

8.1 Časová osa klíčových slov v příspěvcích

V souladu s kriteriem, aby nástroj necílil na jednotlivé uživatele (což by mohlo vést k porušení etických principů) byl nástroj podroben spuštěn na veřejně dostupných datech z Instagramového profilu @fektbrno. Tento profil byl zvolen díky jeho širokému spektru obsahu, ale také proto, že fakulta pořádá každoroční hudební akci Hudba z FEKTu v měsíci září. Tato hudební akce poskytla ideální příležitost pro ověření funkčnosti nástroje v reálných podmínkách a zkoumání jeho schopnosti zachytit a klíčová slova v časové sérii dat.

Pro získání výsledku byly použity následující kroky:

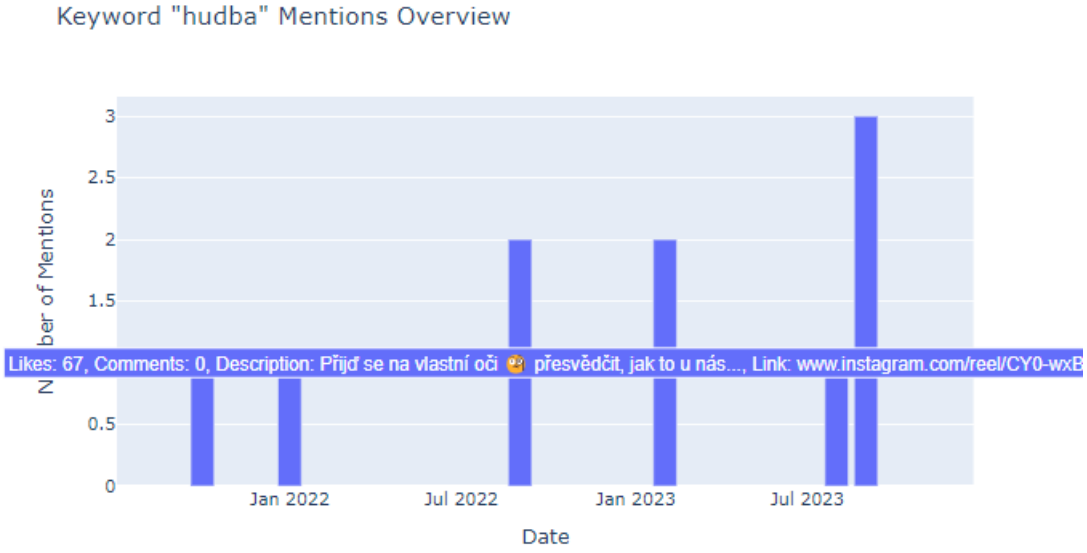
- Přidání názvu profilu pro sběr – Add -> Profile -> Timeline file -> @fektbrno
- Sběr a analýza popisků – Action -> Scrap keywords -> From profile
- Zobrazení časové osy – Action -> Sentiment result -> výběr souboru

output > fekt_vystup.csv

```
1  Profile Name;Keyword;Date;Post Link;Post Description
2  www.instagram.com/fektbrno/;hudba;2023-09-21;www.instagram.com/p/CxdyYuEIIjX/;"Přinášíme fotky z 15. r
3  www.instagram.com/fektbrno/;hudba;2022-09-09;www.instagram.com/reel/CiSHltCq0fa/;"Přijď za námi na @hu
4  www.instagram.com/fektbrno/;hudba;2023-09-14;www.instagram.com/reel/CxKl87cqsCs/;"Nenechej si na začát
5  www.instagram.com/fektbrno/;hudba;2023-09-25;www.instagram.com/reel/CxnFt0NKKc7/;"🔥Hudba z FEKTu 202
6  www.instagram.com/fektbrno/;hudba;2022-01-17;www.instagram.com/reel/CY0-wxBvnJr/;"Přijď se na vlastní
7  www.instagram.com/fektbrno/;hudba;2021-10-05;www.instagram.com/p/CUpuW27IhvC/;"Třináctý ročník ℹ obli
8  www.instagram.com/fektbrno/;hudba;2023-08-09;www.instagram.com/p/CvuacAbIU4d/;"Roj 80 robotických „své
9  www.instagram.com/fektbrno/;hudba;2023-02-13;www.instagram.com/p/ComnmWvIlp8/;"V pátek proběhl druhý a
10 www.instagram.com/fektbrno/;hudba;2022-09-22;www.instagram.com/p/CizHtNEIwZ-/;"Přinášíme první fotky z
11
```

Obr. 8.1: Výstupní soubor profile_fektbrno.csv

Výstupem kroku „sběr a analýza popisku“ je textový soubor na obr. 8.1. Textový soubor byl použit jako vstup pro krok zobrazení časové osy. Výsledná časová osa je na obr. 8.2, a poskytuje kvantitativní analýzu frekvence výskytu klíčového slova „hudba“ ve fakultních příspěvcích v časovém rozmezí od listopadu 2021 do září 2023. Zaznamenány jsou patrné výrazné změny v počtu zmínek klíčového slova, které je třeba podrobněji prozkoumat a posoudit jejich důvod. V letních měsících, konkrétně od července do září, lze pozorovat pravidelné zvyšování počtu příspěvků, což odpovídá pravidelným sezónním trendům, které mohou souviset s akademickými nebo kulturními akcemi na fakultě. Dalším zvyšování počtu příspěvků lze pozorovat v lednu 2021 a 2022. Z pohledu analytika by měly být tyto jevy podroben podrobnější statistické analýze, aby se potvrdila jeho pravidelnost a identifikovaly příčiny.



