Univerzita Palackého v Olomouci

Filozofická fakulta

Katedra politologie a evropských studií

Bc. Richard Andrýsek

Nenávistné projevy na sociálních sítích zpravodajských serverů

Diplomová práce

Vedoucí diplomové práce:

Mgr. et Mgr. Jakub Lysek, Ph.D.

Olomouc 2020

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně na základě uvedených pramenů a literatury.

V Olomouci dne 20. 6. 2020

……

Richard Andrýsek

Poděkování

Na tomto místě bych rád poděkoval vedoucímu své diplomové práce Mgr. et Mgr. Jakubu Lyskovi, Ph.D. za cenné rady a připomínky, které byly velmi přínosné pro napsání mé diplomové práce.

Obsah

Úvod

Hate speech neboli nenávistné projevy, zažívají s rozvojem internetu, komunikace a sociálních sítí boom. Hate speech už neprodukují pouze členové extremistických hnutí, jak tomu bylo dříve, ale i běžní občané, kteří jsou často frustrováni svou životní situací, mají strach z neznámých věcí a svou frustraci ventilují skrze sdílení nenávistných komentářů a příspěvků. Dokud se hate speech týkala převážně omezené skupiny osob, která se dlouhodobě pohybuje na okraji společnosti, nepředstavovala pro demokratickou společnost příliš velké riziko. S rostoucí popularitou sociálních sítí se ale veřejná debata částečně přesunula do virtuálního prostoru, kde může v reálném čase komentovat události téměř každý. Právě zde se v poslední době nejčastěji setkáváme s hate speech. Na to navazuje i problém, že mnoho lidí není s touto problematikou obeznámeno, a tak se poměrně často setkáváme s lidmi obžalovanými z šíření nenávisti, kteří si nebyli při psaní komentářů vědomi, že jednají protiprávně a mohou svým jednáním ublížit ostatním.

Za hate speech jsou převážně považovány útočné projevy, které mají za cíl šířit, podněcovat nebo propagovat nenávist a násilí vůči lidem, kteří sdílí některou z chráněných charakteristik, kterými jsou například rasa, etnický původ, národní původ, náboženská příslušnost, sexuální orientace, kasta, pohlaví, gender, genderová identita a závažná onemocnění nebo postižení. Hate speech tedy zneužívá základního lidského práva – svobody slova, které je v jednotlivých zemích zaručeno v odlišné míře. (Facebook Community Standards 2020).

V posledním desetiletí se o hate speech mluví převážně v souvislosti s online prostředím, kde se s jejími různými formami setkáváme nejčastěji. S hate speech se na internetu můžeme setkat na webových stránkách, sociálních sítích, internetových diskuzích, v emailové komunikaci nebo například při nákupu určitého typu zboží (Výborný 2011: 14–22). V této diplomové práci se věnuji nejaktuálnější problematice hate speech, což je hate speech na sociálních sítích, kde lidé často v příspěvcích a komentářích sdílí své nenávistné myšlenky a vyhrožují ostatním lidem a skupinám.

Aktuálně se většina výzkumů hate speech na sociálních sítích zaměřuje pouze na sociální síť Twitter (Miro-Llinares, Rodrigez-Salsa 2016; ElSherief et al. 2018; Sharma et al. 2018), která má snazší přístup k datům a ve většině západních zemí je velmi populární. Popularita sociálních sítí v České republice se do jisté míry liší od západních států, například

jinde populární Twitter má v České republice pouze 389 tisíc uživatelů. Velká část obyvatelstva používá Facebook (5,3 milionu) a často také Instagram (2,3 milionu), proto jsem se rozhodl zkoumat pouze nenávistní projevy na sociální síti Facebook. (Focus agency 2019) Jako příklad hate speech na sociální síti Facebook lze uvést případ fotky prvňáčků z Teplic z roku 2017, která vyvolala vlnu nenávisti, a za které je trestně stíháno několik autorů nenávistných komentářů (ČTK 2019).

Hate speech bývá často předmětem diskuzí týkajících se svobody slova a demokracie, a to především po druhé světové válce. Míra, do jaké mohou lidé svobodně vyjadřovat své názory, se liší u jednotlivých demokracií. Například ve Spojených státech je přístup k hate speech velmi benevolentní, neexistuje zde žádná regulace a hate speech je chráněna legislativně prvním dodatkem Ústavy Spojených států amerických, který zaručuje svobodu slova, tisku, vyznání a možnost svobodně se shromažďovat. Naopak například Velká Británie nemá psanou ústavu a kriminalizace hate speech zde sahá až do 17. století. Regulace hate speech ve Velké Británii byla nejprve zavedena proto, aby zabránila útokům na panovníka a vládu, až později kvůli rasovým útokům (Rosenfeld 2001: 4–33). V České republice je svoboda slova garantována čl. 17 v Listině základních práv a svobod a pojem hate speech či nenávistných projevů není zakotven v české legislativě. Nicméně v legislativě lze nalézt trestné činy, do kterých lze zahrnout i hate speech.

V posledním desetiletí vstupují do diskuze o svobodě slova a regulaci hate speech také sociální sítě, nadnárodní firmy poskytující komunikační prostředek pro uživatele z celého světa. Na sociální sítě je vyvíjen tlak ze strany států, aby regulovaly nejen nenávistné projevy, ale v poslední době také nepravdivé zprávy a informace. Sociální sítě k těmto požadavkům přistupují převážně zodpovědně a snaží se odstraňovat nevhodné příspěvky a postihovat jejich autory.

Cílem této diplomové práce je analyzovat komentáře na sociální síti Facebook, a to pod příspěvky týkající se politiky, zpravodajství a zahraničních událostí na stránkách zpravodajských serverů iDnes, Aktuálně, Novinky, ČT24 a Parlamentní listy. Hate speech ve zkoumaných komentářích bude hodnocena na vlastní škále se stupni 1 až 5, která určuje intenzitu nenávistného projevu na základě obsahu zprávy a jejího cílení na jednotlivce, nebo skupinu dle chráněných charakteristik. Jedním z cílů je zjistit, zda se v komentářích na vybraných zpravodajských stránkách pod příspěvky vyskytuje hate speech a pokud ano, tak v jaké míře v porovnání s ostatními studiemi. Druhým cílem je zjistit základní charakteristiky

nenávistných komentářů a jejich autorů. U komentáře je cílem zjistit, zda se výrazně liší výskyt nenávistných komentářů na vybraných stránkách zpravodajských serverů a také zda komentáře obsahující vulgarity obsahují častěji hate speech. Dále je cílem zjistit vliv osobnostních charakteristik autorů jako pohlaví, vzdělání a místo bydliště na míru hate speech v komentářích. Dle dosavadního výzkumu a teorií shrnutých v první části práce a shromážděných dat bylo vytvořeno několik testovacích hypotéz:

Hypotéza H1: Zastoupení hate speech v komentářích pod příspěvky zpravodajských serverů bude v porovnání s jinými výzkumy stejné nebo vyšší.

Hypotéza H2: Komentáře mužů obsahují častěji hate speech než komentáře žen.

Hypotéza H3: Komentáře vzdělanějších lidí budou obsahovat méně hate speech.

Hypotéza H4: Uživatelé z méně rozvinutých regionů budou častěji psát hate speech komentáře.

Hypotéza H5: Pod příspěvky renomovaných zpravodajských serverů budou méně často hate speech komentáře.

Hypotéza H6: Vulgární komentáře budou mnohem častěji také obsahovat hate speech.

Pro analýzu hate speech v komentářích pod příspěvky zpravodajských serverů na sociální síti Facebook jsem vytvářel vlastní datový soubor, což mi umožnilo vybrat si, které proměnné budu měřit a zaznamenávat. Vzhledem ke skandálním událostem okolo analytické firmy Cambridge Analytica a zneužívání dat z Facebooku určených pro vědecké účely k účelům komerčním se firma Facebook rozhodla zpřísnit přístup k datům o jednotlivých uživatelích a stránkách na Facebooku (Ma, Gilbert 2019). Protože by bylo příliš složité a pravděpodobně i nelegální získat data z Facebooku automatizovanou formou, rozhodl jsem se pro ruční sběr dat, který mi umožní detailnější analýzu komentářů a profilů, které tyto komentáře píší. Jednotlivé případy byly vybrány náhodným výběrem pod zpravodajskými příspěvky. V datovém souboru bylo zaznamenáno celkem 14 proměnných u celkem 800 případů, zaznamenanými proměnnými jsou: *text komentáře, pohlaví, zpravodajský server, typ zpravodajské zprávy, emoji*[1]*, délka komentáře, reakce na komentář, bydliště, vzdělání, fotografie, pravděpodobný falešný profil, podpora politické strany, vulgarita komentáře a*

[1] Emoji nebo také emotikony je digitální obrázek nejčastěji obličeje, který se přidává ke zprávám v elektronické komunikaci. Emoji zprávám dodávají určité vyjádření emocí a pocitů, které by člověk vyjádřil neverbálně při osobním kontaktu (Cambridge Dictionary 2020)

intenzita hate speech. Všechny údaje zaznamenané v datovém souboru byly získány na základě veřejných údajů uživatelů a stránek na sociální síti Facebook.

Tato diplomová práce přispívá k dosavadnímu výzkumu především svým zaměřením na ve výzkumech opomíjenou sociální síť Facebook, která je zkoumána mnohem méně než například sociální síť Twitter. Práce se dále detailně věnuje jednotlivým charakteristikám hate speech komentářů a autorům hate speech komentářů a rozšiřuje tak dosavadní výzkum hate speech na sociální síti Facebook v Česku, kde je doposud nejdetailnější studie Projevy nenávisti v online prostoru a na sociálních sítích od Člověk v tísni, o.p.s. (Hrdina, Daňková, Kopecká 2017). Dále práce přispívá novou hodnotící škálou pro hate speech, která hodnotí hate speech přísněji než dosavadní hodnotící škály, a kromě obsahu hate speech komentáře bere při hodnocení v potaz i to, na koho je nenávistný projev namířen (jedinec, nebo skupina).

Na témata hate speech na sociálních sítích a analýzy hate speech existuje množství aktuální odborné literatury, ta se ale zaměřuje především na sociální síť Twitter a možnosti automatizovaného sběru dat a analyzování hate speech za pomoci tradičních metod a metod strojového učení (machine learning). Tomuto tématu se věnují například práce autorů Mondal, Silva, Benevenuto (2017); MacAvaney, Yao, Yang, Russell, Gogarian, Frieder (2019); Gitari, Zuping, Damien, Long (2015). Detailnější analýzu hate speech na sociálních sítích pak provedli Miro-Llinares a Rodrigez-Salsa (2016), kteří zkoumali, v jakou denní hodinu se píše nejvíce hate speech komentářů na Twitteru a zda má zkušenost uživatele na Twitteru dle počtu napsaných tweetů vliv na přítomnost hate speech v příspěvcích.

Dosavadní výzkum zahraničních autorů zkoumá především přítomnost hate speech v příspěvcích na sociální síti Twitter a možnosti jejich detekce. Většina zahraničních autorů došla k tomu, že se hate speech v příspěvcích na Twitteru vyskytuje, ale zjištěné procento hate speech příspěvků se v jednotlivých studiích liší, a to v rozmezí od 0,8 % do 11 %. Zahraniční autoři se až na výjimky nezaměřují na pachatele hate speech a kauzalitu přítomnosti hate speech v příspěvcích. Pouze čeští autoři Hrdina, Daňková, Kopecká (2017) analyzují pachatele hate speech a snaží se vymezit jejich osobnostní charakteristiky a zabývají se hate speech více do hloubky. V této diplomové práci navazuji převážně na trojici českých autorů a zkoumám osobnostní charakteristiky pachatelů hate speech a jejich vliv na přítomnost hate speech v komentářích na facebookových stránkách zpravodajských serverů. Dále diplomová práce rozšiřuje výzkum o vliv proměnných vulgarita komentáře a typ zpravodajského serveru na přítomnost hate speech v komentářích.

Diplomová práce je rozdělena do pěti kapitol. V první kapitole jsou uvedeny definice a teorie týkající se hate speech a příbuzných témat svobody slova a hate crime. Dále se první kapitola zaměřuje na hate speech a definice v prostředí internetu a sociálních sítí Facebook a Twitter a na oběti a pachatele hate speech. Druhá kapitola se zaměřuje na dosavadní výzkum hate speech na sociálních sítích a kauzální mechanismy působící na přítomnost hate speech na sociálních sítích. Třetí kapitola se věnuje metodologii výzkumu hate speech na sociálních sítích, je zde představen způsob výběru zpravodajských serverů, sběr dat, technika měření a charakteristika datového souboru. Ve čtvrté kapitole jsou představeny jednotlivé přístupy ke kategorizaci a analyzování hate speech na základě kterých je vytvořena vlastní škála pro kategorizaci hate speech. Samotná analýza, kde jsou zkoumány vlivy jednotlivých proměnných na přítomnost hate speech v komentářích, je obsažena v páté kapitole. Závěr kapitoly pak sumarizuje výsledky analýzy a nastiňuje možnosti pro další výzkum.

1. Teoretické vymezení problematiky hate speech

V této části se věnuji teoretickému vymezení problematiky hate speech a s ní úzce souvisejícími tématy jako je hate crime a svoboda slova. Dále se v této části věnuji jednotlivým definicím hate speech, důraz je kladen na definice, které používají sociální sítě Facebook a Twitter k určování hate speech na svých platformách. Kapitola se také zabývá specifiky hate speech v online prostředí. Poslední část teoretické kapitoly diplomové práce se věnuje problematice obětí a kriminalizaci, a to především v prostředí internetu a sociálních sítí.

1.1. Hate crime

Nenávistné projevy, anglicky hate speech, patří mezi tzv. hate crimes, které jsou v českém prostředí někdy nazývány také jako trestné činy z nenávisti. Definice pojmu hate crimes je nejednotná a často se liší v závislosti na státu a dané národní zkušenosti. V České republice například definovala Policie České republiky trestné činy z nenávisti jako *„trestné činy, které byly spáchané z důvodu nenávisti pachatele pro některou domnělou nebo skutečnou charakteristiku oběti: příslušnost k určité rase, příslušnost k etnické skupině, národnost, politické přesvědčení, vyznání (případně skutečnost, že je bez vyznání).“* (Policie ČR 2020) Dále pak uvádí, které z trestných činů mohou být zároveň činy nenávistnými.

Dle Sun (2006), který vychází z definice Federální legislativy Spojených států amerických, je hate crime definován jako trestný čin, který je na oběť zaměřen na základě skutečné nebo domnělé příslušnosti k určité rase, barvě pleti, náboženství, zdravotnímu postižení, sexuální orientaci nebo národnostnímu původu (Sun, 2006: 597). Mezi vědci a zákonodárci nedošlo k žádnému konsensu na základě kterého by vznikla všeobecná definice, která by hate crime vymezovala. Důvod této neshody spočívá v tom, že sociální normy, kulturní rozdíly a politické zájmy hrají stěžejní roli při vymezení pojmu hate crime. Ani jednotlivé státy Spojených států amerických se neshodly na jednotné definic hate crime (Boeckmann 2002: 208). Mareš v analýze *Problematika hate crime* pro Ministerstvo vnitra České republiky uvádí aspekty, které se vyskytují ve více definicích a je možné je považovat za důvod k nenávisti ke skupině definované na základě skutečných nebo domnělých faktorů:

- rasy, národnosti a etnicity,
- náboženství,
- sexuální orientace (proti gayům, lesbám bisexuálům a transsexuálům),

- genderové příslušnosti,
- fyzického a mentálního postižení,
- věku,
- třídy,
- politické orientaci (Mareš 2011: 4).

Původní koncept hate crime byl zaměřen pouze na fyzické násilí, s postupem času byl tento koncept rozšířen i na verbální projevy. Proto je někdy rozlišováno mezi hate speech (nenávistné projevy) a hate violence (násilí z nenávisti). Toto je ale pouze jedna z možností, jak rozlišovat uskutečnění hate crime. Miroslav Mareš uvádí následující způsoby uskutečnění hate crime:

- fyzické útoky (rozlišujeme dále od náhodných incidentů až po důkladně plánované útoky, od užití fyzické síly až po užití zbraní),
- poškození majetku,
- šikana,
- obtěžování,
- verbální nadávky,
- útočné graffiti a dopisy (Mareš 2011: 5).

Pachatele hate crime je možné zařadit do skupin na základě motivace, kterou má pachatel k činu. Jiří Herczeg ve své práci rozděluje pachatele na ideologicky motivované, xenofobní, ekonomicky a sociálně motivované a na pachatele kriminální (Herczeg 2007: 148–149). S podrobnou typologií pracuje Marcela Moulisová, která vychází z výzkumu amerických kriminologů McDevitta, Levina a Benneta a rozlišuje pachatele na základě čtyř následujících kategorií:

- pachatelé pro vzrušení (thrill),
- pachatelé kvůli obraně (defensive),
- pachatelé kvůli odvetě (retaliatory),
- pachatelé kvůli poslání (mission) (Moulisová 2008: 206).

Hate crimes má dopad nejen na pachatele a oběti, ale i na celou společnost, proto veřejnost a média věnují vyšší pozornost právě těmto činům. Dopady na společnost mohou být sledovány v několika rovinách:

- **dopad na individuum** – psychologické a emocionální poškození a jejich dopad na identitu a pocit méněcennosti oběti,
- **dopad na napadenou skupinu** – zastrašující efekt na skupinu, k níž oběť náleží, protože přítomnost nenávistného zločinu vzbuzuje v členech skupiny pocit zranitelnosti,
- **dopad na jiné zranitelné skupiny** – činy nenávistného charakteru mají dopad i na minoritní skupiny, odlišné od napadené skupiny, zvláště když je nenávist založena na ideologii, která se vymezuje proti více skupinám ve společnosti,
- **dopad na společnost jako celek** – jedná se o největší problém, který způsobují nenávistné činy, v návaznosti na spáchaný nenávistný čin může docházet k rozdělení lidí ve společnosti (Province of Ontario Ministry of Attorney General 2005).

1.2. Svoboda projevu

Svoboda projevu a možnost vyjádřit svůj názor, postoje a stanoviska, která budou ostatními tolerována tvoří jeden ze základních pilířů demokratického státu. Toto tvrzení dokazuje i judikatura Evropského soudu pro lidská práva, který formuloval svobodu slova jako *„jeden ze základů demokratické společnosti, jednu ze základních podmínek rozvoje společnosti i každého člověka“* (ESLP 1976).