Obr. 8.2: Frekvence slova „hudba“ v popise příspěvků

Ze statické analýzy vyplynulo, že v případě pravidelného nárůstů aktivit souvisejících s hudbou v období od července do září koreluje s každoročním konáním Hudby z FEKTu, která zvyšuje zapojení a interakci s veřejností. Tato akce je mar-

ketingově významná, jelikož zvýšená publicita může efektivně přispívat k dosažení větší návštěvnosti této událostí.

Z analýzy lednových výkyvů je zřejmé, že fakulta využívá strategické komunikace zaměřené na potenciální kandidáty během období podávání přihlášek. Tato taktika je navržena tak, aby představila akademické a kulturní možnosti, které fakulta nabízí, a posílila tak její atraktivitu pro nové studenty. Jednou z významných kulturních možností je právě Hudba z FEKTu.

Vzhledem k tomu, že aktivita na fakultním Instagramu vzrostla až v posledním roce, je patrné, že se zvýšil i počet příspěvků, což odráží záměrné zvýšení online prezentace fakulty.

8.2 Sběr komentářů

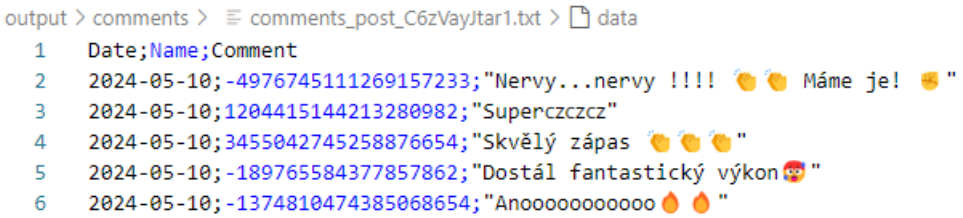
Cílem této části je otestovat fungování nástroje při sběr komentářů pod profilem, příspěvkem, lokací či hashtagem. Pro testování byly nalezeny vhodné profily a příspěvky.

8.2.1 Příspěvek

Pro testování byly vybrány dva příspěvky s potencionálně rozdílným sentimentem.

První použitý příspěvek oznamoval vítězství českého národního týmu v hokeji, U takového příspěvku existuje předpoklad, že sentiment bude pozitivní. Příspěvek byl nalezen na profilu @sport_ct.

Druhý příspěvek se týkal reformy důchodového systému. V posledních měsících se v jedná o diskutované a choulostivé téma, zde je předpoklad, že výsledný sentiment příspěvku bude spíše negativní.



Obr. 8.3: Výstup části sběr komentářů

Pro získání výsledku byly použity následující kroky:

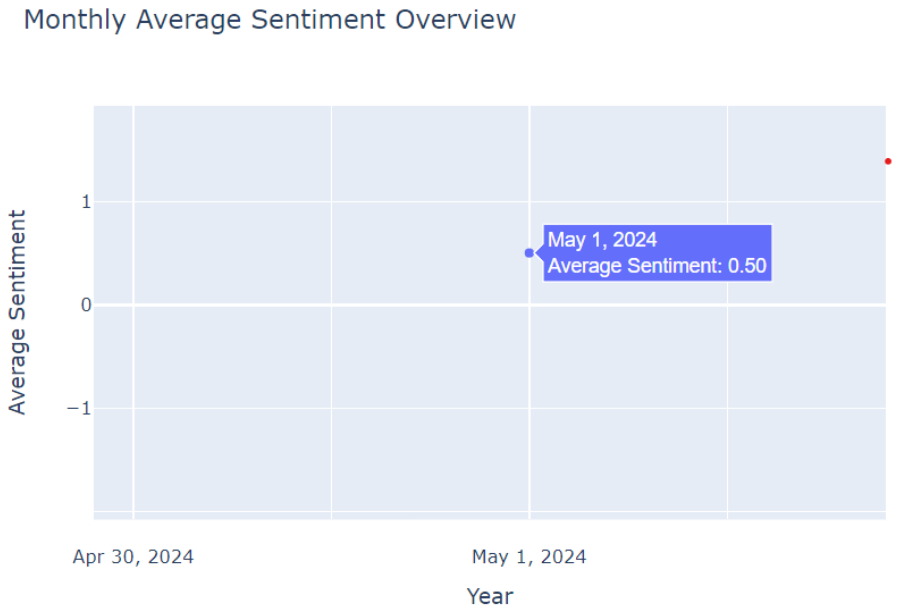
- Přidání odkazu příspěvku – Add -> Post -> odkaz na příspěvek.
- Sběr komentářů – Action -> Scrap comments -> From post.
- Přidání sentimentu – Action -> Analyse sentiment -> AFINN_CZ
- Zobrazení časové osy – Action -> Sentiment result.

output > sentiment_analysis > comments_post_C6zVayJtar1.txt > data

```
1  Date,Name,Cleaned_Comment,Sentiment_Score
2  2024-05-10,-4976745111269157233,nervynervy máme je,
3  2024-05-10,1204415144213280982,super,1.0
4  2024-05-10,3455042745258876654,skvělý zápas,1.0
5  2024-05-10,-189765584377857862,dostál fantastický výkon,1.0
6  2024-05-10,-1374810474385068654,ano,1.0
```

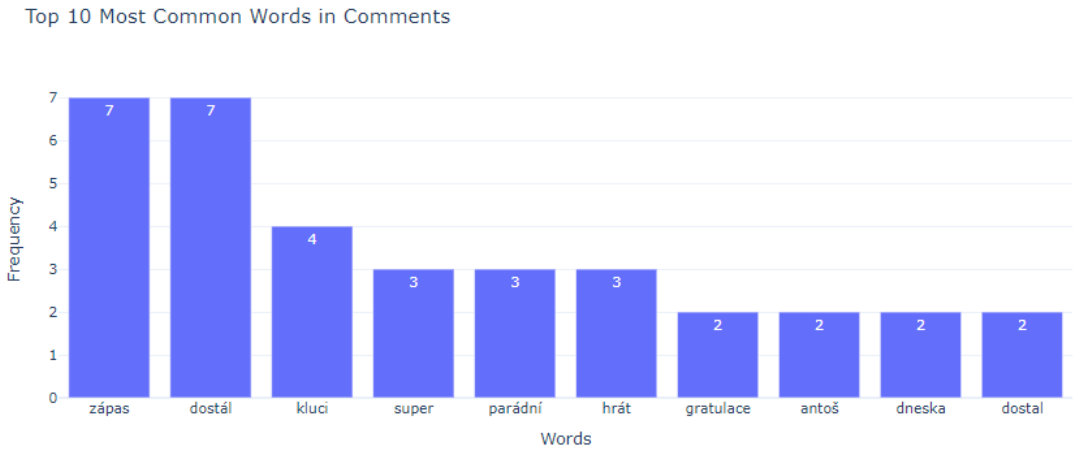
Obr. 8.4: Výstup části přiřazení sentimentu

Výstupní soubor sběru komentářů je zobrazen na obr. 8.3. Podoba souboru po očištění a přiřazení sentimentu je zobrazena obr. 8.4.



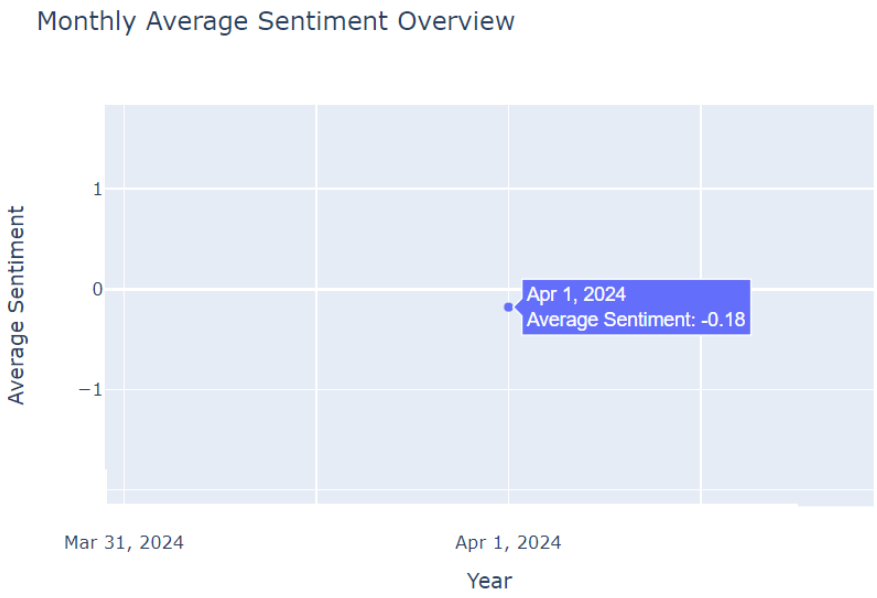
Obr. 8.5: Sentiment vítěství v hokeji

Průměrný výsledek u příspěvku s hokejem je na pozitivní úrovní (+0,5). Výsledek lze vidět na obr. 8.5. Dalším krokem je analýza nejpoužívanějších slov v komentářích. Nejpoužívanější slova je možné vidět ve obr. 8.6. Z těchto výsledku je patrné, že výhra vyvolala pozitivní emoce, nejvíce skloňovaným slovem bylo jméno gólmana. Z tohoto důvodu lze usuzovat, že jeho výkon je veřejností chválen. Komentáře plné slov jako „super“ a „parádní“, které odrážely radost z výhry a celkovou spokojenost s tím, jak hráči („kluci“) bojovali na ledě. Mnozí vyjádřili obdiv slovy „klobouk dolů“, což svědčí o vysokém respektu k hráčům. Výhra byla vnímána jako skvělý týmový úspěch, a proto se mezi fanoušky šířil pozitivní sentiment a hrdost na národní tým.

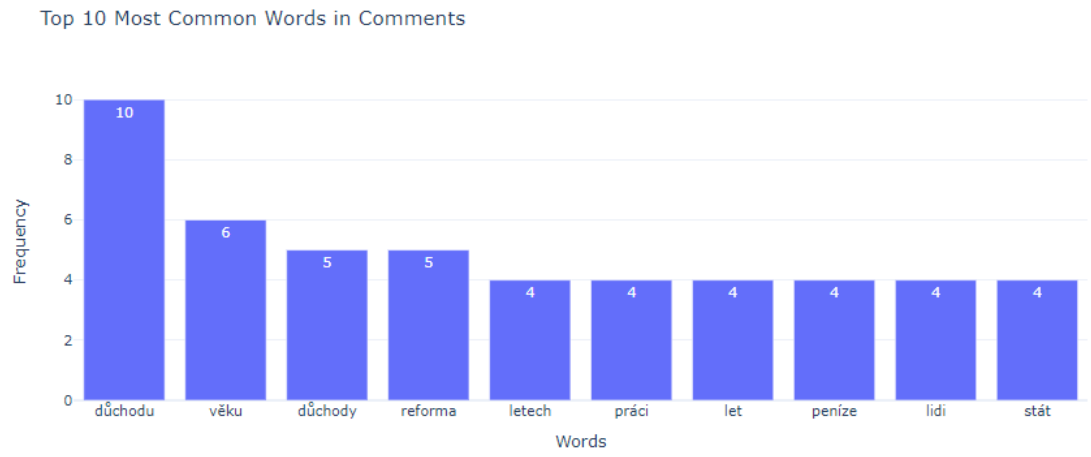


Obr. 8.6: Nejčastější slova - příspěvek hokej

U důchodové reformy je výsledný sentiment negativní (-0,18). Z výsledku lze usuzovat, že nálada společnosti pod tímto příspěvkem je rozporuplná. Výsledek lze vidět na obr. 8.7. Nejčastější slova lze vidět na obr. 8.8. Časté zmínky o „důchodu“ a „důchodcích“ ukazují, že mnoho lidí má obavy z toho, jak změny ovlivní jejich budoucnost. Komentáře se často zaměřovaly na zvýšení „věku“ odchodu do důchodu, což mnozí vnímají jako nespravedlivé vůči starším generacím, které již mnoho let pracují. Kritika směřovala také na „vládu“ a její přístup k reformě. Přičemž lidé vyjadřovali frustraci a pocit, že stát nebere v úvahu jejich potřeby. Celkově se v komentářích odrážely obavy z nejisté budoucnosti a nespokojenost s navrhovanými změnami.