V České republice je svoboda projevu zaručena Listinou základních práv a svobod, přesněji v čl. 17, a je zařazena mezi politická práva. Čl. 17 definuje svobodu projevu jako právo *„svobodně vyjadřovat své názory slovem, písmem, tiskem, obrazem nebo jiným způsobem, jakož i svobodně vyhledávat, přijímat a rozšiřovat ideje a informace bez ohledu na hranice státu“* a nepřipouští možnost jakékoliv cenzury. Listina základních práv a svobod v čl. 17 kromě svobody projevu zaručuje i právo na informace, a to i od státních orgánů a orgánů místní samosprávy, a to přiměřeným způsobem. Svobodu projevu a právo vyhledávat a šířit informace je ale možné omezit zákonem (LZPS 2020).

V oblasti mezinárodního práva je svoboda projevu zakotvena například v Evropské úmluvě o ochraně lidských práv v čl. 10, kde uvádí, že právo na svobodu projevu má každý. Čl. 10 zahrnuje jednak právo na svobodu *„zastávat názory a přijímat a rozšiřovat informace a myšlenky bez zasahování státních orgánů a bez ohledu na hranice“* to se ale nevztahuje na rozhlasové, televizní a filmové společnosti, po kterých může stát vyžadovat jistou formu povolení. Evropská úmluva o ochraně lidských práv ale dodává, že výkon těchto svobod

obsahuje i povinnosti a odpovědnost a může podléhat formalitám, podmínkám a sankcím, které jsou stanoveny zákonem a které jsou nezbytné pro fungování demokratické společnosti a zajištění národní bezpečnosti a fungování demokratického státu (EÚLP 2010). Dále je svoboda projevu zakotvena například v Mezinárodním paktu o občanských a politických právech v čl. 19 anebo také v čl. 19 Všeobecné deklarace lidských práv.

Při konfliktu svobody projevu s jinými základními právy je možné svobodu projevu omezit, to dovoluje například Mezinárodní pakt o občanských a politických právech v čl. 20, kde je zakázána národní, rasová nebo náboženská nenávist, která by mohla vést k podněcování k diskriminaci, nepřátelství anebo násilí (MPOPP 1976). Také Listina základních práv a svobod obsahuje v čl. 17 odstavci 4 možnost omezit svobodu projevu zákonem, jde-li o opatření nezbytná k ochraně práv a svobod druhých, bezpečnost státu, veřejnou bezpečnost, nebo ochranu veřejného zdraví a mravnosti (LZPS 2020).

1.3. Definice hate speech

Při definování pojmu hate speech, česky nenávistné projevy, narážíme na podobný problém jako u definice hate crimes, a tedy nejednotnost definice tohoto pojmu. I přes to, že je pojem často používán ve veřejné, politické, neziskové a akademické sféře, není zde jednotná definice toho, co vše pojem hate speech zahrnuje. Národní i mezinárodní zákony a definice akademiků jsou nejednotné a často vycházejí z místních zkušeností a historicky zakořeněných problémů týkajících se hate speech. V této kapitole se budu věnovat definici hate speech ze dvou perspektiv. Za prvé z pohledu definice hate speech v mezinárodním právu. Za druhé pak představím jednotlivé definice z akademického prostředí, které definují hate speech obsáhleji. Vzhledem k tomu, že se tato práce zabývá analýzou hate speech na sociálních sítích, zařadil jsem do této části i kapitolu, která se věnuje tomu, jak sociální sítě definují hate speech a jak přistupují k příspěvkům obsahujícím hate speech.

Dlouhou dobu nebyl pojem hate speech definován jak v mezinárodním, tak v Evropském právu i přesto, že byl používán v širším významu. První a zatím jediná oficiální mezinárodní definice hate speech vznikla v Radě Evropy v roce 1997 jako doporučení No. R (97) 20 jako *Doporučení výboru ministrů k "nenávistným projevům"* (Council of Europe 2020). Vydané doporučení definuje pojem hate speech jako *„všechny formy projevu, které šíří, podněcují, propagují nebo ospravedlňují rasovou nenávist, xenofobii, antisemitismus nebo jiné formy nenávisti založené na netoleranci, včetně: netolerance vyjádřené agresivním*

nacionalismem a etnocentrismem, diskriminace a nepřátelství vůči menšinám, migrantů a lidem přistěhovaleckého původu." (Council of Europe 1997: 107) Poměrně obsáhlou a aktuální definici hate speech pak přineslo Doporučení No. R 15 *O boji proti nenávistným projevům* vydané v roce 2015 Evropskou komisí proti rasismus a nesnášenlivosti zkráceně ECRI. Doporučení komise definuje hate speech jako výrazy, které používají jednu nebo více z následujících forem projevu: obhajoba, propagace nebo podněcování k pomlouvání, nenávist nebo hanobení jednotlivce či skupiny osob, stejně jako jakékoliv formy obtěžování, urážky, negativní stereotypy, stigmatizace, nebo vyhrožování jednotlivci či skupině. Dále sem patří používání a obhajování jakýchkoliv forem negativních výrazů na základě osobnostní charakteristiky nebo statusu osoby včetně rasy, barvy, jazyka, náboženství nebo víry, věku, postižení, pohlaví, genderu, genderové identity a sexuální orientace (ECRI 2015: 16). Dále je možné za vymezení proti hate speech považovat čl. 20 Mezinárodního paktu o občanských a politických právech, který zakazuje národní, rasovou nebo náboženskou nenávist podněcující k diskriminaci, nepřátelství nebo násilí (MPOPP 1976).

Při definování pojmu hate speech je možné vycházet z mnoha akademických definic, které se liší tím, jak moc obecně na pojem hate speech nahlíží a také v tom, jakou dávají důležitost jednotlivým aspektům. Poměrně jednoduše je hate speech definována v Cambridge Dictionary jako *„veřejný projev, který vyjadřuje nenávist nebo nabádá k násilí vůči člověku nebo skupině založené na něčem, jako je rasa, náboženství, sex nebo sexuální orientace"* (Cambridge Dictionary 2020). Tato definice zahrnuje do hate speech jak člověka jako jednotlivce, tak i skupinu, ale používá pouze malý výčet charakteristik. Více detailní definici hate speech lze najít například v Oxfrod Constitutional Law (2017), kde je hate speech definována jako *„verbální nebo neverbální komunikace, která zahrnuje nepřátelství namířené proti určité sociální skupině, nejčastěji na základě rasy a etnicity (rasismus, xenofobie, anti-semitismus, atd), genderu (sexismus, misogynie), sexuální orientace (homofobie, transfobie), věku a postižení."* Oproti definici Cambridge Dictionary tak obsahuje tato definice navíc i neverbální projevy a vztahuje hate speech pouze na sociální skupinu a už nedefinuje nenávistné projevy proti jedinci.

Autoři Petra Jägera a Pavla Molka (2007: 22) ve své definici rozšiřují osobnostní charakteristiky o politické smýšlení a sociální původ, hate speech tedy definují jako *„projev, jehož cílem je urazit, ponížit, či vyvolat diskriminaci, nenávist nebo násilí proti jednotlivci nebo skupině osob, právě na základě jejich osobních charakteristik, typicky pohlaví, rasy, barvy*

pleti, jazyka, víry a náboženství, politického či jiného smýšlení, národního nebo sociálního původu, příslušnosti k národnostní nebo etnické menšině." S velmi jednoduchou definicí pak přišel Mondal, který definoval hate speech pouze pro prostředí internetu jako *„urážlivý příspěvek, který je zcela nebo zčásti motivovaný zaujatostí autora proti aspektům určité skupiny lidí."* (Mondal et al. 2017) Takto obecně definovaná hate speech umožňuje zahrnout do hate speech velmi široké spektrum příspěvků, které třeba i jen částečně útočí na práva jedince.

Většina definic hate speech se skládá ze tří částí, které definují jednotlivé aspekty, které hate speech obsahuje. Detailnost jednotlivých aspektů hate speech se může v jednotlivých definicích lišit a do jisté míry závisí na zkušenosti autora s hate speech a na problémech a tradicích založených na geografické poloze autora definice. První část definice většinou určuje, zda se do hate speech počítá útok pouze na skupinu, nebo zahrnují do hate speech i útok na jednotlivce, který téměř vždy náleží k více skupinám. Druhou částí je výčet toho, jak může nenávistný výrok útočit na práva jedince či skupiny, je to například nenávist, nabádání k násilí, vyhrožování, urážení, ponižování atd. Třetí část určuje jednotlivé charakteristiky člověka, proti kterým je hate speech namířena, můžou to být například víra, politické názory, rasa, pohlaví a mnoho dalších charakteristik. Jednotlivé definice hate speech se liší v jednotlivých částech, ale většina definic je v jádru velmi podobných a liší se pouze v definování toho, co vše lze považovat za hate speech.

1.4. Jak definují hate speech sociální sítě Facebook a Twitter

Stále více komunikace mezi lidmi probíhá skrze internet a elektronická zařízení, a to v poslední době čím dál více i skrze sociální sítě jako Facebook[2], Instagram[3], Twitter[4] nebo například Snapchat[5]. Sociální sítě umožňují komunikovat miliardám lidí mezi sebou a vyjadřovat se v komentářích nebo příspěvcích. Často tak může na sociálních sítích docházet k veřejným projevům, které obsahují nejen hate speech, ale i násilné fotky a videa, pornografii a sexuální

[2] Sociální síť založená v roce 2004 Markem Zuckerbergem, která má téměř dvě a půl miliardy uživatelů. Facebook častěji používají ženy než muži, nejaktivnější věkové skupiny na Facebooku jsou lidé od 18 do 29 let a také lidé od 30 do 49 let (Chaffey 2020).

[3] Sociální síť založená v roce 2010, která se zaměřuje na sdílení a komunikaci pomocí fotek a videí, kterou používá přibližně miliarda uživatelů. Instagram používají především ženy a mladší lidé od 13 do 29 let (Chaffey 2020).

[4] Sociální síť založená v roce 2006, která umožňuje sdílet a posílat krátké příspěvky mezi uživateli, známé jako tweety. Twitter aktivně používá přibližně 340 milionů uživatelů, častěji používají Twitter muži a mladší věkové skupiny (Chaffey 2020).

[5] Je aplikace spočívající v posílání fotek, které si ostatní uživatelé mohou zobrazit pouze po omezenou dobu, Snapchat byl založen v roce 2011 a aktivně jej používá přibližně 382 milionů uživatelů (Chaffey 2020).

obsah a projevy obsahující nepřiměřenou krutost anebo jiný obsah překračující hranice zákona v daných zemích. Z toho důvodu se sociální sítě brání a vydávají svá pravidla a opatření s kterými každý uživatel sociální sítě musí souhlasit a měl by je respektovat. Pravidla a opatření často obsahují pravidla používání účtu, jeho bezpečnost a zabezpečení a také sekci, kde je popsán všechen nepřípustný obsah, a to jakým způsobem je tento obsah rozpoznáván. V této kapitole se budu věnovat tomu, jak je definována hate speech ve Facebook Community Standards a Twitter Help Center.

Facebook ve svých Community Standards deklaruje, že nepřipouští na Facebooku žádnou formu hate speech, protože skrze ni může docházet k zastrašování, vylučování ze společnosti a v krajních případech i k násilí v reálném světě a aktivně takovéto příspěvky maže a trestá jejich autory. Facebook definuje hate speech ve svých pravidlech jako přímý útok na skupinu lidí na základě toho, čemu Facebook říká chráněné charakteristiky, jsou jimi rasa, etnický původ, národní původ, náboženská příslušnost, sexuální orientace, kasta, pohlaví, gender, genderová identita a závažná onemocnění nebo postižení. Facebook také poskytuje ochranu před hate speech některým případům týkajících se skupiny lidí se statusem imigranta. Pojem "útok" zde Facebook definuje jako násilný nebo dehumanizující projev, prohlášení o méněcennosti nebo výzvu k vyloučení nebo segregaci (Facebook Community Standards 2020).

Facebook ale nepřistupuje ke všem projevům hate speech stejně, snaží se přistupovat k hate speech příspěvkům vždy v kontextu v kterém je příspěvek napsán a také za jakým záměrem je hate speech v příspěvku použito. Za hate speech tak nejsou považovány příspěvky, které sdílejí hate speech někoho jiného za účelem upozornit na jeho nevhodné vyjádření. Dále také není považováno za hate speech, pokud se používá označení, které vylučuje pohlaví za účelem přijetí do zdravotnických nebo podpůrných skupin. Takovou skupinou mohou být například kojící matky, jako skupina určená výhradně ženám. Pokud z příspěvku nebo textu obsahující hate speech není patrný záměr, Facebook takovéto příspěvky automaticky označuje za hate speech a jsou smazány (Facebook Community Standards 2020).

V pravidlech a zásadách sociální sítě Twitter je zakázána jakákoliv forma nenávistného chování, které je definováno jako propagace násilí nebo přímý útok a vyhrožování ostatním lidem na základě jejich rasy, etnicity, národnostního původu, kasty, sexuální orientace, pohlaví, genderové identity, náboženské příslušnosti, věku, zdravotního postižení nebo vážného onemocnění. Na základě těchto kategorií nejsou povoleny účty, jejichž primárním účelem je poškozovat ostatní uživatele na sociální síti. Dále Twitter zakazuje používání nenávistných

obrázků a znaků na profilových obrázcích a profilových hlavičkách. Dále je zakázáno mít v profilovém jménu, nebo v popisu profilu jakékoliv projevování nenávisti vůči osobě nebo skupině osob (Twitter Help Center 2020).

1.5. Hate speech v online prostředí

Online prostředí je jedním z kanálů, které jsou vážně zasaženy problematikou hate speech, nejedná se však o jednu stránku nebo sociální síť, kde se nenávistné projevy vyskytují. Webových stránek, kde se setkáme s hate speech je nespočet, ne všechny stránky jsou ale zaměřeny na šíření nenávisti, některé jen umožňují se svým uživatelům vyjádřit. Uživatelé v tomto případě zneužijí například komentářovou sekci k šíření nenávisti. Štěpán Výborný (2011: 14–22) ve své rigorózní práci rozděluje zdroje hate speech na internetu do sedmi skupin:

- oficiální stránky politických stran a přidružených sdružení,
- mobilizační weby,
- informační propagandistické servery,
- sociální sítě (Facebook, Twitter, YouTube),
- internetové diskuze,
- propaganda využívající e-mailovou komunikaci,
- stránky umožňující prodej a nákup nenávistného zboží.

Výborný (2013: 23) ve své knize rozšiřuje zdroje hate speech o další dvě skupiny, a to o portály zaměřené na reprodukci a sdílení hudby a videí, tuto skupinu vyčlenil zvlášť ze skupiny sociální sítě, kam by se tyto portály jako např. YouTube daly zařadit. Druhou skupinou jsou internetové počítačové hry, které umožňují komunikaci mezi hráči, a může tak docházet k nenávistným projevům, propagaci ideologií a mýtů. Hráči počítačových her utvářejí komunity, které se mnohem častěji vymezují proti určité skupině a mohou smazávat hranici mezi fantazií a realitou (British Institute of Human Rights 2012: 22–23).

Oficiální stránky a prezentace politických stran a k nim přidružených sdružení slouží k prezentaci činnosti, cílů a ideologií dané strany či sdružení. Prezentují zde své politické názory a informují o svém politickém programu a kampani. V České republice do této kategorie primárně spadá krajně pravicový portál Dělnické strany sociální spravedlnosti (dsss.cz), která kromě své primární stránky využívá k propagaci také Dělnické listy (delnickelisty.cz). Také krajně levicové strany v České republice využívají webové prezentace k propagaci svých

ideologií, z politických stran sem patří například Komunistická strana Čech a Moravy, či více radikální Komunistická strana Československa. Z přidružených organizací lze pak zmínit internetovou prezentaci Komunistického svazu mládeže (ksm.cz) a jejich deník Mladá Pravda (Výborný 2011: 15–16).

Mobilizační weby mají za cíl vést propagaci myšlenek a ideologií a zároveň mobilizovat příslušníky dané skupiny k akci, například účastí na demonstraci, občanskou angažovaností anebo bojem proti ideovým nepřátelům. Na mobilizačních webech mohou být také informace o již proběhlých akcích. Do této skupiny lze zařadit webové servery jako Svobodná mládež (revolta114.blogspot.com), Antifašistická akce (antifa.cz), Kolektivně proti kapitálu (protikapitalu.org) anebo dříve fungující Národní odpor[6], AntifaCZ[7] či Třídní válka[8] (Výborný 2011: 16–17).

Informační propagandistické servery na svých webových stránkách informují propagandisticky o aktuálním dění, problémech a věnují se také zahraničním záležitostem a historii. Takovéto webové stránky upravují informace a nahlíží na danou problematiku tak, aby korespondovala s idejemi krajní pravice nebo levice. Zprávy na těchto webových stránkách mohou prezentovat nenávistné postoje a pokusit se překrucovat zprávy tak, aby vedly k posílení radikálních názorů. Z webových stránek, které Výborný (2011: 18) zmiňuje aktuálně fungují stránky Zvědavec (zvedavec.org), Délský potápěč (deliandiver.org), Společnost česko-kubánského přátelství (kubadnes.cz) a Good night white pride (gnwp.cz).

Sociální sítě využívají pravicoví i levicoví extremisté k šíření svých názorů, komunikaci a získávání nových příznivců. V České republice se nejčastěji s extrémistickými skupinami a názory setkáme na sociální síti Facebook a také na sociálních sítích YouTube a Twitter. Členové těchto skupin pravidelně aktualizují své profily, sdílí své příspěvky a zakládají facebookové skupiny, které slouží k šíření jejich myšlenek. Facebookové skupiny šířící nenávist lze rozdělit do tří kategorií:

[6] Národní odpor je neonacistická skupina, která šíří nenávist a popírá některé demokratické principy. Nejznámější členem skupiny je Filip Vávra, zakladatel hnutí (Viktora 2010).

[7] Antifa.cz neboli Antifačistická akce (AFA) se zabývá propagačním a informačním bojem proti všem autoritářským ideologiím a formám útlaku (Antifa.cz 2020).

[8] Třídní válka je levicové komunistické revue, působící na internetu od roku 2009, které začalo zřejmě pod vlivem levicové revolty v Řecku (Bastl et al. 2011: 52).

- **skupina pro sympatizanty** – ve skupině jsou sympatizanti myšlenek a tvoří ji aktivní členové hnutí, kteří zde probírají své názory a koordinují akce.
- **prezentace sdružení** – jedná se o oficiální facebookové stránky sdružení, které slouží k prezentaci sdružení a komunikaci s vlastními členy i ostatními uživateli na sociálních sítích,
- **skupiny širokého ohlasu** – nejsou spjaty s žádnou konkrétní skupinou, ale slouží k šíření nenávistných projevů, často bývají blokovány správci sociálních sítí (Výborný 2011: 18–20).