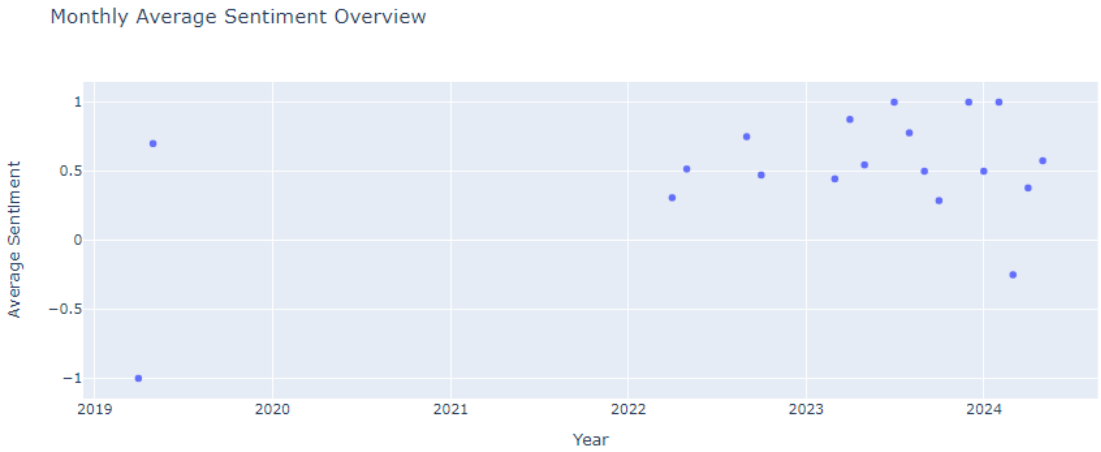
Obr. 8.7: Sentiment reforma důchodového systému



Obr. 8.8: Nejčastější slova - příspěvek reforma důchodového systému

8.2.2 Testování profilu, hashtagu a lokace

Cílem části bylo podrobit nástroj testování na profilu, kde postupně projde všechny příspěvky z kterých následně extrahuje komentáře. Sběr byl proveden pod profilem @majalesbrno, který má 370 příspěvků. Jedná se o oficiální profil hudebního festival v Brně, který se koná každoročně a cílí na mladší generaci, která často komentuje příspěvky svých oblíbených či neoblíbených interpretů.



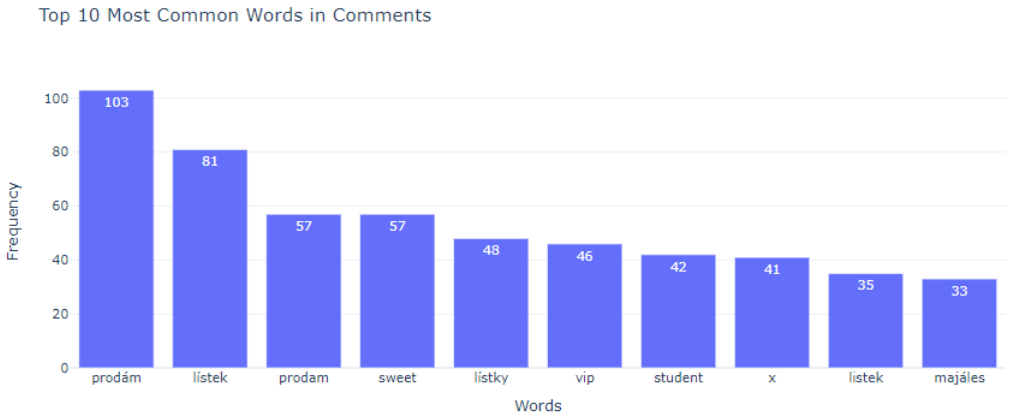
Obr. 8.9: Sentiment - profil @majalesbrno

Z výsledné časové osy sentimentu čase na obr. 8.9. Je patrné, že převládá pozitivní sentiment. Lidé často sdílejí své pozitivní zážitky z akcí Majáles. Uživatelé v komentářích také často oslavují oznámení vystoupení jejich oblíbeného interpreta na festivalu. Uživatelé také chválí organizátory za dobře uspořádané akce a celkovou atmosféru festivalu. Profil často sdílí vizuálně atraktivní a zajímavý obsah, což vede k pozitivním reakcím a komentářům od sledujících. Dalším důvodem pozitivního

sentimentu je časté pořádání soutěží o vstup zdarma, kde lidé často označují své přátele, což zvyšuje počet komentujících a pozitivní sentiment v komentářích.

Z časové osy je patrný chybějící sentiment pro období 2019-2022. Absence komentářů v letech 2019-2022 je pravděpodobně způsobena pandemií COVID-19. V této době byla většina velkých veřejných akcí, včetně festivalů, zrušena nebo odložena, což vedlo k nižší aktivitě na sociálních sítích a menšímu počtu komentářů.

Nejčastější slova v komentářích jsou zobrazena na obr. 8.10. Slova „prodám“ a „lístek“ naznačují, že lidé často využívají komentáře k prodeji vstupenek. „Sweet“, „vip“ a „student“ ukazují na zájem nebo prodej speciálních typů lístků pro mladší generaci. „X“ je používáno při prodeji lístků mezi uživateli a vyjadřuje počet lístků. Z výsledku nejčastěji objevujících slov je patrné, že jsou komentáře využívány z hlavní části právě k prodeji lístků.



Obr. 8.10: Nejčastější slova - profil @majalesbrno

8.2.3 Testování slovníkového přístupu

Cílem bylo porovnat tři anglické slovníky: AFINN, VADER a TextBlob. Cílem je zjistit, který z těchto slovníků poskytuje nejpřesnější hodnocení sentimentu a zda je slovníkový přístup v praxi použitelný. Slovníky byly testovány dvou dostupných datasetech.

První dataset obsahoval nenávistné komentáře ze sociálních sítí. V tomto datasetu byl sentiment komentářům přidělen manuálně a nabýval hodnotu N pro negativní (nenávistný projev) a P pro pozitivní sentiment (pozitivní projev). [47] Výběr datasetu byl motivován rostoucím zájmem o problematiku nenávistných výroků na sociálních sítích v poslední době a cílem bylo ověřit, jak slovníky dokáží problematiku zpracovat. Sentiment přiřazený slovníky byl porovnán vůči sentimentu přiřazeným člověkem, výsledek popisuje tab. 8.1.

Slovník	**Úspěšnost (%)**	**Označené**	**Neoznačené (%)**
AFINN	40.95	3347	24.70
VADER	40.15	2911	21.48
TextBlob	37.01	4083	30.13

Tab. 8.1: Porovnání slovníků nenávistné komentáře

Druhý dataset obsahoval textové recenze aplikace Isntagram z Google Play. Aplikaci hodnotili uživatele od 1 do 5 hvězdiček. [48] Hodnocení 4 a 5 hvězdiček bylo klasifikováno jako pozitivní, hodnocení třemi hvězdičkami jako neutrální a hodnocení 1 a 2 hvězdičky jako negativní. I v případě druhého datasetu byl sentiment přiřazený slovníky porovnán vůči sentimentu přiřazeným člověkem při hodnocení aplikace, výsledek popisuje tab. 8.2.

Slovník	**Úspěšnost (%)**
AFINN	54.96
VADER	57.25
TextBlob	51.58

Tab. 8.2: Porovnání slovníků - recenze z obchodu play

Jedním z hlavních důvodů nízké přesnosti slovníků u nenávistných komentářů je obtížnost identifikace sarkasmu a satiry. Tyto formy vyjadřování často používají pozitivní slova k vyjádření negativního sentimentu, nebo naopak. Tyto důvody mohou vést k nesprávné interpretaci kontextu. Například věta: „Skvělá práce, opravdu!“ může být sarkastická a ve skutečnosti znamenat opak.

Rozmanitost jazyka používaného na sociálních sítích také snižuje přesnost těchto slovníků. Lidé používají slang, zkratky, emotikony a další nestandardní formy vyjadřování, které slovníky nemusí správně interpretovat.

Dalším faktorem je kontext, který slovníky často neberou v úvahu. Význam některých slov závisí na předchozích nebo následujících větách, což je pro slovníky obtížné správně interpretovat. Například věta „Nemyslím si, že je to špatný člověk, ale...“ může být negativní v kontextu celého textu, ale slovník tuto často přehledné.

Výsledky ukazují, že sentimentové slovníky AFINN, VADER a TextBlob dosáhly vyšší přesnosti při analýze sentimentu v datasetu, který obsahuje hodnocení aplikace Instagram. Hlavímy důvody jsou konkrétní kontext a použité menšího počtu sarkasmu a satiry. Negativní komentáře často obsahují slova jako bug, crash, problem, které slovníky dokáží lépe identifikovat. Pozitivní recenze obsahují slova jako good, love nebo best, což také přispívá k přesnosti.

Výsledkem testování nejsou ideální. Přesnost sentimentových slovníků zůstává omezená a pro dosažení lepších výsledků by bylo nutné použít pokročilejší přístupy založené na strojovém učení. I tam však úspěšnost dosahuje maximálně kolem 80 %. Tyto výsledky ukazují, že sentimentové slovníky nejsou dostatečně přesné pro komplexní analýzu sentimentu a pro použití v praxi a je třeba použít sofistikovanější metody a modely.

8.3 Přednost, slabina, příležitost a hrozby

Pro lepší pochopení získaných dat byla provedena analýza přednost, slabina, příležitost a hrozby.

- **Silné stránky** – data z Instagramu jsou široká a rozmanitá, zahrnují různé typy popisků příspěvků a komentářů, což umožnuje analyzovat trendy. Data jsou aktuální a relevantní, poskytují přesný obraz o současných náladách a názorech uživatelů. Různorodé zdroje dat (profily, lokace a hashtag) umožňují detailnější analýzu.
- **Slabé stránky** – data z Instagramu nemusejí plně reprezentovat širší populaci, protože se zaměřují pouze na uživatele této platformy. V případě používají sarkasmu a ironie analýza sentimentu nemusí vždy odrážet skutečné pocity a názory uživatelů.
- **Příležitosti** – data mohou být využita pro hlubší analýzy, jako je identifikace skrytých vzorců nebo analýza vlivů. V případě rozšíření na další sociální sítě by data mohla poskytnout komplexnější pohled na veřejné mínění. Data mohou být také základem pro vytvoření efektivnějších marketingových kampaní na specifické cílové skupiny.
- **Hrozby** – jednou z hlavních hrozeb je rychlé zastarávání dat, protože trendy na sociálních sítích se mění. Další hrozbou je možnost zneužití získaných dat, například pro šíření dezinformací nebo narušení soukromí uživatelů. Rovněž může dojít ke zkreslení výsledků, pokud se objeví náhlé a masivní kampaně, které uměle nafouknou určité trendy nebo sentimenty.