Internetové diskuze v poslední době obsahují stále více hate speech, která jednak úzce souvisí například s romskou problematikou nebo s uprchlickou krizí a migranty. Diskuze lze rozdělit do dvou kategorií, a to na diskuze, které jsou navštěvovány především příslušníky extrémních uskupení a na diskuze, které navštěvuje běžná populace například v komentářové sekci zpravodajských serverů. S rostoucí popularitou sociálních sítí se internetové diskuze přesunuly z fór a komentářových sekcí právě na sociální sítě. Často se tak v diskuzích na sociálních sítích můžeme setkat s nesnášenlivými výroky, a to i přes to, že většina velkých stránek na Facebooku má své správce, kteří mažou nejradikálnější příspěvky (Výborný 2011: 20–21).

Propaganda využívající e-mailovou komunikaci, zneužívá e-mailovou komunikaci k šíření nenávistných zpráv a hoaxů[9]. Na rozesílání hoaxů se podílí samotní uživatelé e-mailů, kteří uvěří nepravdivým nebo zavádějícím zprávám, které v nich vyvolají potřebu informovat i ostatní známé. Hlavní problém hoaxů spočívá v uživatelích elektronické pošty, kteří si neověřují informace v e-mailech a dále přeposílají zprávy často s poplašným nebo nenávistným obsahem. Poslední skupinou, kde se může vyskytovat hate speech na internetu, je **prodej a nákup zboží** do kterého spadají internetové obchody, které prodávají oblečení a další předměty s nenávistnou tématikou. Jejich cílem je za pomocí předmětů rozšířit své ideje, a především díky prodeji získat finanční zdroje pro svou další činnost (Výborný 2011: 21–22).

V diplomové práci se zabývám především skupinou nenávistné projevy na sociálních sítích, a to pouze na síti Facebook. Další zkoumanou skupinou z výše zmíněných je skupina

[9] Hoax je nejčastěji poplašná zpráva, která varuje před neexistujícím nebezpečím, nejčastěji se tento pojem používá ve spojitosti s internetovou komunikací (Hoax.cz 2020).

internetové diskuze, u kterých se zaměřuji pouze na vybrané diskuze pod příspěvky zpravodajských serveru na sociální síti Facebook.

1.6. Oběti hate speech

Oběti hate speech jsou příslušníci určité skupiny na základě společných charakteristik. Cílem pachatele je oběť právě proto, že vykazuje jisté charakteristiky, jakými jsou například rasa, náboženské vyznání, nebo politická příslušnost, které není možné ovlivnit. Právě proto je pro oběť těžké jakýmkoliv způsobem předcházet hate speech, některé charakteristiky jako sexuální orientace nebo náboženské vyznání lze však skrývat. Stejně jako u dopadů hate crime na oběti v kapitole 1.1. tak i u hate speech může být dopad nejen na danou oběť jako jedince, ale také na skupinu, ke které náleží a potažmo i další skupiny, které se mohou cítit ohroženy a v krajních případech může mít hate speech dopad i na celou společnost.

V České republice se nejčastěji obětí hate speech na internetu dle projektu HateFree Culture stávají Romové, a to i častěji než v posledních letech aktuální téma okolo uprchlíků (iDnes.cz 2017). Další skupinou, která se stává obětí hate speech jsou muslimové. Vyšší intenzitu nenávistných projevů proti této skupině je možné sledovat v poslední době, a to především po útocích na francouzský satirický časopis Charlie Hebdo a dalších teroristických útocích (Zavoral 2015). Od roku 2015, kdy vrcholila uprchlická krize do Evropy, se čím dál častěji stávají obětí nenávistných projevů i migranti. Na sociální sítí Facebook dle výzkumu nevládní organizace Člověka v tísni vykazovalo až 80 % všech zkoumaných příspěvků o uprchlících a muslimech negativní postoj (Fendrych 2016). Příkladem závažných nenávistných projevů proti romským, arabským a vietnamským občanům může být i případ fotky dětí z teplické školy, pod kterým se objevila řada nenávistných komentářů, za které některým autorům hrozily 2 roky odnětí svobody s podmíněným odkladem. Komentáře vyzývaly k použití granátu na děti, nebo ke zplynování dětí (České noviny 2019).

1.7. Kriminalizace hate speech v České republice

Kriminalizace hate speech je už jen z pohledu omezování svobody jednotlivců složitá, pokud ji zákonodárci omezí, setkají se často s nesouhlasem veřejnosti, které kriminalizací takovýchto projevů omezují svobodu slova. Míra kriminalizace hate speech se liší v závislosti na státu. Spojené státy americké například hate speech nepostihují vůbec, naopak státy v Evropě včetně České republiky kriminalizují hate speech poměrně často, a to i na sociálních sítích.

V rámci českého právního řádu nenajdeme přímou definice hate speech (nenávistného projevu). V trestním právu nejsou tresty za hate speech přímo vymezeny, ale je možné je nalézt v některých trestných činech, které jsou uvedeny v trestním zákoníku převážně v části Trestné činy narušující soužití lidí. Dle § 352 je to násilí proti skupině obyvatel a proti jednotlivci – v odstavci 1 je zahrnuto i pouhé vyhrožování usmrcením, ublížením na zdraví nebo způsobením škody velkého rozsahu. Odstavec 2 pak dále vymezuje, že pokud je pohnutka pachatele namířena i jen domněle proti rase, příslušnosti k etnické skupině, národnosti, politickému přesvědčení nebo vyznání je trestní sazba vyšší. Dále je vyhrožování v § 353 – nebezpečné vyhrožování, kde v odstavci 1 je postihováno vyhrožování usmrcením, těžkou újmou na zdraví nebo jinou těžkou újmou takovým způsobem, že to může u oběti vzbudit důvodnou obavu. Pohoršující v tomto případě je, pokud pachatel spáchá čin v odstavci jedna jako člen organizované skupiny (Trestní zákoník 2009).

Dále lze najít hate speech v trestných činech v § 355 – hanobení národa, rasy, etnické nebo jiné skupiny osob. Trestný čin je v odstavci 1 vymezen jako veřejné hanobení některého národa, jeho jazyka, některé rasy nebo etnické skupiny nebo skupiny osob pro jejich skutečnou nebo domnělou rasu, příslušnost k etnické skupině, národnost, politické přesvědčení, vyznání nebo proto, že jsou skutečně nebo domněle bez vyznání. Vyšší sazba je, pokud pachatel spáchá trestný čin minimálně se dvěma osobami nebo jej spáchá prostřednictvím tisku, filmu, rozhlasu, televize nebo veřejně přístupné počítačové sítě. Dále sem lze zařadit i § 356 – podněcování k nenávisti vůči skupině osob nebo k omezení jejich práv a svobod. Takovýto trestný čin obsahuje veřejné podněcování k nenávisti vůči některému národu, rase, etnické skupině, náboženství, třídě nebo jiné skupině osob nebo k omezování práv a svobod jejich příslušníků. Stejně jako v § 355 je vyšší sazba určena, pokud je trestný čin spáchán hromadnými sdělovacími prostředky (Trestní zákoník 2009).

Hate speech lze najít také v trestním zákoníku v části *Trestné činy proti lidskosti,* a to především v § 402 – apartheid a diskriminace skupiny lidí a § 403 – založení, podpora a propagace hnutí směřujícího k potlačení práv a svobod člověka (Trestní zákoník 2009). Osobně si myslím, že kriminalizace hate speech v České republice je poměrně přísná. Dle zákonného nastavení mohou být postihovány i osoby, jejichž projevy jsou svým obsahem pouze nekorektní, ale neobsahují přímou nenávist vůči osobám nebo skupině osob na základě chráněných charakteristik. Dle výzkumu veřejného ochránce práv 2020 na téma *Nenávistné projevy na internetu a rozhodování českých soudů* bylo z celkového počtu 47 případů 49 % se

skutkovou podstatou **podněcování k nenávisti vůči skupině osob nebo k omezování jejich práv a svobod** (§ 356 TZ). Zhruba pětina případů obsahovala **hanobení národa, rasy, etnické nebo jiné skupiny osob** (§ 355 TZ) a přes 20 % případů mělo jako skutkovou podstatu **násilí proti skupině obyvatel a proti jednotlivci** (§ 352 TZ). Celkem 91 % všech sledovaných případů nenávistných projevů na internetu skončilo tak, že byl pachatel odsouzen a pouze ve 2 % případů byl pachatel zproštěn obžaloby. Nejčastějším trestem bylo podmíněné odnětí svobody s určením zkušební doby (63 % případů), dále byl často udělen peněžitý trest s určením náhradního trestu v rozmezí od 5 000 Kč až do 30 000 Kč (Výzkum veřejného ochránce práv 2020)

2. Hate speech na sociálních sítích

Sociální sítě, jakými jsou Facebook, Instagram nebo Twitter používá v dnešní době značná část české populace. Nejčastěji tráví lidé svůj čas na sociálních sítích právě na Facebooku, kde je celkem 5,3 milionu uživatelů. Celkově sociální sítě v České republice používá 52,6 % žen a 49,4 % mužů (Michl 2019). Do prostředí online sociálních sítí se v posledních letech přesunula značná část nenávistných projevů, dříve bylo pro tyto projevy charakteristické například hajlování v hospodě, či vytetovaný hákový kříž na ruce. V dnešní době je pro nenávistné projevy charakteristická masovost a částečná anonymita sociálních sítí, za kterou se schovává řada útočníků (Nahodil 2019).

Zásadní nárůst hate speech na sociálních sítích v České republice nastal v roce 2015, kdy v Evropě vrcholila uprchlická krize. Hate speech se v této době zaměřovala proti potenciálním migrantům překračujícím hranice EU s kterými úzce souvisel i boj proti islamizaci naší společnosti. Projevy proti migrantům jsou v dnešní době jen malou částí hate speech na sociálních sítích, daleko častěji jsou přítomny hate speech komentáře vůči institucím Evropské unie. V České společnosti dlouhodobě panuje proti-romská nálada, která se projevuje i v článcích a komentářích na sociálních sítích (Cakl 2019).

Následující kapitola je rozdělena do dvou částí. První část se věnuje dosavadnímu výzkumu četnosti hate speech na sociálních sítích, a to především na sociální síti Twitter. Druhá část se věnuje dosavadnímu výzkumu v oblasti charakteristiky pachatelů hate speech na sociálních sítích na základě dostupných informací o jednotlivých uživatelích a příspěvcích.

2.1. Přítomnost hate speech na sociálních sítích

O tom, že se na sociálních sítích můžeme setkat s agresivními projevy a hate speech projevy není pochyb, ale o jak velký fenomén se skutečně jedná? V České republice proběhl pouze omezený počet studií věnujících se hate speech na sociálních sítích, proto se kapitola zaměřuje převážně na zahraniční analýzy hate speech, které se kvůli snadnějšímu přístupu k datům věnují většinou pouze sociální síti Twitter.

V prostředí Českého Facebooku provedli Hrdina, Daňková, Kopecká (2016) analýzu hate speech v období června a července 2015, když vrcholila uprchlická krize. Zaměřili se pouze na komentáře obsahující klíčová témata, jakými jsou například islám nebo migranti. Celkem autoři shromáždili na téma *islám* přibližně 20 000 komentářů od přibližně 10 000 autorů. Z tohoto počtu autorů bylo pouze 230 autorů velmi aktivních a napsali za sledované období více než 10 komentářů obsahující téma islám. Z 230 velmi aktivních autorů jich 184

bylo označeno za extremistické a islamofobní. Dále bylo zjištěno, že přibližně 80 % komentářů obsahující islám bylo v tomto směru negativních a útočných a pouze 20 % komentářů se vyjadřovalo o islámu neutrálně, pozitivně anebo vyjádření nebylo relevantní. U tématu migranti bylo ve sledovaném období červenec až srpen publikováno 60 000 komentářů z kterých byl vybrán vzorek 377 autorů. Z toho vzorku 62 % autorů vyjadřovalo v komentáři negativní až nenávistný projev vůči migrantům, 31 % autorů vyjadřovalo v komentáři neutrální postoj a pouze 5,5 % autorů bylo vůči migrantům pozitivních. Z této studie je patrné, že přítomnost hate speech na českých sociálních sítích je značná, je ale nutné vzít v potaz, že byla vybrána témata, která česká společnost vnímá velmi kriticky a je zde větší pravděpodobnost, že budou komentáře obsahovat nenávistné projevy.

V zahraničí se přítomnosti hate speech na sociálních sítích věnuje podstatně vyšší pozornost a byla již zveřejněna řada odborných článků a publikací. Vzhledem k vyšší popularitě sociální sítě Twitter a snazší dostupnosti dat prostřednictvím Twitter API[10] se většina studií zaměřuje právě na sociální síť Twitter a u nás populárnější síť Facebook se v zahraničních studiích až na výjimky nezkoumá.

Dánští autoři Waseem a Hovy (2016) se ve své studii zaměřují na možnosti predikce hate speech na sociální síti Twitter a zaměřují se na predikci za pomoci demografické, geografické a lexikální distribuce. Výzkum je postaven na datech z celkem 136 052 tweetů sesbíraných během dvou měsíců. Z těchto tweetů bylo celkem anotováno 16 914 tweetů od 1 236 uživatelů, které v 31,7 % případů (5 355 tweetů) obsahovaly hate speech z čehož 3 383 tweetů bylo označeno jako sexisticky ofenzivní a 1 972 tweetů za rasistické. Tweetů, které byly neutrální a neobsahovaly známky hate speech bylo 68,3 %.

Další výzkum příspěvků na Twitteru provedli Miró-Llinares a Rodriguez-sala (2016), kteří ve své studii analyzovali soubor Twitterových příspěvků týkající se útoku na satirický časopis Charlie Hebdo. Ve studii analyzovali celkem 282 397 tweetů, které byly automaticky klasifikovaný na základě manuálně kódovaného korpusu dat. Jako hate speech bylo klasifikováno pouze 2 304 příspěvků, což tvoří přibližně 0,8 % ze všech analyzovaných tweetů.

Vědci Burnap a Williams (2015) ve své studii zkoumali na 450 000 tweetů související s událostí vraždy vojáka Britské Armády Lee Rigbyho v roce 2013 v Londýnské části Woolwich. Pro anotaci tweetů využili machine learning metody, kde jako vstupní data pro učení využili manuálně anotovaný korpus tweetů. V náhodně vybraném manuálně anotovaném

[10] Twitter API je platforma, skrze kterou mohou uživatelé získávat rozsáhlá data o tweetech. Umožňuje sledovat data o vlastních tweetech, nebo vyhledávat aktivní témata a trendy na celém Twitteru (Developer Twitter 2020).

korpusu dat bylo zaznamenáno přibližně 11 % příspěvků obsahující hate speech. Při použití machine learning metod se přesnost anotování hate speech zlepšila o 7 %, což znamená, že přibližně 3 000 tweetů obsahující nenávistné a antagonistické názory bylo manuální anotací určeno chybně.

Davidson et al. (2017) se zaměřují na automatickou detekci ofenzivních a hate speech příspěvků na sociální síti Twitter. Z celkového počtu 85,4 milionů tweetů náhodně manuálně nakódovali 24 802 tweetů, přičemž každý tweet byl kódovaný alespoň třemi pracovníky společnosti CrowdFlower. Manuálně kódovaný datový soubor obsahoval 5 % ofenzivních nebo hate speech tweetů na kterém se shodla většina hodnotitelů a pouze 1,3 % tweetů na kterých se hodnotitelé shodli jednomyslně.

Z prezentovaných studií zabývajících se přítomností hate speech na sociálních sítích je patrné, že se přítomnost hate speech značně liší mezi jednotlivými studiemi. Většina studií vychází ze svého unikátního datového souboru, který byl sesbírán na základě určitého klíče (události), jenž je pro každou studii unikátní. Také použitá metodika pro rozpoznávání hate speech se u jednotlivých studií liší. Je proto velmi obtížné míru hate speech porovnávat mezi jednotlivými studiemi. Ve vybraných studiích bylo procento hate speech příspěvků v rozmezí od 0,8 % u Twitter příspěvků s tématem útoku na Charlie Hebdo až po 80 % nenávistných příspěvků na téma islám na facebookových stránkách českých zpravodajských serverů v období uprchlické krize v roce 2015.

2.2. Pachatelé hate speech na internetu a sociálních sítích

Proč lidé píší nenávistné projevy a kdo jsou lidé na sociálních sítích a internetu, kteří nenávistné projevy píší? Motivy pro psaní hate speech se u jednotlivých skupin pachatelů liší, motivaci pachatelů můžeme podobně jako u hate crimes v kapitole 1.1. zařadit do kategorií – vzrušení, obrana, odveta a poslání. Nenávistné projevy byly dříve spojovány převážně s extrémně pravicovými subjekty, nyní jsou ale stále častěji pachateli i běžní občané (Český helsinský výbor 2012: 13). Nenávistné projevy běžných občanů na internetu a sociálních sítí analyzovali Hrdina, Daňková, Kopecká (2016) v rámci projektu neziskové organizace Člověka v tísni, zaměřili se na sociální síť Facebook, kde analyzovali nenávistné diskuze obsahující hate speech hesla jako islám, migranti, podpora a Afrika. Sběr dat byl realizován od června do srpna v roce 2015, v období, kdy v Evropské unii vrcholila uprchlická krize.

Z analýzy komentářů vyplynulo, že mnohem častěji jsou autory nenávistných komentářů na téma migrace a islámu na sociální síti Facebook muži a to v 67 % případů a pouze v 33 % případů byly autorem komentářů ženy. Průměrný věk nejaktivnějších producentů hate speech v komentářích na sociálních sítí je mezi 35 až 50 lety. Zanedbatelnou skupinu autorů hate speech tvořily lidi mladší 25 let a 60 let a starší, kteří byli autory nenávistných komentářů jen ve výjimečných případech. Vliv vzdělání se nepodařilo prokázat, protože relevantní údaje byly pouze u 29 osob, kde převažovalo středoškolské vzdělání a několik lidí s vysokoškolským vzděláním převážně technického a ekonomického charakteru. U autorů hate speech nelze určit ani regionální zakotvení, nejvíce autorů pocházelo z Prahy (22 sob) a Brna (10 osob) další města byly zastoupeny náhodně s výskytem po celé ČR. Součástí výzkumu z roku 2017 byla i analýza komentářů pod zprávami na zpravodajských serverech. Zde až 70 % příspěvků obsahovalo hate speech. Nejčastěji se hate speech objevila na portálu iDnes a jeho subdoménách, dále následovaly zpravodajské servery Novinky a Parlamentní listy. Podíl nenávistných komentářů zde většinou přesahoval 70 % z celkového počtu komentářů v diskuzi (Hrdina, Daňková, Kopecká 2016: 11–15).