Závěr

Diplomová práce byla zaměřena návrh a implementaci nástroje pro sběr informací z otevřených zdrojů na sociální síť Instagram. Cílem teoretické části bylo porozumět současné situaci a na základě získaných poznatků vytvořit nástroj reflektující zadané potřeby.

Teoretická část obsahuje popis disciplín spojených se zpravodajstvím a zahrnuje vysvětlení zpravodajského cyklu. Zpravodajský cyklus je klíčovým prvkem nejen pro Open Source Intelligence (OSINT), ale i pro všechny ostatní zpravodajské obory. V teoretické části jsou popsány služby, techniky a nástroje pro sběr veřejně dostupných dat. V teoretické části se také objevuje popsání různých metod určování sentimentu v komentářích a detailní rozbor sociální sítě Instagram.

V praktické části byl vytvořen nástroj pro sběr a analýzu veřejně dostupných dat z platformy Instagram. Vytvořený nástroj umožňuje sběr dat z vybraných uživatelských profilů, lokací, hashtagů a příspěvků. Mezi sbíraná data patří popisky a komentáře příspěvků. Nástroj následně umožňuje analýzu sentimentu a vizualizaci výsledků prostřednictvím interaktivních grafů. Díky interaktivních grafům lze monitorovat a analyzovat trendy v příspěvcích a komentářích.

Testování nástroje na skutečných datech prokázalo jeho schopnost efektivně rozpoznat a analyzovat trendy. Analýza ukázala, že nástroj je schopen identifikovat sezónní trendy nebo zvýšenou aktivitu spojenou s konkrétními událostmi pomocí analýzy popisků příspěvků. Dále proběhla analýza sentimentu komentářů při příspěvcích s pozitivním i negativním obsahem, což umožnilo srovnat reakce veřejnosti na různé typy událostí. Výsledky potvrdily, že nástroj dokáže rozpoznat pozitivní i negativní emoce, což může pomoci při predikci veřejného mínění a hledání potenciálních nebezpečí.

Vytvořený nástroj úspěšně splnil všechny požadavky a je efektivním řešením pro monitorování a vyhodnocování dat z Instagramu. Testování sentimentu pomocí různých slovníků ukázala, že výsledky tohoto přístupu ne vždy postačují pro použití v praxi, což naznačuje nutnost použití pokročilejších metod. V budoucnu by bylo vhodné nástroj dále aktualizovat a vylepšovat, například integrací metod strojového učení pro přesnější analýzu sentimentu.

Literatura

[1] ZETOCHA, Karel. *Úvod do studia zpravodajských služeb, Vojenské rozhledy.* Online. In: Vojenské rozhledy. 2006. Dostupné z URL: <https://www.vojenskerozhledy.cz/kategorie-clanku/teorie-a-doktriny/uvod-do-studia-zpravodajskych-sluzeb>. [cit. 2023-10-24].

[2] HAVRÁNEK, Daniel. *Právní rámec spolupráce zpravodajských služeb a Policie České republiky.* In: Správní právo. ISSN 0139-6005. Praha: Ministerstvo vnitra ČR, 2014, 2014, roč. 47, č. 3, s. 167-179. [cit. 2023-10-24].

[3] ČESKO. *Zákon č. 153/1994 Sb. Zákon o zpravodajských službách České republiky.* Online. In: Zákony pro lidi. AION CS, 1994-2023. Dostupné z: <https://www.zakonyprolidi.cz/cs/1994-153>. [cit. 2023-10-24].

[4] Maltego Team. *Understanding the Different Types of Intelligence Collection Disciplines.* Online, blog. 27. 11. 2022 Dostupné z URL: <https://www.maltego.com/blog/understanding-the-different-types-of-intelligence-collection-disciplines/>. [cit. 2023-10-27]

[5] BAZZELL, Michael. *Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information.* 2023. ISBN 979-8794816983. [cit. 2023-10-25].

[6] BAKER, L. Rae a HOFFMAN, Micah. *Deep Dive: Exploring the Real-world Value of Open Source Intelligence.* 2023. ISBN 978-1119933243. [cit. 2023-10-25]

[7] Sinwindie. *OSINT & The Intelligence Cycle Part IV: Analysis and Production* Online. In: secjuice. 8. 11. 2020. Dostupné z URL: <https://www.secjuice.com/osint-the-intelligence-cycle-part-iv-processing-raw-intelligence/>. [cit. 2023-10-27].

[8] PASTOR-GALINDO, J.; NESPOLI, P.; MÁRMOL, G. F. A PÉREZ, M. G. *The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends.* In: IEEE Access. s. 10282 - 10304. 2020. Dostupné z URL: <https://doi.org/10.1109/ACCESS.2020.2965257>. [cit. 2024-05-10].

[9] NORDINE, Justin. *OSINT Framework.* Online. In: osintframework. 2022. Dostupné z URL: <https://osintframework.com/s>. [cit. 2023-10-28].

[10] FREDA, Anthony. *Google Dorks: What Are They and How Are Google Hacks Used?* Online, blog. In: AVG Signal Blog. 14. 11. 2022. Dostupné

z URL: <https://www.avg.com/en/signal/google-dorks#:~:text=Google%20dorking%2C%20also%20called%20Google,reveal%20hidden%20parts%20of%20websites>. [cit. 2023-10-28]

[11] Maltego Support. *What is Maltego?* Online. In: Maltego FAQs. 8. 4. 2024 <https://docs.maltego.com/support/solutions/articles/15000019166-what-is-maltego->. [cit. 2024-05-10].

[12] Maltego Team. *Pricing.* Online. In: Maltego. 2024. Dostupné z URL: <https://www.maltego.com/pricing-plans/>. [cit. 2024-05-10].