Dle Hrdiny, Daňkové a Kopecké (2016) lze definovat demografický profil „řadového" producenta hate speech na téma migrace a islám následovně. *„Více než dvě třetiny producentů hate speech tvoří muži. Producenti hate speech se rekrutují z věkové kategorie cca 35–50 let, mladší lidé jsou mezi nimi zastoupeni spíše výjimečně. Nelze u nich určit specifické regionální zakotvení. Mnozí z nich mají středoškolské či vysokoškolské vzdělání technického či ekonomického zaměření."*

Vliv pohlaví na hate speech analyzovali také autoři Waseem a Hovy (2016), ti se zaměřili na hate speech příspěvky na sociální sítí Twitter. Pohlaví u uživatelů určovali dle jména a fotky uživatele, protože Twitter nevyžaduje při registraci účtu zadání pohlaví a dalších demografických údajů. Z celkového počtu analyzovaných hate speech příspěvků bylo pouze u 52,36 % uživatelů zjištěno pohlaví. Z celkového vzorku příspěvků byli nejčastěji autory hate speech příspěvků muži a to v 50,08 % případů, 47,64 % případů hate speech nemělo identifikované pohlaví a pouze u 2,26 % příspěvků byly autorkami ženy. I přes vysoký počet případů bez určení pohlaví je zde patrné, že muži jsou mnohem častěji pachateli hate speech než ženy, často jsou také pachateli hate speech osoby, které vystupují pod smyšlenou identitou a nelze tak určit jejich pohlaví.

Odlišný přístup k analyzování pachatelů hate speech na internetu a sociálních sítích zvolili autoři Miro-Llinares a Rodrigez-Salsa (2016), kteří se zaměřili na sociální síť Twitter. V rámci své analýzy zkoumali jak zkušený a sledovaný je uživatel, který píše hate speech komentáře. Nejčastěji psali hate speech středně zkušení uživatelé, kteří měli mezi 100 - 10 000 sledujících, naopak malý podíl hate speech v příspěvcích měli uživatelé s více než 10 000 sledujícími. Dále autoři analyzovali, v jakou denní dobu nejčastěji uživatelé píší hate speech, zde vyšlo, že 30 % hate speech komentářů bylo napsáno mezi 16:00 - 20:00 a 27,35 % mezi 13:00 - 16:00. Naopak nejméně komentářů bylo napsáno v noci a to pouze 3,52 % a v ranních hodinách (8:00 - 13:00) pouze 18,87 %.

V analytické části diplomové práce navazuji na některé dosavadní výzkumy jako je míra hate speech na sociálních sítích a vliv pohlaví, místa bydliště a vzdělání na přítomnost hate speech a rozšiřuji dosavadní výzkum o proměnné typ zpravodajského serveru a vulgárnost komentáře a zkoumám jejich vliv na přítomnost hate speech v komentářích na vybraném vzorku komentářů pod příspěvky zpravodajských serverů na sociální síti Facebook.

U proměnné typ zpravodajského serveru lze očekávat vyšší míru hate speech u méně renomovaných médií, kde je často neochota editorů a správců moderovat diskusi a dohlížet na její obsah (Cakl 2019). Naopak dle Netrvalové (2017) renomované zpravodajské servery jako Aktuálně a iDnes přistupují k čtenářským diskuzím na svých webech zodpovědně a pečlivě sledují a moderují komentářovou sekci. Lze tedy předpokládat, že tyto renomované zpravodajské servery budou mít obdobný přístup i na facebookových stránkách. To částečně potvrzuje i studie Hrdiny, Daňkové a Kopecké (2016) kde se nejčastěji vyskytovala hate speech na portálu iDnes, Novinky a Parlamentní listy. U portálu iDnes bychom měli dle Netrvalové očekávat spíše menší přítomnost hate speech a více moderovanou diskuzi, naopak vyšší přítomnost hate speech na serverech Novinky a Parlamentní listy podporuje tvrzení, že méně renomované a moderované diskuze na stránkách zpravodajských serverů budou obsahovat více hate speech. Vulgarismy by měly mít na přítomnost hate speech pozitivní vliv, vzhledem k tomu, že velká část hate speech zároveň obsahuje jistou formu vulgarismů. Tento předpoklad potvrzuje i studie Holgate et al. (2018), ve které vylepšují machine learning metodu pro detekci hate speech o kategorii vulgarismů, díky které se povede správně detekovat o 6,1 % více hate speech komentářů než bez využití kategorie vulgarismů.

3. Metodologie výzkumu

V této kapitole je představen výběr jednotlivých zkoumaných případů zpravodajských serverů, vymezení sledovaných proměnných, technika sběru dat a výběr vhodných kvantitativních metod pro výzkum. Dále se kapitola věnuje základní charakteristice datového souboru facebookových komentářů, lingvistické analýze hate speech a rozpoznávání trollích a falešných profilů na sociálních sítích.

3.1. Výběr zpravodajských serverů

Pro analýzu hate speech v komentářích na Facebooku jsem zvolil zpravodajské servery, které denně informují o tuzemských i zahraničních událostech a komentují tak aktuální dění. Jednotlivé zpravodajské servery byly vybrány tak, aby reprezentovaly co největší okruh čtenářů, názorů a vlastníků. Při výběru zpravodajských serverů byl brán ohled na měsíční návštěvnost zpravodajského webu, počet fanoušků na sociální síti Facebook, vlastníky zpravodajských serverů a rating dle Nadačního fondu nezávislé žurnalistiky. Pro lepší reprezentativnost vzorku zpravodajských serverů jsem se rozhodl vybrat i zástupce médií, které na svých webových stránkách a sociálních sítích zveřejňují falešné zprávy[11] a také zástupce veřejnoprávních médií.

Pro hodnocení transparentnosti a profesionality jednotlivých médií jsem se rozhodl použít rating Nadačního fondu nezávislé žurnalistiky, který na škále A až C hodnotí komerční média. Rating využívá metodu kvalitativního výzkumu s ručně kódovaným obsahem zpravodajských serverů na základě kritérií jako snadno dostupné informace o redakci a organizaci, jasně označené výchozí zdroje článku, důležitost zpráv není uměle zvyšována, články nezneužívají stereotypy, reklamy a komentáře jsou jasně odděleny od zpravodajského sdělení a zpravodajský server se vyhýbá zavádějícím titulkům. (NFNZ, 2020)

Pro analýzu hate speech v komentářích na Facebooku jsem zvolil pětici zpravodajských serverů iDnes, Novinky.cz, Aktuálně, ČT24 a Parlamentní listy. V následujících odstavcích vysvětlím, proč jsem jednotlivé zpravodajské servery vybral, dále jsou v tabulce 3.1 uvedeny nejdůležitější charakteristiky vybraných zpravodajských portálů.

Dle webové služby SimilarWeb.com, která dlouhodobě sleduje návštěvy největších webových stránek je nejnavštěvovanějším zpravodajským sever v České republice iDnes, který

[11] Falešné zprávy (Fake news) jsou příběhy, které se tváří jako zpravodajské sdělení, které se často virálně šíří internet nebo je sdíleno alternativními médii. Falešné zprávy se nejčastěji používají k ovlivnění veřejného mínění nebo jako vtip (Cambridge Dictionary 2020).

jen za prosinec 2019 měl 84,25 milionů návštěv. Zpravodajský server iDnes spravuje společnost MAFRA, a.s. patřící firmě Agrofert a.s., která je nyní ve svěřenském fondu AB private trust I a II (Kurzy.cz, 2020). Na Facebooku má stránka iDnes 219 tisíc sledujících, kteří tvoří dle služby SimilarWeb.com 4,73 % ze všech návštěvníků iDnes. Dle ratingu NFNZ je největším nedostatkem horší dostupnost informací o redakci, a ne vždy jasně označené výchozí zdroje článku, proto byl server iDnes zařazen do kategorie B+. Server iDnes jsem pro analýzu vybral především kvůli vysoké návštěvnosti webových stránek, vysokému počtu fanoušků na Facebooku a také proto, že je součástí mediální společnosti MAFRA, a.s.

Tabulka 3.1: Přehled zkoumaných zpravodajských serverů

Zpravodajský server	Měsíční návštěvnost webu	Návštěvnost z Facebooku	Počet Facebook fanoušků	Rating dle NFNZ	Majoritní vlastníci
Aktuálně	24.78 milionů	10.64 %	113 134	A	Economia, a.s.
ČT24	5.39 milionů	20.03 %	394 706	-	Veřejnoprávní
iDnes	84.25 milionů	4.73 %	219 663	B+	MAFRA, a.s.
Novinky.cz	76.23 milionů	3.99 %	145 021	A-	Silky s.r.o. a Seznam.cz
Parlamentní listy	6.9 milionů	7.59 %	48 723	C	OUR MEDIA a.s.

Zdroj: NFNZ, Facebook a SimilarWeb.com, prosinec 2019

Druhým nejnavštěvovanějším zpravodajským serverem jsou Novinky.cz, které mají měsíčně na svém webu mezi 73 až 88 miliony návštěv. Novinky.cz vlastní z 58,4 % firma Silky s.r.o. a z 33,6 % Seznam.cz Iva Lukačoviče (Kurzy.cz, 2020). Facebookové stránky Novinky.cz mají 145 tisíc a dle služby SimiliarWeb.com pouze 3,99 % celkové návštěvnosti webu Novinky.cz je skrze Facebook. Novinky.cz byly zařazeny do kategorie A-, a to především kvůli hůře dostupným informacím o redakci a nedostatečnému odkazování na zdroje přes hypertextový odkaz. Server Novinky.cz jsem vybral kvůli vysoké návštěvnosti na webu a také

kvůli silnému propojení s vyhledávačem Seznam.cz, jehož měsíční návštěvnost je dle SimilarWeb.com 259 milionů.

Aktuálně je český on-line deník, který vydává společnost Economia, a.s. vlastněná podnikatelem Zdeňkem Bakalou do které patří i další média jako Hospodářské noviny nebo portál Centrum.cz (Kurzy.cz, 2020). Web Aktuálně je šestým nejnavštěvovanějším zpravodajským webem v České republice s 24,78 miliony přístupy za prosinec 2019. Facebookové stránky Aktuálně mají 113 tisíc fanoušků, traffic z Facebooku na webové stránky Aktuálně.cz tvoří 10,64 % z celkové návštěvnosti. Z vybraných zpravodajských serverů se Aktuálně v ratingu NFNZ umístilo nejlépe a získalo nejvyšší možné hodnocení A. Nižší stupeň hodnocení server získal pouze v kritériu práce s online odkazy, ověření zpráv z více zdrojů a nejednoznačné odlišení reklamního sdělení. Server Aktuálně jsem do analýzy vybral jako zástupce mediální společnosti Economia, a.s. a zároveň jako představitele nejlépe hodnoceného média v oblasti transparentnosti a profesionality.

Při výběru veřejnoprávního média pro analýzu hate speech v komentářích na sociální síti Facebook jsem měl na výběr ze tří veřejnoprávních médií v České republice, a to Český rozhlas, Česká televize a Česká tisková kancelář. Při výběru mezi těmito médii rozhodovala především aktivita na Facebooku a počet fanoušků, zde jasně dominovala zpravodajská služba ČT24, která denně sdílí desítky zpravodajských článků a reportáží a sleduje ji 394 tisíc lidí. Oproti tomu Facebookový profil ČTK sleduje pouze 13 tisíc lidí a profil Český rozhlas 79 tisíc lidí. Webové stránky ct24.ceskatelevize.cz mají za prosinec 2019 celkem 5,39 milionů přístupů, což je poměrně málo v porovnání s návštěvností sledovaných komerčních zpravodajských serverů. Z celkového počtu 5,39 milionu přístupů na web ČT24 bylo 20,03 % provedeno skrze Facebook.

Parlamentní listy jsou kontroverzní zpravodajský server, který v minulosti vydal dezinformační článek o migrantech překračující české hranice a podle analýzy z Masarykovy univerzity používá manipulativní techniky ve svých článcích (ČT24 2020). Dle služby SimilarWeb.com je návštěvnost webu Parlamentní listy mezi 6,6 až 7,9 miliony, z toho přibližně 7,59 % stránky navštíví skrze Facebook. Facebookovou stránku Parlamentních listů sleduje 48 tisíc uživatelů. Parlamentní listy vydává společnost OUR MEIDA a.s., kterou vlastní Jan Holoubek, Michal Voráček, Jiří Čermák a společnost WCV World Capital Ventures Cyprus Limited sídlící na Kypru, kterou z části vlastní senátor Ivo Valenta (Kurzy.cz 2020). V ratingu NFNZ dostaly Parlamentní listy hodnocení C. Problém vidí autoři ratingu v obtížně dostupných informacích o organizaci, nedostatečném využívání online zdrojů jako hypertextový odkaz, malém počtu ověřování zpráv z více zdrojů, zneužívání stereotypů a

zobecňování tvrzení a používání zavádějících titulků. Zpravodajský server Parlamentní listy jsem vybral jako zástupce médií, která ne vždy uvádí na svých stránkách ověřené informace. Ze zpravodajských serverů někdy označovaných také jako dezinformační mají právě Parlamentní listy nejvyšší návštěvnost na svém webu a aktivní komunitu na svých facebookových stránkách.

3.2. Sběr dat a zaznamenávání proměnných

Data potřebná pro analýzu hate speech na sociálních sítích zpravodajských serverů byly sesbírány na stránkách zpravodajských serverů Aktuálně, Novinky.cz, iDnes, ČT24 a Parlamentní listy na sociální síti Facebook. Sběr dat proběhl ve vybraných dnech od 30. dubna 2019 do 19. ledna 2020, nejintenzivnější období pak proběhlo v období duben až květen 2019 a prosinec až leden 2020. Data o jednotlivých zprávách, komentářích a uživatelích se sbírala vždy v komentářové sekci pod sdíleným příspěvkem. Ze všech sdílených příspěvků na sociálních sítích vybraných zpravodajských serverů byly vybrány pouze ty, které se nějak týkají domácí nebo zahraniční politiky a příbuzných témat. Příspěvky a komentáře pod příspěvky byly sbírány vždy nejdříve 12 hodin od zveřejnění proto, aby měli uživatelé dostatek času reagovat na komentář. Data byla sesbírána ručně, tak aby bylo možné zaznamenat všechny dostupné informace o uživatelích, kteří píší komentáře pod příspěvky zpravodajských serverů na Facebooku.

Komentáře pod příspěvky byly vybrány pravděpodobnostním výběrem na základě generování náhodného čísla v daném rozmezí. Facebook v roce 2019 změnil způsob zobrazování komentářů pod příspěvky velkých facebookových stránek. Zobrazování komentářů nyní závisí na následujících faktorech:

- **integrita** – Facebook nezobrazuje komentáře porušující komunitní pravidla,
- **názor uživatelů** – při zobrazování komentářů hraje roli to, zda si uživatel již dříve nepřál některé typy komentářů nezobrazovat,
- **angažovanost** – Facebook se snaží své uživatele přimět k nějakému typu akce, proto zobrazuje příspěvky s více reakcemi a komentáři na lepších pozicích,
- **vlastní nastavení** – uživatel si sám může vybrat, zda chce zobrazit příspěvky chronologicky, nebo vybrat jen ty nejvíce relevantní anebo pouze nejnovější (Vozková 2019).

Pro co největší relevantnost sesbíraných komentářů jsem se rozhodl pro sběr komentářů založit nový účet na Facebooku tak, aby vybrané komentáře nebyly ovlivněny předchozím používáním sociální sítě. Komentáře byly seřazeny dle výchozího nastavení stránek, tedy dle největší relevance. Seřazení komentářů dle relevance odpovídá nejvíce tomu, co vidí běžný návštěvník jednotlivých stránek zpravodajských serverů. Komentáře, které algoritmus vyhodnotí jako méně relevantní jsou většinou napsány delší dobu od zveřejnění příspěvku a mají velmi nízký počet lajků a zobrazení. Při výběru komentářů byly vynechány komentáře, které obsahovaly pouze obrázek, gif, nebo video, ty by bylo velmi obtížné zařadit do studie a určit u nich míru nenávistného projevu. Dále byly vybírány pouze primární komentáře, tedy komentáře, které reagují přímo na příspěvek zpravodajského serveru, komentáře reagující na další komentáře byly při sběru dat vynechány.

Získaná data o každém z komentářů lze rozdělit do tří kategorií dle toho, zda se data týkají příspěvku, komentáře, nebo autora komentáře. V kategorii o příspěvku jsem zaznamenával proměnné druh zpravodajského serveru, odkaz na příspěvek a typ zprávy (domácí nebo zahraniční). U komentáře jsem kromě samotného obsahu komentáře zaznamenával délku komentáře, souhrnný počet reakcí na komentář a vulgaritu komentáře. O autorech komentáře jsem zaznamenával proměnné pohlaví, místo bydliště, vzdělání, profilová fotografie, pravděpodobnost fake profilu a podpora politické strany. Při sběru dat jsem vždy vycházel pouze z toho, co jednotliví uživatelé zveřejňují na svých profilech, pravdivost jimi zveřejněných informací jsem dále nijak neověřoval. Jednotlivým proměnným se budu detailněji věnovat v následujících podkapitolách, proměnné hate speech a jejímu kódování je pak věnována kapitola 3.3.

3.3. Kategorizace sledovaných proměnných

Zkoumaný datový soubor facebookových komentářů obsahuje celkem 800 případů, ale pouze 172 případů má zaznamenáno jak bydliště, tak i vzdělání a další pro analýzu podstatné proměnné. V datovém souboru bylo zaznamenáno celkem 15 proměnných. Hlavní nezávisle proměnnou v datovém souboru je kvalitativní ordinální proměnná hate speech, které bude věnována následující kapitola. Zbylým 14 sledovaným proměnným se věnuji v této kapitole, popíšu jejich škálu, účel a způsob měření. Proměnné v hlavním datovém souboru jsem rozdělil do tří kategorií.