[13] ShadowDragon LLC. *ShadowDragon SocialNet.* Online. In: Maltego. 2023. Dostupné z URL: <https://www.maltego.com/transform-hub/socialnet/>. [cit. 2023-11-03].

[14] Social Links Inc. *Social Links Professional.* Online. In: Maltego. 2023. Dostupné z URL: <https://www.maltego.com/transform-hub/social-links-pro/>. [cit. 2023-11-03].

[15] TABATABAEI, Fahimeh; WELLS, Douglas. OSINT in the Context of Cyber-Security. Open Source Intelligence Investigation: From Strategy to Implementation, Online. 2016, 213-231.

[16] PASTOR-GALINDO, J.; P. NESPOLI; MÁRMOL G. F. A PÉREZ M. G. *OSINT is the next Internet goldmine: Spain as an unexplored territory.* Online. In: ResearchGate. 2019. Dostupné z URL: <https://www.researchgate.net/publication/333703698_OSINT_is_the_next_Internet_goldmine_Spain_as_an_unexplored_territoryd>. [cit. 2023-11-03].

[17] mohdshariq. *Sherlock – Hunt Username on Social Media Kali Linux Tool.* Online. In: GeeksforGeeks. 20. 04. 2021 Dostupné z URL: <https://www.geeksforgeeks.org/sherlock-hunt-username-on-social-media-kali-linux-tool/>. [cit. 2023-11-03].

[18] LOPEZ, Martina. *5 free OSINT tools for social media.* Online. In: welivescurit: by ESET. 31. 4. 2023. Dostupné z URL: <https://www.welivesecurity.com/2023/05/31/5-free-osint-tools-social-media/>. [cit. 2023-11-03].

[19] SINGRODIA, V.; MITRA, A. a PAUL, S. *A Review on Web Scrapping and its Applications.* Online. In: 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. 2019. s. 1-6, Dostupné z URL: <https://doi.org/10.1109/ICCCI.2019.8821809>. [cit. 2024-05-10]

[20] baeldung. *Web Crawling vs. Web Scraping.* Online. In: Baeldung. 18. 3. 2024. Dostupné z URL: <https://www.baeldung.com/cs/web-crawling-vs-web-scraping>. [cit. 2024-04-22]

[21] BHADUARIA, Saumya a JUSTIN, Joseph. *Cookie based protocol to defend malicious browser extensions.* Online. In: 2019 International Carnahan Conference on Security Technology (ICCST). In: IEEE Access. 2019. Dostupné z URL: <https://ieeexplore.ieee.org/abstract/document/8888425?casa_token=h4UHeSX8Y8YAAAAA:FjyAohSA00M_Tn-cmu5LuD1pl4w1Pz6V4AZGoX4aBwXhuw56e2DzvyT_ycbGYqAMNrog6D_AAw>. [cit. 2023-10-27]

[22] FIELDS, Abby a SHAFFER, Dan. *Website Crawling 101: The Beginner's Guide to Web Crawlers.* Online, blog. In: SEO: Basics. 2023. Dostupné z URL: <https://www.seo.com/basics/how-search-engines-work/crawling/#:~:text=Website%20Crawling%20101%3A%20The%20Beginner%E2%80%99s,and%20deliver%20relevant%20content>. [cit. 2023-10-28]

[23] BELLO-ORGAZ, G.; JUNG, J. J. a CAMACHO, D. *Social big data: Recent achievements and new challenges.* Online. In: Inf. Fusion. s. 45–59. 2016. Dostupné z URL: <https://doi.org/10.1016/j.inffus.2015.08.005>. [cit. 2024-04-10].

[24] MARMOL, F. G.; PEREZ, M. G. a PEREZ, G. *Reporting offensive content in social networks: Toward a reputation-based assessment approach.* In: IEEE Access. s. 32–40. 2014. <https://doi.org/10.1109/MIC.2013.132>. [cit. 2024-04-10].

[25] KANDIAS, M.; MITROU, L.; STAVROU, V. a GRITZALIS, D. *Which side are you on? A new panopticon vs. privacy.* Online. In: 2013 International Conference on Security and Cryptography (SECRYPT). 2013. Reykjavik, Iceland. s. 1–13 <https://ieeexplore.ieee.org/document/7223159>. [cit. 2024-04-10].

[26] HANSON, James. *The Admiralty Code: A cognitive tool for self-directed learning.* Online. In: ijlter. 2015 . Dostupné z URL: <https://www.ijlter.org/index.php/ijlter/article/view/494/234>. [cit. 2023-10-27].

[27] KINZIE, Kody. *What is OSINT?* Online, blog. In: VARONIS: blog. 2022. Dostupné z URL: <https://www.varonis.com/blog/what-is-osint>. [cit. 2023-11-03].

[28] ELDRIDGE, Alison. *Instagram* Online. In: Britanica. 2023. Dostupné z URL: <https://www.britannica.com/topic/Instagram>. [cit. 2023-11-03].

[29] Lets_0SINT. *Instagram OSINT.* Online. In: Medium. 11. 3. 2023. Dostupné z URL: <https://medium.com/@Lets_0SINT/instagram-osint-eb5f7a5222>. [cit. 2023-11-03].

[30] MOSERRI, Adam. *Shedding More Light on How Instagram Works.* Online, blog. In: Instagram: Blog. 8. 6. 2021. Dostupné z URL: <https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works>. [cit. 2023-11-04].

[31] Meta. *Instagram Graph API.* Online. In: Meta for Deveopers. 2023. Dostupné z URL: <https://developers.facebook.com/docs/instagram-api>. [cit. 2023-11-04].

[32] ROSANENSI, M.; MADANI, M.; WANGGONO, R. T. P.; SETYANTO, A.; SELAMETO, A. A. a et al."*Analysis Sentiment And Tourist Response To Rinjani Mountain Tour Based On Comments From Photo Upload In Instagram.* Online. In: 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE). Yogyakarta, Indonesia. 2018. s. 184-188 Dostupné z URL: <https://doi.org/10.1109/ICITISEE.2018.8720960> [cit. 2023-11-03]

[33] CHEN, LC.; LEE, CM. a CHEN, MY. *xploration of social media for sentiment analysis using deep learning.* Online. In: Soft Comput 24. 2020. Dostupné z URL: <https://doi.org/10.1007/s00500-019-04402-8>. [cit. 2023-11-03].

[34] JAIN, K. a KAUSHAL, S. *A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis* Online. In: 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). Noida, India 2018. s. 483-487. Dostupné z URL: <https://doi.org/10.1109/ICRITO.2018.8748793>. [cit. 2024-03-03].

[35] HE, Q. *Recent Works for Sentiment Analysis using Machine Learning and Lexicon Based Approaches* Online. In: 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). Wuhan, China. 2022. s. 422-426. Dostupné z URL: <https://doi.org/10.1109/AEMCSE55572.2022.00090>. [cit. 2024-03-03].

[36] MEHTA, Pooja a PANDYA, Shamil. *A review on sentiment analysis methodologies practices and applications.* Online. In: International Journal of Scientific and Technology Research. 2020. s. 601-609. Dostupné z URL: <https://doi.org/10.1109/AEMCSE55572.2022.00090>. [cit. 2024-03-03].

[37] GAUTAM, Geetika a YADAV, Divakar. *Sentiment analysis of twitter data using machine learning approaches and semantic analysis.* Online. In: Seventh International Conference on Contemporary Computing (IC3). 2014. Dostupné z URL: <https://doi.org/10.1109/IC3.2014.6897213>. [cit. 2024-03-03].

[38] HEARST, M.A. *Support vector machines.* Online. In: IEEE Access. 1998. s. 18-28. Dostupné z URL: <https://doi.org/10.1109/5254.708428>. [cit. 2024-03-03].

[39] JAIN, K. a KAUSHAL, S. *A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis.* Online. In: 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). Noida, India. 2018. s. 483-487. Dostupné z URL: <https://doi.org/10.1109/ICRITO.2018.8748793>. [cit. 2024-03-03].

[40] KULDEEP, Rana. *Selenium Introduction.* Online. In: ArtOfTesting. 2023. Dostupné z URL: <https://artoftesting.com/selenium-introduction>. [cit. 2023-12-01].

[41] LORIA, Steven a contributors. *TextBlob: Simplified Text Processing.* Online. In: TextBlob: readthedocs Dostupné z URL: <https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>.

[cit. 2024-05-10].

[42] ATHAR, AFAF. *Sentiment Analysis: VADER or TextBlob?.* Online. In: Analytics Vidhya 4. 8. 2022. Dostupné z URL: <https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/>. [cit. 2024-05-10].

[43] PERRY, O. Patrick. Online. In: corpus Dostupné z URL: <http://corpustext.com/reference/sentiment_afinn.html>. [cit. 2024-05-10].

AFINN Sentiment Lexicon

[44] FINN, Å. Nielsen. *afinn* Online. In: Github. Dostupné z URL: <https://github.com/fnielsen/afinn>. [cit. 2024-05-10].

[45] COUNTWORDSFREE. *The list of stop words* Online. Dostupné z URL: <https://countwordsfree.com/stopwords/czech>. [cit. 2024-05-10]

[46] ŘEZNÍČEK, Vilém. *Analýza sentimentu - databáze českých slov s polaritou.* Online, blog. 14. 7. 2018 Dostupné z URL: <https://blog.root.cz/hadoop-kdy-uz-ma-cenu-o-nem-uvazovat-a-kdy-jeste-

n/analyza-sentimentu-databaze-ceskych-slov-s-polaritou/>. [cit. 2024-05-10]

[47] NYANTUDRE, A. Alban Wendyellé. *Hate Speech Detection curated Dataset* Online. 2023 Dostupné z URL: <https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset>. [cit. 2024-05-13]

[48] SALONI, Jhalani. *Instagram Play Store Reviews.* Online. 29. 8. 2023 Dostupné z URL: <https://www.kaggle.com/datasets/saloni1712/instagram-play-store-reviews>. [cit. 2024-05-13]

Seznam symbolů a zkratek

API	Application Programming Interface – Rozhraní pro programování aplikací
CSV	Comma-Separated Values – Hodnoty oddělené čárkami
CSS	Cascading Style Sheets – Kaskádové styly
DOC	Document – Dokument
GDPR	General Data Protection Regulation – Obecné nařízení o ochraně osobních údajů
GEOINT	Geospatial Intelligence – Geoprostorové zpravodajství
GUI	Graphical User Interface – Grafické uživatelské rozhraní
HUMINT	Human Intelligence – Lidské zpravodajství
ID	Identifier – Identifikátor
IP	Internet Protocol – Internetový protokol
MASINT	Measurement and Signature Intelligence – Zpravodajství z měření a podpisů
OSINT	Open Source Intelligence – Zpravodajství z otevřených zdrojů
PDC	Primary Domain Controller – Primární doménový kontroler
PDF	Portable Document Format – Přenosný formát dokumentů
SIGINT	Signals Intelligence – Zpravodajství z elektronických signálů
SOCMINT	Social Media Intelligence – Zpravodajství ze sociálních médií
SVM	Support Vector Machine – Stroj podpůrných vektorů
URL	Uniform Resource Locator – Jednotný lokátor zdrojů
XLS	Excel Spreadsheet – Excelová tabulka
XPATH	XML Path Language – Jazyk pro dotazování v XML

Seznam příloh

A Obsah elektronické přílohy

`ukazka.mp4` ..... videoukázka funkčnosti nástroje