V první kategorii jsou proměnné týkající se sdíleného příspěvku na oficiální facebookové stránce vybraných zpravodajských serverů. Důležitou sledovanou proměnnou v této kategorii je typ zpravodajského serveru, což je nominální kategorická proměnná, která zaznamenává, pod kterým zpravodajským webem byl komentář napsán. Proměnná typ zpravodajského serveru může nabývat hodnot iDnes, Novinky.cz, Aktuálně, ČT24 a Parlamentní listy. Další sledovanou proměnnou je odkaz na příspěvek, který je ve formě hypertextového odkazu na příspěvek, proměnná nebude použita v kvantitativní analýze, ale je zaznamenávána pro potřeby kvalitativní analýzy a zjištění v jakém kontextu byl komentář pod příspěvkem psán. Další sledovanou proměnnou je typ příspěvku, jedná se o dichotomickou proměnnou, která může nabývat hodnot 1 – zpráva se týká domácího dění, 2 – zpráva se týká zahraničního dění. Zprávy byly kódovány na základě zařazení na zpravodajském webu do rubrik "domácí" a "zahraniční" a dále také podle svého obsahu. Zprávy, které komentovaly dopad zahraničního dění do českého prostředí byly zařazeny jako zprávy týkající se domácího prostředí.

Druhá kategorie zahrnuje proměnné, které se přímo týkají napsaných komentářů pod příspěvkem na facebookových stránkách zpravodajských serverů. Proměnné v druhé kategorii vychází z textu samotného komentáře včetně použitých emotikonů. Text komentáře je využit k lingvistické klasifikaci hate speech a pro analýzy kvalitativního charakteru. Z obsahu komentáře vychází dichotomická proměnná emotikony, která kontroluje, zda se v daném komentáři nachází alespoň jeden emotikon. Mezi emotikony jsou počítány jak vestavěné Facebook emotikony přímo v aplikaci, tak i emotikony napsány za pomocí znaků. Další dichotomickou proměnnou je přítomnost vulgarismů, která indikuje to, zda komentář obsahuje nějakou z forem vulgarismů. Pro definici vulgarismů jsem využil Nový encyklopedický slovník češtiny, který spravuje Centrum zpracování přirozeného jazyka Masarykovy univerzity. Vulgarismy jsou zde definovány jako *„výrazy vyjadřující negativní a zároveň emotivní postoj mluvčího k člověku či věci, nesoucí expresivní odstín hrubosti či obhroublosti."* (Jelínek, Vepřek 2017*)* Zahrnují se zde výrazy související se sexuálním aktem či lidskými genitáliemi (mrdat, čůrák, teplouš), dále pak výrazy urážející ženu nebo její chování (běhna, kurva, vykopávka), výrazy spojené s vylučováním a toaletou (hovno, sračka, hajzl), dále pak výrazy podle zvířat (vůl, svině, prase, prasopes, koňomrd) a odumřelých částí rostlin (pařez, stará větev). Do vulgárních výrazů jsou zahrnuty i výrazy rasistické (negr, tatar, křovák), šovinistické (židák, čobol) a mající původ v náboženství (hergot, krucifix, sakra, ježišmarjá). Patří sem i výrazy označující zastánce obecně negativně vnímaných názoru (komouš, nácek,

bolševik), dále pak označující duševní i fyzické vady (idiot, kretén, mrzák, kripl, špekoun, kostra, plešoun) (Jelínek, Vepřek 2017). Za vulgarismy byly považovány i výrazy, které pisatel komentáře sám cenzuroval, ale z kontextu šlo poznat, že jde o slovo vulgární (např. de**l). U každého komentáře v datovém souboru byl změřen počet znaků včetně mezer a zaznamenán jako číselná proměnná délka komentáře. Dále byla vytvořena proměnná reakce na komentář, která zaznamenává, jaké byly souhrnné reakce na komentář. Proměnná je tvořena součtem všech Facebook reakcí, které jde na komentář dát, jsou to reakce typu to se mi líbí, super, haha, paráda, to mě mrzí, a to mě štve. Jednotlivé typy reakcí na komentář nejsou ve zkoumaném datovém souboru zahrnuty. Do proměnné reakce na komentář nijak nezasahují reakce na komentář formou dalšího komentáře. Proměnnou reakce na komentář lze považovat za měřítko popularity daného komentáře, ať už se jedná o popularitu pozitivní nebo popularitu negativní.

U uživatelů, kteří psali komentáře pod sledovanými příspěvky se zaznamenávalo několik proměnných, výběr proměnných vycházel z běžně dostupných informací, které uživatelé sdílejí na svých facebookových profilech. I přes výběr běžně dostupných informací se nepodařilo u všech případů zaznamenat všechny informace, a to především proto, že část uživatelů údaje jako bydliště nebo vzdělání neuvádí a část uživatelů si své osobní profily nastavuje jako soukromé.

Proměnnou, která byla zjistitelná téměř u všech případů je kategorická proměnná pohlaví, která byla určována podle jména a profilové fotky. U některých profilů nebylo možné určit pohlaví, především proto, že se jednalo o facebookové stránky typu *Kolik je v Pardubicích stupňů* nebo *Čeští elfové*. Takovéto profily byly v datovém souboru označeny jako 0, muži jako 1 a ženy jako 2. Kategorická proměnná bydliště byla zaznamenávána na úrovni krajů, jako místo bydliště bylo zaznamenáno aktuální místo pobytu uvedené na profilech uživatelů. U uživatelů, kteří měli jako místo pobytu uvedené město v zahraničí byl zaznamenán pouze stát a budou do analýzy vstupovat jako kategorie zahraničí. Další proměnnou v datovém souboru je vzdělání, které je vedeno jako ordinální proměnná a byla zaznamenána celkem u 225 případů. Kategorie vzdělání jsou rozděleny na Základní škola + fiktivní škola, Střední odborné učiliště, Střední škola a Vysoká škola. Jako vzdělání je u uživatelů bráno studium na nejvyšší vzdělávací instituce, není zde nijak ošetřeno, zda uživatel vzdělání dokončil nebo nikoliv a zda vůbec do uvedené školy chodil. Do kategorie Základní škola + fiktivní škola jsou zařazeny i všechny fiktivní neakreditované školy typu *Vysoká škola života* a *Vysoká škola pro idioty*. Uživatelé takovýchto titulů jsou často uživateli s nižším vzděláním, kterým při zadávání

vzdělání do profilů na sociálních sítích nebylo příjemné zadávat nižší vzdělání, proto se často na jejich profilech objevuje jako vzdělání Vysoká škola života. Vzhledem k těmto okolnostem bylo toto vzdělání zařazeno do kategorie společně se základním vzděláním.

U každého z profilů na sociální síti Facebook byl zaznamenán i grafický obsah profilové fotky a zařazen do jedné z kategorií. Kategorie pro proměnnou profilová fotka jsou lidská fotografie (reálná fotografie člověka s jasně rozpoznatelným obličejem), obrázky a fotografie věcí a známých osobností (kreslený obrázek nebo fotografie obsahující předmět, krajinu, známého herce atd.) a bez profilové fotky (defaultní obrázek Facebook pro profil bez profilové fotky). Vzhledem k častému výskytu fotek zvířat jako profilových fotek jsem se rozhodl pro profilovou fotku se zvířetem vlastní kategorii, nejčastěji se pak mezi zvířaty na profilové fotce objevovali psi a kočky.

Někteří uživatelé na facebooku vyjadřují své politické preference a názory, nejčastěji se tak děje před volbami, kdy za pomocí nejrůznějších nástrojů mohou uživatelé upravovat své profilové fotky a vkládat si tak loga politických stran, hnutí i jednotlivých kandidátů do profilové fotky a vyjádřit tak svoji podporu. Pro zaznamenání těchto preferencí jsem se rozhodl vytvořit kategorickou proměnnou podpora politické strany. Podpora politické strany byla zaznamenána, pokud měl uživatel na svém profilu logo nebo jiné propagační materiály strany, hnutí nebo jednotlivců. Nejčastěji uživatelé podporují politické strany, hnutí, spolky, ale také například jednotlivé kandidáty na prezidenta.

Falešné profily a komentáře internetových trollů v diskusích jsou dnes již nedílnou součástí většiny sociálních sítí. Autoři jako Jessikka Aro píší dokonce o tom, že propaganda za pomocí internetových trollů s falešnými profily v diskusích je novým způsobem vedení války. Nejčastěji se setkáváme s tímto pojmem v souvislosti s Ruskou informační válkou, která se za pomocí placených, anonymních a agresivních komentátorů na sociálních sítích snaží ovlivnit veřejnou debatu a útočí na psychiku a myšlení lidí (Aro 2016). Problém internetových trollů, dezinformací a informační války nemá pouze Finsko a Pobaltské země, ale týká se i České republiky a dalších zemí, kde se především v diskuzích na sociálních sítích objevují podezřelé profily, které se snaží rozdělovat společnost. Proto jsem se rozhodl sledovat podezřelé uživatele a zahrnout je jako proměnnou do analýzy hate speech. Rozpoznání falešného profilu ale není vůbec snadné, existují zde metody, které detekují falešné profily na základě aktivity většího počtu účtu, které vykazují podobné charakteristiky. Další možností je analyzovat individuální účty a hodnotit je na základě dostupných informací na profilu uživatele (Romanov et. al 2017).

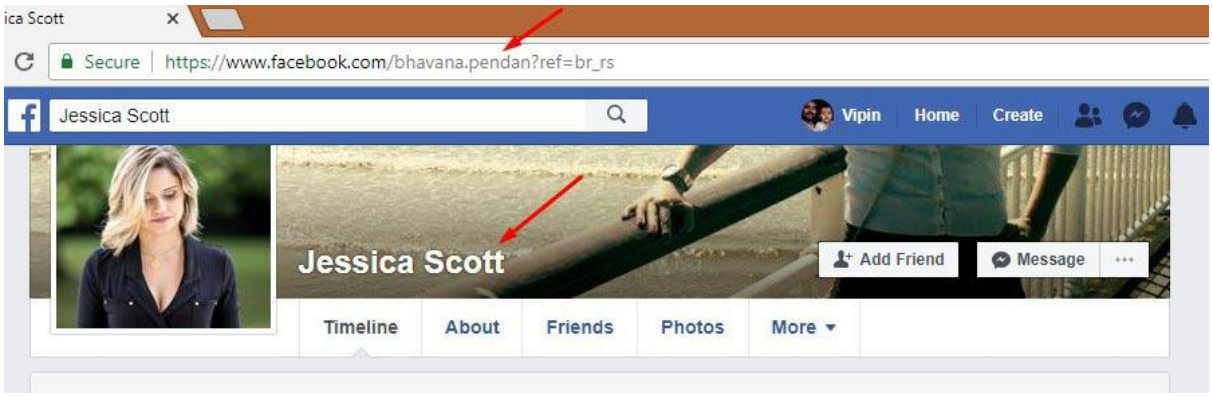
Pro analyzování fake účtů píšící komentáře pod příspěvky zpravodajských serverů jsem se rozhodl použít přístup zaměřující se na jednotlivé profily, který je více vhodný na menší počet případů. Profily na sociálních sítích lze označit za falešné, pokud vykazují některé z následujících charakteristik:

- profil má pouze několik fotek, které většinou ani nejsou fotky skutečných lidí, nebo se jedná o fotky herců a slavných osobností,
- profil byl vytvořen v posledních dvou letech (platí pouze u starších osob) a nevykazuje žádné znaky dlouhodobého používání,
- profil s vámi nemá žádné společné přátele ani zájmy, a i přes to se vás snaží přidat do přátel,
- falešný profil přidaný do přátel většinou nevykazuje žádnou aktivitu a nesnaží se o žádnou interakci s vaším účtem (Aronovich 2018).

Výše zmíněné charakteristiky se týkají účtů na sociálních sítích, které se vás snaží přidat do přátel především za účelem phishingového útoku. Pro analyzování falešných profilů ve zkoumaném datovém souboru je důležité se více zaměřit na charakteristiky dostupné přímo na falešném facebookovém profilu. Falešný profil můžeme rozpoznat podle následujících znaků:

- profilová fotka – falešné profily mají většinou jen pár fotek, které jsou navíc většinou fotkami celebrit nebo vůbec neobsahují lidskou tvář,
- informace o uživateli – uživatelé na facebooku o sobě často sdělují informace kde bydlí, co studovali a kde pracují, u falešných profilů většinou tyto informace chybí, nebo jsou vyplněny nesmyslnými informacemi,
- přátelé – pokud uživatel nemá žádné přátele anebo až příliš přátel z ciziny, je zde pravděpodobnost, že se jedná o falešný účet,
- rozdílné jméno profilu a URL – falešné profily často mění své jména, nebo se může jednat o odcizený účet, který si změnil své jméno, ukázku rozdílného jména profilu a jména v URL je na obrázku 3.1,
- sdílené příspěvky – pokud má uživatel málo sdílených příspěvků, sdílené příspěvky odkazují pouze na několik stránek, sdílí velké množství příspěvků v poslední době jedná se pravděpodobně o falešný profil,
- interakce s uživatelem – dalším znakem falešného profilu je nedostatek interakce s ostatními uživateli, účet tak má málo komentářů a lajků pod svými příspěvky (Polston 2018).

Obrázek 3.1: Rozdíl v profilovém jménu a v URL



Zdroj: Polston 2018

Profily jsem v diplomové práci analyzoval za pomocí kritérií dle Vince Polston a vytvořil kategorickou proměnnou pravděpodobný falešný profil, která hodnotí profily na škále od 1 do 3. Profily označené jako 1 jsou zaručeně profily skutečných lidí, obsahují dostatek fotek dané osoby, jsou aktivní a mají mnoho přátel, profily v kategorii 2 jsou pravděpodobné falešné profily, které splňují pouze část z určených kritérií. Profily ve třetí kategorii jsou velmi pravděpodobně falešné, takovéto profily splňují většinu určených kritérií, často nemají na profilu fotku reálné osoby, byly založeny nedávno a mají pouze pár přátel nebo žádné, aktivita pod jejich příspěvky je velmi malá. Tato analýza falešných profilů ale není stoprocentní, někteří uživatelé si chrání své soukromí a dovolují zobrazovat svůj profil cizím lidem pouze v omezené formě, tyto případy mohou být zahrnuty do třetí kategorie i přes to, že majitel takového profilu je reálná osoba.

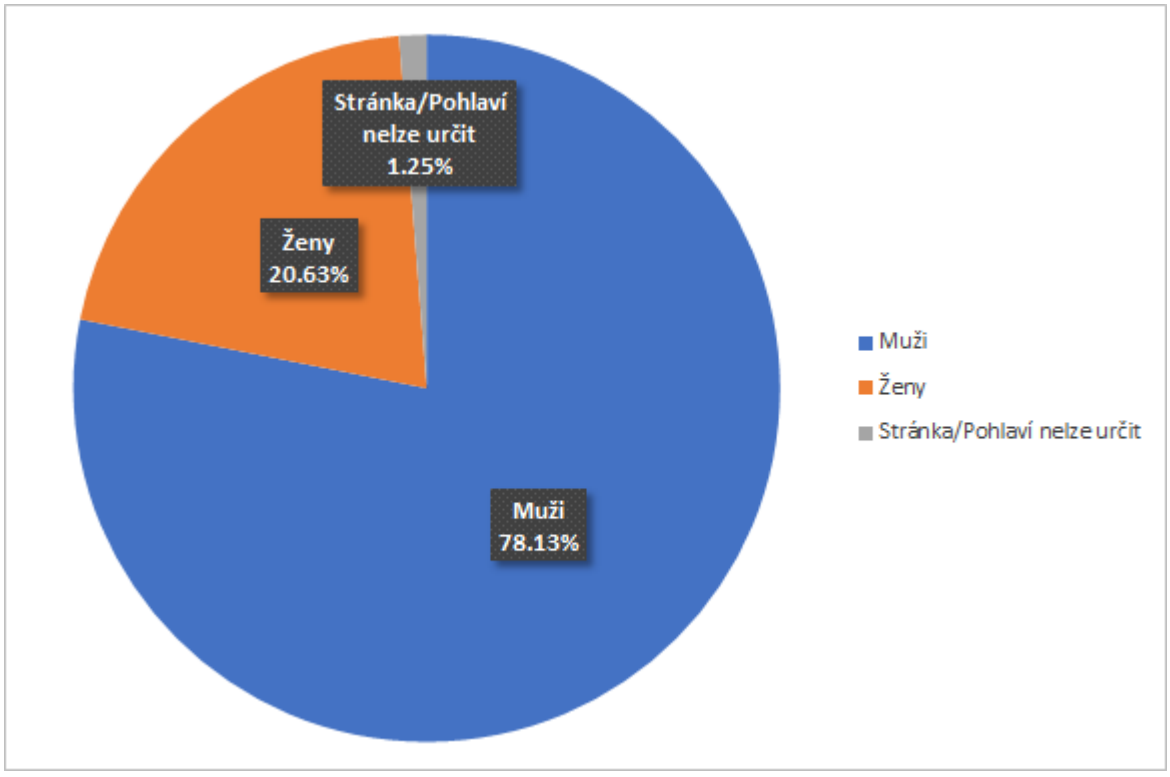
3.4. Kdo komentuje příspěvky na Facebooku

Na základě analyzovaného datové souboru lze identifikovat komentátory na sociálních sítích zpravodajských serverů a určit tak jejich hrubý socio-demografický profil. V této kapitole provedu deskriptivní analýzu dat, která byla sesbírána na profilech komentátorů, kteří komentovali příspěvky vybraných zpravodajských serverů na sociální síti Facebook. Zaměřím se na pohlaví komentátorů, vzdělání, místo bydliště, podporu politické strany, typ profilové fotografie a pravděpodobnost falešného profilu. Podrobné analýze komentářů, které obsahují hate speech se věnuji v kapitole 4.

Pod příspěvky vybraných zpravodajských serverů převažují komentáře od profilů, které jsou anebo se vydávají za profily mužů. Z celkového počtu 800 komentářů bylo těch od mužů 625, tedy 78,1 %. Jedná se o poměrně vysoké procento zastoupení mužů, při srovnání například

se studií Projevy nenávisti v online prostoru na sociálních sítích (Hrdina, Daňková, Kopecká 2017), kde autoři zkoumali pouze komentáře, které obsahovaly klíčová slova spojená s nenávistnými projevy, počet mužský profilů zde byl 60 %, což je o 18,1 % méně než v případě zkoumaného datového souboru. Komentářů od ženských profilů bylo 165, tedy 20,6 %. U ostatních komentářů buďto nešlo určit pohlaví pisatele, nebo komentář psal někdo skrze stránku na Facebooku, kde je pohlaví autora nedohledatelné, takovýchto komentářů bylo v souboru dat 10, tedy 1,3 %. Převaha mužských profilů v tomto výzkumu je dána především charakterem zpravodajských příspěvků, kde byly vybírány zprávy týkající se domácí i mezinárodní politiky, ekonomie a financí, o tyto témata mají mnohem menší zájem ženy než muži. Naopak ženy se zajímají více o témata týkající se kriminálního zpravodajství, kultury, zdraví a víry (Rosentiel 2008).

Graf 3.1: Zastoupení pohlaví v datovém souboru facebookových komentářů

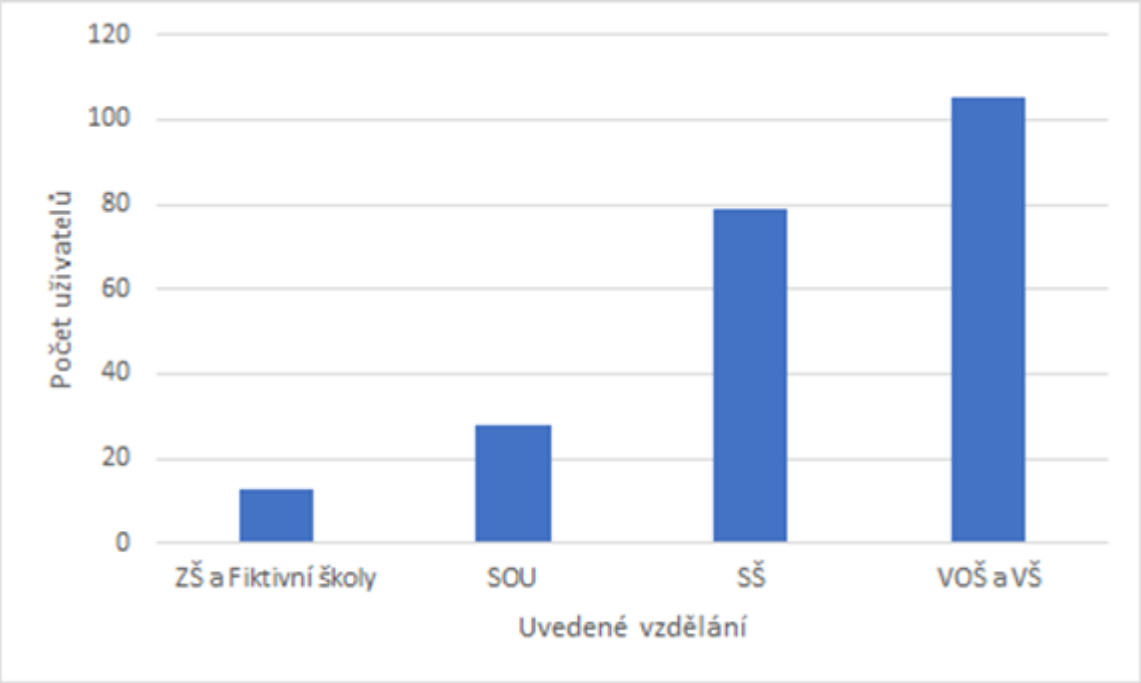


Zdroj: vlastní

Vzdělání na svém Facebookovém účtu uvedlo pouze 225 z celkového počtu 800 komentářů. Uvedené vzdělání na profilu nemusí vždy odpovídat dosaženému vzdělání autora komentáře. Nejčastěji uživatelé komentující pod příspěvky zpravodajských serverů uváděli, že mají vyšší odborné nebo vysokoškolské vzdělání, takových profilů bylo celkem 105, tedy 46,7 % ze všech komentátorů, kteří měli na profilu uvedené vzdělání. Středoškolské vzdělání mělo

na profilu uvedeno 79 případů což tvoří 35,1 % ze všech uvedených vzdělání. Jako své vzdělání uvedlo střední odborné učiliště 28 účtů, které tvoří 12,4 %. Nejméně uživatelé uváděli základní vzdělání, do kterého byly zařazeny i fiktivní školy jako Vysoká škola života a Vysoká škola pro idioty s.r.o. takovéto uvedené vzdělání mělo pouze 13 účtů, což tvoří pouze 5,8 % z komentátorů, kteří uvedli své vzdělání. Nízké procento základních škol a středních odborných učilišť lze vysvětlit tím, že uživatelé se mnohdy za své nízké dosažené vzdělání stydí, a proto na svých sociálních sítích raději žádné neuvedou, nebo uvádí výše zmíněné fiktivní vzdělání jakými je nejčastěji Vysoká škola života.

Graf 3.2: Uvedené vzdělání sledovaných účtů na Facebooku

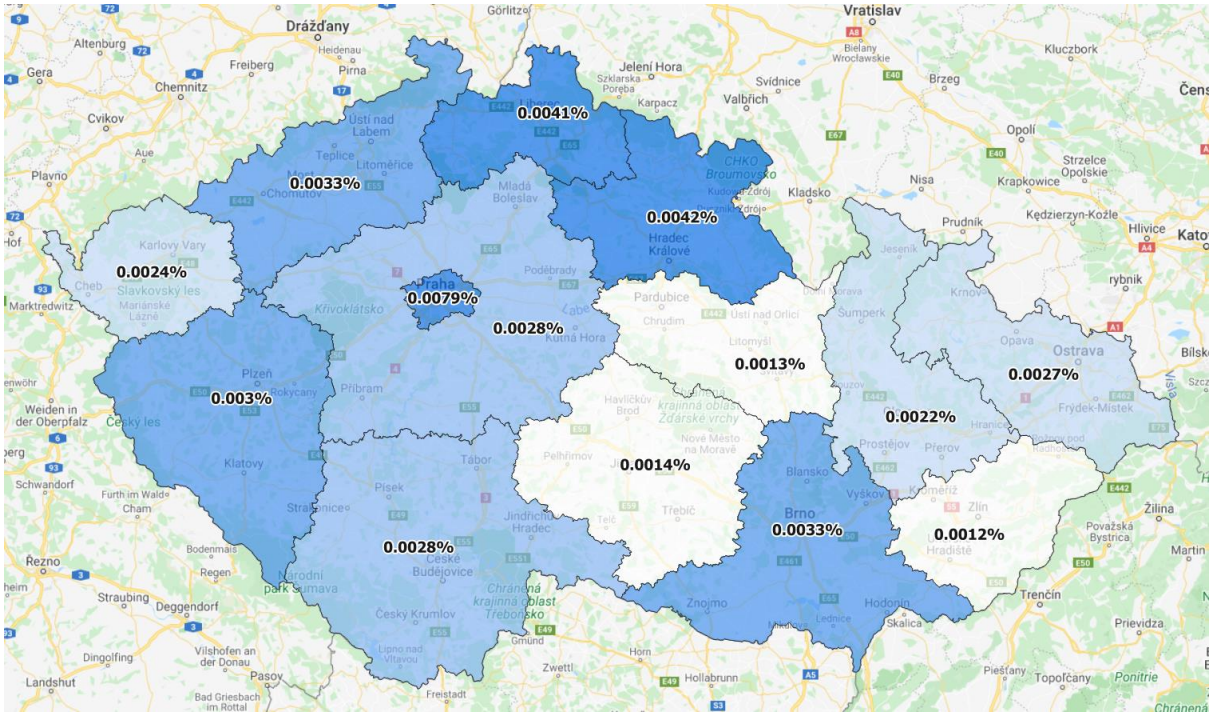


Zdroj: vlastní

Bydliště na profilu bylo uvedeno celkem u 401 komentářů, nejčastěji lidé uváděli, že bydlí v Praze, kterou uvedlo jako své bydliště 103 uživatelů, tedy 25,7 %. Poměrně často uváděli komentující uživatelé na svých profilech jako bydliště zahraničí, celkem bylo 42 případů (asi 10,5 %), které jako místo bydliště uváděli zahraničí. Mezi nejčastěji uváděné kraje, kromě Prahy, patří také Středočeský kraj (39 případů), Jihomoravský kraj (39 případů) a Moravskoslezský kraj (32 případů). Samozřejmě absolutní počty případů nejsou vypovídající, proto jsem absolutní počet případů z jednotlivých krajů přepočetl jako procento obyvatel v daném kraji, které jsem následně vnesl do mapy na obrázku 2.2. Největší procento případů vůči obyvatelům v kraji má Praha, která dosahuje 0,0079 %. Vyšší zastoupení je i na severu Čech v Ústeckém, Libereckém a Královéhradeckém kraji, kde se procento případů pohybuje mezi

0,0033 % a 0,0042 %, a také v Jihomoravském kraji, kde je hodnota 0,0033 %. Naopak nejnižší procentuální hodnoty jsou v kraji Vysočina a Pardubickém a Zlínském kraji, kde se pohybují mezi 0,0012 % a 0,0014 %.

Obrázek 3.2: Mapa poměru počtu uživatelů vůči celkovému počtu obyvatel kraje

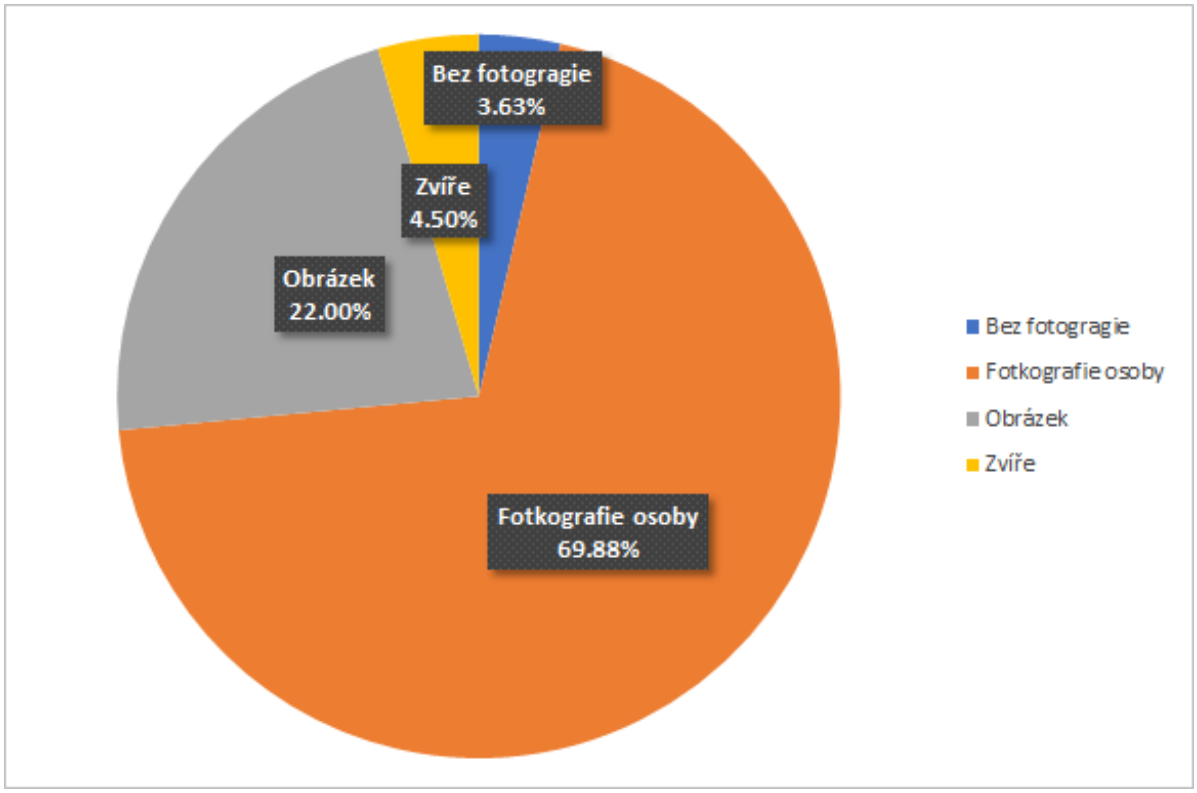


Zdroj: vlastní (zpracováno v QGIS)

Podpora politické strany, hnutí nebo neziskové organizace formou loga nebo sloganu na profilové fotce byla poměrně častým jevem, z celkového počtu 800 komentářů jich 109 mělo na svém profilu logo nebo slogan některé z organizací nebo osobností. Nejčastěji uživatelé na svých profilech vyjadřovali podporu zájmovému spolku Milion chvilek, který od roku 2018 pořádá demonstrace namířené proti vládě a premiéru Andreji Babišovi. Podporu spolku Milion Chvilek na svém profilu deklarovalo 22 případů, což je 20,2 % ze všech komentářů, které vyjadřovali politickou podporu. Mezi politickými stranami nejčastěji lidé podporovali Českou pirátskou stranu, kterou podporovalo 21 facebookových profilů. Vyšší podporu mezi stranami zaznamenala i strana Svoboda a přímá demokracie, kterou na svém profilu podporovalo 11 uživatelů, 5 uživatelů podporovalo nově vzniklé hnutí Trikolóra hnutí občanů. Mezi prezidentskými kandidáty měl největší podporu Jiří Drahoš (10 uživatelů) a Miloš Zeman (3 uživatelé). Ostatní prezidentští kandidáti a politické strany a hnutí měli pouze jednotky podporovatelů.

U každého z profilů byl zaznamenán typ profilové fotky, kromě fotografií osob uživatelé jako profilovou fotku často používají i obrázek z filmu, video hry, karikatury, ilustrace, vlajky a jiné. Často uživatelé používají fotky zvířat jako profilové fotky, nejčastěji pak psů a koček. Celkem v 559 případech měli uživatelé na profilové fotce reálnou osobu, která zároveň nebyla herec nebo jiná slavná osobnost. Obrázek jako profilovou fotku mělo nastaveno 176 profilů a zvíře v profilové fotce mělo 36 profilů z celkového počtu 800 sledovaných případů. Bez obrázku či fotografie bylo 29 případů, přehledné zastoupení jednotlivých typů profilové fotky je v grafu 2.3.

Graf 3.3: Typ profilové fotky uživatelů



Zdroj: vlastní

Pravděpodobnost toho, že se jedná o falešný profil byla hodnocena na škále 1-3, dle kritérií v kapitole 3.3. Profilů označených jako ověřené profily s historií, určitým počtem fotek a jasným správcem účtu bylo 434, tedy 54,3 % případů. U 187 případů bylo jisté podezření, že by se mohlo jednat o falešný účet, především proto, že účet nesplňoval část sledovaných kritérií. Profilů označených jako falešné bylo 179, což je 22,4 % ze všech profilů, takovéto profily většinou neměly jako profilovou fotku osobu a na profilu se nenacházely žádné fotky osob, kterým by daný facebookový profil mohl patřit. Hodnocení falešného profilu je velmi subjektivní záležitost a lidé si často chrání své soukromí a nemají veřejně přístupný profil,

proto je potřeba brát pravděpodobnost falešného profilu jako doplňující proměnnou pro tento výzkum.

3.5. Charakteristika komentářů na facebookových stránkách zpravodajských serverů

V této kapitole se věnuji deskriptivní analýze proměnných, které se týkají obsahu sledovaných komentářů, jsou jimi typ zpravodajského sdělení, používání emoji, délka komentáře, počet reakcí na komentář a vulgárnost komentáře.

Z celkového počtu 800 komentářů bylo 593 komentářů napsáno pod příspěvkem, který se věnoval tuzemskému zpravodajství a politické situace, 207 komentářů bylo napsáno pod příspěvky věnující se zahraničnímu zpravodajství. Použití jednoho nebo více emoji se objevilo v celkem 205 komentářích, což tvoří 25,9 % všech komentářů. Nejčastěji uživatelé používali v komentářích některou z forem smějícího se emoji, kterým doplňovali převážně ironické nebo sarkastické komentáře.

Průměrná délka komentářů pod příspěvky byla 160,87 znaků. Nejkratší komentář měl pouze 8 znaků, naopak nejdelší komentář měl 1212 znaků. Medián v délce komentářů byl 117, z histogramu délky komentářů je patrné, že uživatelé píší nejčastěji komentáře o délce 40 až 240 znaků, delší komentáře jsou spíše výjimkou. Počet reakcí na komentář se pohyboval v rozmezí od 0 do 460 s průměrem 23,82 a mediánovou hodnotou 9,5. Nejčastěji komentář získal pouze jednu reakci, a to v 63 případech a 90 % komentářů získalo méně než 57 reakcí. Vyšší počet reakcí na komentář je tak ojedinělý, že komentářů, které získali více než 100 reakcí bylo v celém datovém souboru pouze devět. Vulgárnost komentářů byla hodnocena poměrně přísně a za vulgární komentáře byly označeny i fráze obsahující přirovnání ke zvířatům, nebo méně vulgární nadávky a výrazy. I přes takto přísné hodnocení bylo za komentáře obsahující vulgární výrazy označeno pouze 45 případů, což je pouze 5,6 % ze všech komentářů.

4. Analýza hate speech komentářů na sociálních sítích

V následující kapitole jsou nejprve představeny vybrané kvalitativní a kvantitativní přístupy detekování komentářů obsahující hate speech na sociálních sítích, následně jsou představeny některé možnosti analyzování a kategorizování hate speech. V poslední části kapitoly se zabývám zvolenými způsoby analýzy a kategorizací proměnné hate speech pro účely této diplomové práce.

4.1. Metody detekování hate speech

V posledních letech se objevila řada výzkumů zabývající se detekováním hate speech a dalšího nevhodného chování v online prostoru. Termín hate speech bývá často spojován s útočnými, urážlivými projevy a kyberšikanou především na sociálních sítích a dalších internetových stránkách umožňující diskusi mezi návštěvníky. Existující metody detekce hate speech a příbuzných problémů se primárně zaměřují na klasifikaci dokumentů (komentářů, příspěvků, článků). Metody detekování hate speech lze rozdělit do dvou kategorií: první kategorie se spoléhá na manuální hodnocení dat, které je následně zpracováno algoritmy jako Support Vector Machines a Naive Bayes. Tyto způsoby detekování jsou označovány jako *klasické metody* (classic methods), tyto metody používají ve svých pracích například Burnap a Williams (Burnap, Williams 2015) nebo Davidson et al. (Davidson et al. 2017). Druhá kategorie nazývána *deep learning metody* (deep learning methods) reprezentuje více aktuální přístup, kdy detekuje hate speech za pomoci deep learning paradigmat, neuronových sítí a učení se z prvotních dat (Zhang, Luo 2018: 1–3). Tento přístup využívají například Nobata et al. (2016) nebo Park a Fung (2017).

Klasické metody vyžadují nejprve manuálně navrhnout a nakódovat jednotlivá kritéria pro detekci hate speech, které následně použijí klasifikátoři k manuální detekci hate speech. Pro detekci nejen hate speech existuje několik funkcí, jsou jimi *Simple Surface Features, Word Generalization, Sentiment Analysis, Lexical Resources, Linguistic Features, Knowledge-Based Features, Meta-Information a Multimodal Information* (Schmidt, Wiegand 2017). Následně se na datový soubor použijí klasifikátory, které umožní automaticky ohodnotit větší množství případů za pomoci algoritmů na základě manuálně ohodnoceného souboru hate speech. Nejpoužívanější algoritmus je Support Vector Machines (SVM), mezi další populární algoritmy patří Naive Bayes, Logistická Regrese a Random Forest (Zhang, Luo 2018: 3).

Deep learning metody využívají umělé neuronové sítě k naučení se abstraktních funkcí ze surových dat a na základě skládání mnoha vrstev tak vznikne klasifikace pro detekci hate speech. Jako základ pro deep learning metody mohou být jak surová data, tak i různé formy kódovaných dat, například některé z *klasických metod*. Model využívající deep learning nepoužívá vložená data přímo ke klasifikaci hate speech, ale prvně se naučí dle dostupných dat svou vlastní definici na základě, které následně hodnotí, zda případ obsahuje hate speech či nikoliv. Tato metoda je čím dál více používána pro rozpoznávání hate speech a dosahuje také lepších výsledků než metody klasické. Nejpopulárnějšími deep learning architekturami pro rozpoznávání hate speech jsou Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory network (LSTM) (Zhang, Luo 2018: 3–4).

4.2. Kategorizace a analýza hate speech

V předchozí kapitole jsem se zaměřil na detekování hate speech na sociálních sítích, jednalo se tedy pouze o to, zda hate speech přítomna je nebo není. Tato kapitola si klade za cíl rozšířit toto detekování a představit vybrané kategorie, dle kterých je dále možné hate speech dělit. Před tím, než zde budou prezentovány jednotlivé přístupy k měření a kategorizaci hate speech v online prostředí, je nutné definovat, co je pod pojmem hate speech nebo někdy také harm speech myšleno. Detailněji se pojmu hate speech a jeho definicím věnuji v kapitole 1.3., pro účely této kapitoly jsem zvolil více obecnou definici, která definuje hate speech jako *„urážlivý příspěvek, který je zcela nebo zčásti motivovaný zaujatostí autora proti aspektům určité skupiny lidí.“ (Mondal et al. 2017)* Takto široká definice zahrnuje nejen kriminální zločiny z nenávisti, ale také některé behaviorální a fyzické aspekty, které nemusí být nutně kategorizovány jako zločiny. (Mondal et al. 2017)

V rámci odborné literatury se nejčastěji setkáváme s kategorizací na základě toho, jaká je cílová skupina nenávistného sdělení. Takovou kategorizaci využívají i Mondal, Silva a Benvenuto (Mondal et al. 2017), kteří za pomoci analýzy struktury vět zkoumají data ze sociálních sítí Twitter a Whisper. Jejich výzkum je postaven na vyhledávání nenávistných vět na základě předem stanovených větných konstrukcí, nejčastěji používaná větná konstrukce má tvar *Já <intenzita> <záměr uživatele> <cíl hate speech>*. V analýze Twitter postů se nejčastěji objevovalo slovní spojení “I hate Nigga” (Já nesnáším negry). Jednotlivé cíle hate speech byly zařazeny do kategorií, dle toho, na jakou skupinu byl nenávistný projev zaměřen. Kategorie, které používají autoři jsou shrnuty v tabulce 4.1 společně s příklady cílů hate speech u jednotlivých kategorií. Kategorizaci založenou na cílové skupině nenávistného projevu

používají i ElSherief, Kulkarni, Nguyen, Wang a Belding (ElSherief et al. 2018), kteří ve své práci analyzují příspěvky na Twitteru a rozdíly mezi přímým a generalizovaným používáním hate speech. Vytváří tak další možné kategorizování hate speech na přímou a generalizovanou. Za pomocí SAGE vybrali ke každé z kategorií založených na cílové skupině pět nejpoužívanějších slov v nenávistných projevech, a to jak u přímého, tak generalizovaného hate speech. V práci autoři dělí kategorie podobně jako Mondal et al. na archaismy, postižení, gender, víra, třída, etnicita, národnost, a sexuální orientace.

Tabulka 4.1: Kategorie hate speech dle cíle nenávistného projevu

Kategorie	**Příklad cílů nenávistných projevů**
Rasa	negři, černí lidé, bílí lidé, asiaté
Chování	nejistí lidé, pomalý lidé, citliví lidé
Fyzické dispozice	obézní lidé, malí lidé, fyzicky přitažliví lidé
Sexuální orientace	homosexuálové, heterosexuálové
Třída	lidé pocházející z ghett, bohatí lidé
Gender	těhotné ženy, sexističtí lidé
Etnicita	číňané, indiáni, pákistánci
Postižení	mentálně postižení lidé, bipolární lidé
Víra	muslimové, židé, křesťané
Ostatní	opilí lidé, povrchní lidé

Zdroj: Mondal et al. 2017

S odlišným přístupem ke klasifikování hate speech přišli autoři Sharma, Agrawal a Shrivastava (Sharma et al. 2018), kteří se na kategorizaci hate speech dívají z více filozofického úhlu pohledu a snaží se ji pochopit na základě emocí a slovesa nenávist. Ve své práci vychází z teorie nenávisti od Karin Sternberg a zaměřují se při zkoumání hate speech na emoce a pocity. Hate speech rozdělují do jednotlivých tříd dle toho, jaký byl nenávistný záměr z pohledu pisatele hate speech. Nabízí tak opačný pohled na klasifikování hate speech než Mondal et al. nebo ElSherief et al., rozdělují hate speech do tří tříd (Class I, Class II a Class III), které jsou seřazeny sestupně. Na obrázku 4.1 je názorně ukázána škála se zvýrazněnými hranice mezi jednotlivými třídami a příklady hate speech, které byly nalezeny na sociálních sítích (extrémismus, vyhrožování nebo trolling a sarkasmus). Pro rozdělení hate speech do jednotlivých kategorií autoři stanovily následující pokyny:

Class I:

- Podněcuje nenávistnou řečí k násilným činům nad rámec samotného nenávistného sdělení.
- Může být zaměřena jak na určitou skupinu, tak i na jednotlivce. Nenávistné a násilné chování vůči skupině je hodnoceno jako závažnější stupeň hate speech, než nenávistné a násilné chování na úrovni jednotlivce.
- Z kontextu je patrné, že úmysl řečníka je ranit city nebo páchat násilné činy vůči určité skupiny nebo jednotlivci.

Class II:

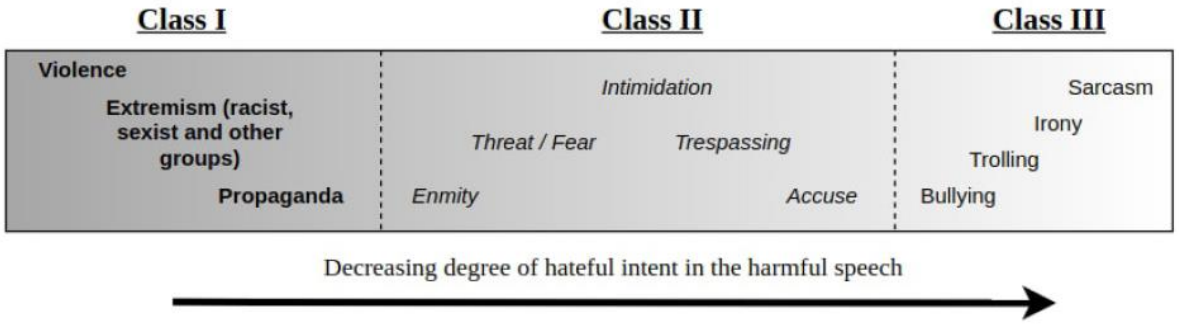
- Online žertování a provokování, které obsahuje obviňování, vyhrožování, používání agresivních/provokativních frází k vyjádření nesouhlasu a vyzývání k slovním soubojům.
- Násilné projevy jsou méně závažné než v Class I, které mohou ranit city oběti, ale nepodněcují k násilným činům.
- Hate speech sdělení se pohybuje mezi lingvistickým násilím a zastrašováním jedince v online prostředí. Sdělení může být vysoce provokativní a je častěji směřováno vůči jedinci než skupině nebo ideologii.

Class III:

- Mírně provokativní charakter sdělení většinou zaměřený na jednotlivce než na skupinu.
- Jsou používány více vulgární a provokativní výrazy a obsah sdělení je spíše trollení, ironie a sarkasmus než projev nenávisti.
- Nepřímo zraňuje city a obsahuje méně závažné projevy nenávisti než v Class I a Class II.

Autoři Sharma, Agrawal a Shrivastava (Sharma et al. 2018) svoji kategorizaci hate speech aplikují na anglicky psané příspěvky na Twitteru, které nejprve kvalitativně hodnotí a zařazují do jednotlivých tříd na základě kterých provádí automatické testování hate speech za pomoci machine learningu a testují trojici klasifikačních systémů (Naive Bayes, Support Vector Machines a Random Forest) a jejich přesnost při určování hate speech tříd.

Obrázek 4.1: Spektrum ukazující třídy, s jakým záměrem je pácháno hate speech



Zdroj: Sharma et al. 2018

Autoři Miro-Llinares a Rodrigez-Salsa (2016: 406–415) se ve své práci věnují analyzování násilných a nenávistných projevů na Twitteru. Analýza se zaměřuje pouze na příspěvky a konverzace týkající se útoku na redakci satirického časopisu Charlie Hebdo, byly proto vybírány pouze příspěvky označené hashtag[12] #CharlieHebdo, #JeSuisCharlie a #StopIslam. Pro rozlišení a kategorizaci příspěvků na příspěvky obsahující hate speech, násilné příspěvky a neutrální příspěvky autoři použili kvantitativní analýzu dat. Vybrané komentáře vyhodnotily tři páry hodnotitelů v každém z následujících pěti kritérií:

- Kritérium 1: Vážné urážky, ponižující výrazy nesporného charakteru namířené proti určitým nebo neurčitým osobám a určitým nebo neurčitým skupinám.
- Kritérium 2: Vyvolávání pozitivního přístupu k násilí na lidech určité skupiny ospravedlněné tím, že se jedná o obranu, glorifikaci, trivializace, podněcování, indukce, porozumění, radost atd.
- Kritérium 3: Přisuzování urážlivých výrazů konkrétním jednotlivcům, jejich veřejné ponižování, způsobování závažných nepříjemných pocitů. Také se zde řadí případy kriminálních aktů a vážného napadení.
- Kritérium 4: Projevy nenávisti nebo pohrdání zaměřené vůči určitým skupinám, speciálně těm, které se cítí ohrožené, bojí se omezení svých práv a jsou terčem určité nesnášenlivosti. Patří sem zejména ty výrazy, které tyto skupiny hanlivě uráží, nebo žádají po omezení práv těmto skupinám.
- Kritérium 5: Hanebné výrazy a nevhodně komentované události, které mohou zapříčinit u obětí takovéto hate speech silné nepříjemné pocity, zejména se jedná o výrazy, které

[12] Hashtag je slovo nebo skupina slov začínající symbolem mřížky (#), nejčastěji se s použitím hastagu můžeme setkat na sociálních sítích jako Twitter, Instagram nebo Facebook, kde slouží pro spojení různých příspěvků s podobným tématem (Objevit.cz 2015).

projevují nenávist vůči lidem, nebo svým obsahem zcela dehumanizují určitou skupinu. Patří sem vtipy a černý humor, zvláště pak vztažený vůči událostem nenásilného charakteru (přirozená, náhodná smrt) a způsobují velkou bolest i nepřímým obětem (Miro-Llinares, Rodrigez-Salsa 2016: 407–408).

Hodnocení příspěvků a komentářů bylo hodnoceno dle výše zmíněných kritériích ve čtyřech vlnách tak, aby hodnocení dat bylo validní a reliabilní. Za pomoci Kappa Testu vytvořili index shody mezi jednotlivými hodnotiteli, který dosahoval vysoké reliability (Kappa = 0,91). Výsledky kvalitativní analýzy byly použity pro rozsáhlejší kvantitativní analýzu, která obsahovala 282 397 tweetů, z kterých bylo 2 304 obsahovalo jistou formu hate speech a násilí. Vybrané komentáře dále autoři rozdělují dle vlastní taxonomie do čtyř kategorií: podněcování k násilí (violent incitement), osobní útok (personal offence), podněcování k diskriminaci (discrimination incitement) a kolektivní útok (collective offence). Autoři na základě analýzy hate speech zjišťují možné prediktory, které by přítomnost hate speech mohly indikovat, zaměřují se například na čas zveřejnění příspěvku, počet sledujících, nebo použitý hashtag (Miro-Llinares, Rodrigez-Salsa 2016). Dále pak lze zmínit autory Gambäck a Sikdar (Gambäck, Sikdar 2017), kteří kategorizují hate speech dle obsahu sdělení do následujících kategorií: neobsahuje hate speech, obsahuje sexistickou hate speech, obsahuje rasistickou hate speech a obsahuje jak rasistickou, tak sexistickou hate speech.

Možností, jak klasifikovat hate speech je mnoho a vzhledem k tomu, že každý člověk vnímá hate speech subjektivně, je těžké určit pouze několik kategorizací, podle kterých hodnotit hate speech. Často se používá kategorizace dle pachatelů hate speech a jejich obětí, nebo také podle toho, zda se nenávistný projev soustředí na jednotlivce nebo skupinu osob.

4.3. Vybrané způsoby pro analyzování hate speech na sociální síti Facebook

Pro analyzování hate speech v datovém souboru facebookových komentářů jsem se rozhodl použít vlastní škálu hate speech, na základě, které budou hodnoceny komentáře. Komentáře budou analyzovány a hodnoceny ručně. Pro tuto formu hodnocení jsem se rozhodl jednak proto, že pro klasifikaci a skórování hate speech není dostupný žádný model pro český jazyk, ani dostupná data, z kterých by bylo možné klasifikátor vyrobit. Analyzovaný datový soubor obsahuje pouze 800 případů, proto není obtížné takovýto počet případů klasifikovat ručně, čímž je zaručena vysoká kvalita výsledku.

Vzhledem k možnosti zaujatosti při hodnocení datového souboru byly vybráni čtyři hodnotitelé, kteří nezávisle na sobě hodnotili ručně hate speech a kategorizovali ji dle předem zvolených kritérií. Každý z hodnotitelů klasifikoval 200 náhodně vybraných komentářů, které byly vybrány tak, aby zastupovaly všechny sledované zpravodajské servery. Pro vytvoření kategorií a kritérií jsem se rozhodl zvolit přístup intenzity dopadu hate speech na oběť, který používají i Miro-Llinares a Rodrigez-Salsa (Miro-Llinares, Rodrigez-Salsa 2016). Kromě intenzity dopadu hate speech jsem do kategorií zakomponoval i to, zda je hate speech směřována vůči jednotlivci anebo vůči určité skupině, projevy proti jednotlivcům jsou pak hodnoceni jako méně závažné a činy proti skupině jako více závažné. Dále jsem se rozhodl zahrnout i projevy, které jsou spíše ofenzivního charakteru a většina definic by takové projevy neoznačila za hate speech. Vytvořená kategorizace hate speech na základě výše zmíněných charakteristik vypadá následovně:

- **První kategorie:** Komentář neobsahuje žádnou formu hate speech, ani jiné náznaky agresivního chování vůči jedinci či skupině.
- **Druhá kategorie:** Komentář obsahuje hanlivé výrazy, urážky, ponižování nebo obecnou výzvu k páchání násilí proti jednotlivé osobě, které takové jednání může způsobit psychickou újmu.
- **Třetí kategorie:** Komentář obsahuje hanlivé výrazy, urážky, ponižování nebo obecnou výzvu k páchání násilných činů proti skupině osob založené na základě osobnostních charakteristik (rasa, chování, fyzické dispozice, sexuální orientaci, třída, gender, etnicita, postižení, víra a politická příslušnost), které mohou způsobit psychickou újmu. Dále zde patří podněcování nebo vyhrožování násilím vůči jednotlivé osobě.
- **Čtvrtá kategorie:** Komentář obsahuje nenávistné projevy, pohrdání nebo diskriminování vůči určitým skupinám, speciálně vůči těm, které se mohou cítit ohroženy nebo se bojí omezení svých práv (Romové, Židé, transgender atd.). Dále sem patří vyvolávání pozitivního přístupu k násilí na lidech určité skupiny, které si následně pachatel ospravedlňuje. Podněcování k násilným činům proti příslušníkům určité skupiny.
- **Pátá kategorie:** Komentář obsahuje nesporné projevy nenávisti vedené především vůči lidem určité skupiny anebo svým obsahem zcela dehumanizují určitou skupinu. Takové to činy pak mohou způsobit skupině osob těžkou psychickou újmu a mohou se dotknout i nepřímo zúčastněných osob a skupin. Dále sem patří přímé vyhrožování násilím nebo smrtí určité skupině, nejčastěji na základě některé z nenávistných ideologií.

Celkem bylo hodnoceno 800 komentářů z facebookových stránek vybraných zpravodajských serverů z čehož bylo celkem 650 komentářů zařazeno do první kategorie, tedy neobsahovaly žádné náznaky hate speech. Patří sem převážně komentáře, které vyjadřují postoj pisatele ke sdílenému příspěvku nebo jej hodnotí, dále do této kategorie spadají komentáře, které nesouvisí s příspěvkem a zároveň nikoho neuráží. Do této kategorie byly zařazeny komentáře jako *"A daň ze vzduchu by to chtělo, a daň z alergie na řepku"* nebo například *"Vztah Prahy a Pekingu by měl být pragmatický, realistický a kritický."*

Komentářů, které obsahovaly jakoukoliv formu hate speech bylo 150, nejvíce případů spadá do druhé a třetí kategorie. Naopak žádný z komentářů neobsahoval velmi agresivní a dehumanizující hate speech, která by vyhrožovala násilím nebo smrtí určité skupině, proto nebyl žádný komentář zařazen do páté kategorie. Nejvíce případů hate speech bylo zařazeno do druhé kategorie, která zahrnuje pouze páchání nenávistných projevů vůči jedné osobě, celkem do této kategorie bylo zahrnuto 80 komentářů. Nejčastěji šlo o nadávky a útoky vůči lidem, kteří byli předmětem sdíleného zpravodajského příspěvku a časté byly také nenávistné projevy mezi jednotlivými diskutujícími. Do druhé kategorie byly zařazeny komentáře jako *"To máte odkud Pavle, Vy děcko nevychivané/spratku jeden."* nebo *"Myslím, že je na čase aby byla po Kremlíkovi odvolána (klidně dnes) i Maláčová. Ta žena neumí počítat. Na to ale Hamáček nemá koule."*

Do třetí kategorie bylo zařazeno celkem 63 komentářů, ty obsahovaly hate speech proti skupině osob anebo podněcovaly a vyhrožovaly násilím určité osobě. Jako hate speech byly zařazeny i komentáře, které sarkasticky útočily na určitou skupinu, která spadá do některé z chráněných charakteristik. Komentáře v třetí kategorii, které se vymezují proti skupině osob jsou například *"Ať tam mají i hákové kříže, když už mají hidžíb"* nebo *"Upozorňujeme místní buranské čecháčky, že uprchlíci jsou pro moderní společnost větším přínosem, než oni. Všechny protievropské a protiuprchlické komentáře budou nahlašovány a předány panu Hamáčkovi."* Dalšími příklady komentářů z této kategorie mohou být výhružná a násilná sdělení vůči jednotlivci jako *"Srandisto! A pro pos , nás je více my si vás najdeme , víme kde bydlíte 🤣🤣🤣🤣."*

Do čtvrté kategorie bylo zařazeno pouze 7 komentářů, které obsahovaly velmi nenávistnou a agresivní formu hate speech. Nejčastěji se jednalo o nenávistné komentáře, které se vymezovaly vůči uprchlíkům a zdůrazňovaly, že je potřeba se postarat nejprve o občany České republiky a časté byly také komentáře vymezující se agresivně vůči politické straně nebo

hnutí. Spadají zde komentáře jako *"Sou to magori ,já bych tuhle celou fetackou bandu zavřela na doživotí celou Piratskou stranu a jim podobný.* 😠😠😠😠😠*"* a nebo například *"Českým členům rodinám pořádně pomocí neumí ale cizincům který jsou ještě celý život jen učení k nebezpečnosti by pomáhat chtěli když oni se i ve 3 letech jsou schopný třeba odpálit a buch vi co všecko."*

Z celkového počtu 800 komentářů bylo 150 komentářů označeno za nenávistné, což tvoří 18,75 % ze všech komentářů. Četnost nenávistných komentářů ve zkoumaném datovém souboru je relativně vysoká, je to pravděpodobně tím, že komentáře byly sbírány pod příspěvky týkající se domácího a zahraničního zpravodajství a politiky, které mezi lidmi často vzbuzují emoce a objevují se zde často osobní názory, nebo rozepře mezi jednotlivými skupinami diskutujících. Podrobné analýze toho, kdo píše hate speech komentáře se věnuji v následující kapitole.

5. Analýza nenávistných projevů na sociálních sítích zpravodajských serverů

V předchozích kapitolách byly vysvětleny základy problematiky hate speech, jak lze hate speech hodnotit a jak byly konstruovány jednotlivé proměnné a sběr dat. Analýza nenávistných projevů bude provedena na datovém souboru 800 komentářů z facebookových stránek zpravodajských serverů Idnes.cz, Aktuálně.cz, Novinky.cz, ParlamentníListy.cz a ČT24. Tato kapitola se za pomoci kvantitativní analýzy snaží zjistit, které sledované faktory vysvětlují přítomnost hate speech v komentářích. Závisle proměnnou je zde míra hate speech v komentáři, která byla hodnocena na škále 1 až 5, jednotlivá hodnotící kritéria jsou podrobně vysvětlena v kapitole 3.3. Nezávislé proměnné vstupující do analýzy jsou pohlaví, vzdělání a bydliště autora komentáře, dále pak zpravodajský server a vulgárnost komentáře.

První část této kapitoly se věnuje četnosti hate speech ve zkoumaném datovém souboru a porovnání s četností hate speech v obdobných studiích. Další podkapitoly se pak věnují jednotlivým nezávisle proměnným a jejich vlivu na hate speech, poslední část kapitoly pak shrnuje výsledky analyzování hate speech v komentářích na facebookových stránkách zpravodajských serverů.

5.1. Výskyt nenávistných projevů na sledovaných facebookových stránkách zpravodajských serverů

Analyzovaný datový soubor facebookových komentářů ze stránek zpravodajských serverů má celkem 800 případů z kterých bylo 150 případů označeno za komentáře obsahující hate speech, což je 18,75 % z celkového počtu komentářů. Pro představu o tom, zda se jedná o vysoké zastoupení hate speech v datovém souboru jsem se rozhodl porovnat toto procento s podobnými studiemi. Porovnání míry hate speech mezi jednotlivými studiemi je ale obtížné, každá se zaměřuje na jiný typ příspěvků z jiného časového období a události a používá jiná kritéria pro hodnocení hate speech. Je proto potřeba brát tato srovnání pouze jako orientační. Miró-Llinares a Rodriguez-sala (2016) ve své studii analyzovali příspěvky na Twitteru týkající se události útoku na francouzský satirický časopis Charlie Hebdo, kde z celkového počtu 282 397 tweetů bylo označeno za nenávistné pouze 2 304, které tvořily přibližně 0,8 % ze všech tweetů. Hate speech u jednotlivých tweetů byla hodnocena na vybraných tweetech týmem hodnotitelů z jejichž hodnocení se následně vytvořily pravidla pro automatickou klasifikaci hate speech. Burnap a Williams (2015) ve své studii Kybernetické Hate Speech na Twitteru označili jako nenávistné projevy přibližně 11 % tweetů z celkového počtu 450 000 zkoumaných tweetů.

Davidson et al. (2017) vybrali ve své studii z celkem 85,4 milionů tweetů náhodně 25 000 tweetů, z kterých bylo 24 802 ručně kódovaných alespoň třemi pracovníky CrowdFlower. V tomto ručně kódovaném datovém souboru bylo 5 % označeno za tweety, které obsahují hate speech. I mezi jednotlivými studiemi hate speech na sociální síti Twitter jsou značné rozdíly v procentech příspěvků označených jako nenávistné, a to v rozmezí od 0,8 % do 11 %. Studií věnující se nenávistnému obsahu na sociální síti Facebook je podstatně méně, podařilo se mi nalézt pouze studii autorů Hrdina, Daňková, Kopecká (2016), která se věnuje analýze hate speech na českém Facebooku, ale pouze na komentářích, které obsahovaly klíčová slova migrace, islám a Afrika. Ze studie vyplynulo, že u analyzovaných komentářů byla přítomnost hate speech až 80 %. Jedná se o poměrně vysoké procento hate speech v komentářích, které je do jisté míry způsobeno výběrem klíčových slov a témat s nimi spjatých, ty v české společnost velmi silně rezonují a část společnosti je vnímá velmi kriticky.

V porovnání s výše uvedenými studiemi je 18,75 % hate speech případů ve zkoumaném datovém souboru poměrně velké procento, pouze studie Facebook komentářů na téma islám, migrace a Afrika zaznamenala vyšší procento hate speech případů. Procento hate speech u ostatních studií prováděných na Twitteru se pohyboval v rozmezí od 0,8 % do 11 %. Takto vysoký rozdíl v zastoupení hate speech může být jednak způsoben jinými platformami, způsobem hodnocení, a především výběrem případů. Ve zkoumaném datovém souboru byly vybírány komentáře pod zprávami, které se týkají především politických témat, které často mezi lidmi vytváří rozepře, ve kterých lidé častěji mohou použít nenávistné výrazy. Také zvolená kritéria hate speech jsou nastavena poměrně přísně, především proto, že zahrnují i nenávistné projevy vůči jednotlivci, které část definic nepovažuje za hate speech.

Datový soubor obsahuje 800 komentářů s rozdílnou intenzitou nenávistného projevu, proto bylo zvoleno 5 kategorií, které indikují míru hate speech v komentáři. Do první kategorie byly zařazeno celkem 650 komentářů, které neobsahovaly žádný druh hate speech. Druhá kategorie zahrnuje hate speech proti jednotlivci, tento typ hate speech byl v komentářích zaznamenán nejčastěji, a to celkem u 80. případů. Ve třetí kategorii jsou případy hate speech proti skupině osob na základě některé z chráněných charakteristik, takovýchto komentářů bylo 63. Čtvrtá kategorie zahrnuje případy agresivní hate speech, které mohou oběti způsobit závažnou újmu, pouze 7 komentářů bylo zařazeno do této kategorie. Do páté kategorie, která zahrnuje dehumanizující hate speech a výhružky smrtí skupině osob nebyl zařazen žádný z

komentářů. Přehled procentuálního zastoupení jednotlivých kategorií je přehledně v tabulce 5.1.

Tabulka 5.1: Přehled kategorizace hate speech

Kategorie hate speech	**Počet případů**	**Procento**
Neobsahuje hate speech	650	81,3 %
Hate speech vůči jednotlivci	80	10 %
Hate speech vůči skupině	63	7,9 %
Agresivní hate speech vůči skupině	7	0,9 %
Velmi agresivní hate speech a výhružky smrtí skupině	0	0 %
Celkem	800	100 %

Zdroj: vlastní

Ve zkoumaném datovém souboru bylo zaznamenáno podstatně větší procento hate speech než v podobných studiích, to je pravděpodobně způsobeno mnoha faktory, jakými jsou například zvolená kritéria pro klasifikaci hate speech, nebo typ sledovaných facebookových stránek. Ač bylo zaznamenáno větší procento případů hate speech, nejednalo se, až na výjimky, o velmi agresivní hate speech. Většina případů hate speech spadala do druhé a třetí kategorie, opravdu agresivních hate speech komentářů bylo pouze 7 a žádný z komentářů nevyhrožoval skupině osob násilím nebo smrtí, nebo jiným způsobem dehumanizoval některou ze skupin na základě chráněných charakteristik.

5.2. Vliv pohlaví na přítomnost hate speech v komentářích

Tato část kapitoly se zabývá pouze vlivem pohlaví na přítomnost hate speech v komentářích pod příspěvky vybraných zpravodajských serverů. Proměnná pohlaví je v této analýze rozdělena to tří kategorií a to muž, žena anebo pohlaví nelze určit. Pro analyzování vlivu pohlaví na přítomnost hate speech jsem zvolil kontingenční tabulku a pro určení síly vztahu jsem zvolil koeficient Cramerovo V.

Největší zastoupení nenávistných komentářů je u mužů, to je ale zapříčiněno tím, že muži celkově častěji komentovali příspěvky na sociální síti Facebook. Pokud se ale podíváme na procentuální zastoupení hate speech, tak největší procento hate speech vykazuje kategorie pohlaví nelze určit a to celkem 30 % všech komentářů zařazených do této kategorie obsahuje

hate speech. U mužů je to 19 % a u žen pouze 17 % komentářů napsaných ženským profilem obsahovalo hate speech.

Tabulka 5.2: Kontingenční tabulka hate speech – pohlaví

		Nelze určit	Muži	Ženy	Celkem
Neobsahuje hate speech	Počet	7	506	137	650
	Očekávaný počet	8,1	507,8	134,1	650
	% Pohlaví	70,00 %	81,00 %	83,00 %	81,30 %
Hate speech vůči osobě	Počet	1	67	12	80
	Očekávaný počet	1	62,5	16,5	80
	% Pohlaví	10,00 %	10,70 %	7,30 %	10,00 %
Hate speech vůči skupině	Počet	1	50	12	63
	Očekávaný počet	0,8	49,2	13	63
	% Pohlaví	10,00 %	8,00 %	7,30 %	7,90 %
Agresivní hate speech vůči skupině	Počet	1	2	4	7
	Očekávaný počet	0,1	5,5	1,4	7
	% Pohlaví	10,00 %	0,30 %	2,40 %	0,90 %
Celkem	Počet	10	625	165	800
	% Pohlaví	100,00 %	100,00 %	100,00 %	100,00 %

Zdroj: vlastní

V kategorii hate speech vůči osobě měli muži vyšší zastoupení o 0,7 % oproti očekávanému zastoupení, naopak ženy v této kategorii měli méně hate speech, než se očekávalo a to o 2,7 %. Zastoupení kategorie pohlaví nelze určit bylo stejné, jako očekávané zastoupení, tedy 10 %. U kategorie hate speech vůči skupině je výrazně vyšší zastoupení pouze u kategorie pohlaví nelze určit, které ale může být způsobeno poměrně malým vzorkem případů v kategorii pohlaví nelze určit. Mírně vyšší počet hate speech komentářů oproti očekávanému počtu zde mají muži, a to pouze o 0,1 %. U žen je zde zastoupení nižší a to o 0,6 % oproti očekávanému zastoupení. V kategorii agresivní hate speech vůči skupině mají nadprůměrně vyšší zastoupení kategorie pohlaví nelze určit a ženy, které zde mají o 1,5 % vyšší zastoupení, než je očekávaná hodnota. Muži zde naopak mají o 0,6 % nižší zastoupení. Je ale nutno zmínit, že v této hate speech kategorii bylo jen 7 případů, proto výsledky v této skupině jsou spíše náhodné, než že by ukazovaly vliv pohlaví na tuto kategorii hate speech. Koeficient Cramerovo

V vyšel pro data v kontingenční tabulce 0,107, což značí střední asociaci mezi proměnnými, signifikance u Cramerova V vyšla 0,006. Pro inferenční statistiku je nutné zvolit neparametrický test, jelikož zkoumané proměnné nemají normální rozdělení. Pro zkoumaná data jsem zvolil Kruskal-Wallis test, který je alternativou one-way ANOVY pro neparametrická data. Nulová hypotéza říká, že jednotlivé kategorie pohlaví mají shodnou distribuci hate speech. Signifikance neboli hodnota p vyšla pro zkoumaná data 0,521, díky čemuž se nulová hypotéza přijímá. Na úrovni spolehlivosti 0,05 (95 %) nebyl prokázán statisticky významný rozdíl mezi distribucí hate speech u jednotlivých kategorií pohlaví.

Z provedené analýzy na datech facebookových komentářů je patrné, že nejčastěji píší hate speech účty, které nemají uvedené pohlaví. Jedná se o účty, které komentují jako facebooková stránka anebo osobní účty, u kterých dle jména ani fotky nelze určit pohlaví. O 2 % častěji byl autorem nenávistného komentáře muž než žena, nejvíce je rozdíl znát na kategorii hate speech vůči osobě. Výjimkou je kategorie agresivní hate speech vůči skupině, kde mnohem častěji byly zastoupeny ženy, tato kategorie ale obsahuje jen málo případů hate speech, pro ověření tohoto výsledku by bylo potřeba robustnější soubor facebookových komentářů.

5.3. Vliv úrovně vzdělání na přítomnost hate speech v komentářích

V této části budu analyzovat vliv uváděného vstupně vzdělání na facebookových profilech na přítomnost hate speech v komentářích. Proměnná vzdělání byla zaznamenána celkem u 225 případů a je rozdělena do čtyř kategorií na základní školy a fiktivní školy, střední odborná učiliště, střední školy a vyšší odborné školy a vysoké školy. Pro analýzu vlivu vzdělání na hate speech v komentářích jsem zvolil kontingenční tabulku a pro ověření síly vztahu mezi proměnnými koeficient Kendallovo tau-b. Nejčastěji měli uživatelé na svém profilu uvedené vzdělání na VOŠ a VŠ a to celkem 105 případů, 79 případů mělo uvedeno SŠ vzdělání, pouze 28 uživatelů mělo uvedeno vzdělání na SOU a vzdělání na ZŠ nebo fiktivní školu uvedlo jen 13 uživatelů. Mezi jednotlivými kategoriemi je velký nepoměr v počtu případů, proto budu v analýze pracovat s jednotlivými kategoriemi převážně v procentech.

Procento nenávistných projevů se u každé ze zkoumaných kategorií vzdělání výrazně liší. Největší procento bylo zaznamenáno u kategorie ZŠ a fiktivní školy, kde 30,8 % komentářů obsahovalo hate speech. U uživatelů uvádějících jako vzdělání SOU je procento hate speech komentářů 25 %, což je podstatně méně než u kategorie ZŠ a fiktivní školy.

Uživatelé, kteří uvedli jako vzdělání střední školu napsali komentář obsahující hate speech pouze v 17,7 % případů a uživatelé s uvedenou VOŠ a VŠ psali nenávistné komentáře jen v 14,3 % případů. Lze zde tedy pozorovat, že se stoupajícím vzděláním klesá procento hate speech komentářů, nejvíce patrný rozdíl je pak mezi SOU a SŠ kde je rozdíl mezi těmito dvěma kategoriemi 7,3 %. Pro detailnější pochopení vlivu vzdělání na hate speech je potřeba se ale podívat, jak jsou v jednotlivých kategoriích hate speech zastoupeny dané stupně vzdělání.

Tabulka 5.3: Kontingenční tabulka hate speech – vzdělání

		ZŠ a Fiktivní školy	**SOU**	**SŠ**	**VOŠ a VŠ**	**Celkem**
Neobsahuje hate speech	Počet	9	21	65	90	185
	Očekávaný počet	10,7	23	65	86,3	185
	% Vzdělání	69,20 %	75,00 %	82,30 %	85,70 %	82,20 %
Hate speech vůči osobě	Počet	1	4	10	8	23
	Očekávaný počet	1,3	2,9	8,1	10,7	23
	% Vzdělání	7,70 %	14,30 %	12,70 %	7,60 %	10,20 %
Hate speech vůči skupině	Počet	3	2	4	7	16
	Očekávaný počet	0,9	2	5,6	7,5	16
	% Vzdělání	23,10 %	7,10 %	5,10 %	6,70 %	7,10 %
Agresivní hate speech vůči skupině	Počet	0	1	0	0	1
	Očekávaný počet	0,1	0,1	0,4	0,5	1
	% Vzdělání	0,00 %	3,60 %	0,00 %	0,00 %	0,40 %
Celkem	Počet	13	28	79	105	225
	% Vzdělání	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %

Zdroj: vlastní

V kategorii hate speech vůči osobě má SOU a SŠ vyšší procento hate speech komentářů, než je očekávaná hodnota a to o 3,1 % respektive o 2,5 %. Naopak v této kategorii mají menší procento hate speech komentářů uživatelé s VOŠ a VŠ a uživatelé kteří spadají do kategorie ZŠ a fiktivní školy. U kategorie VOŠ a VŠ bylo procento hate speech komentářů o 2,6 % nižší, než bylo očekávané procento, kategorie ZŠ a fiktivní školy měla zastoupení hate speech komentářů v této kategorii o 2,5 % nižší oproti očekávaným hodnotám. V kategorii hate speech

vůči skupině má největší procento hate speech kategorie ZŠ a fiktivní školy, která měla o 16 % vyšší zastoupení oproti očekávanému zastoupení. Naopak nejnižší zastoupení zde měla kategorie SŠ následovaná kategorií VOŠ a VŠ, zde byl rozdíl oproti očekávanému zastoupení o 2 % respektive o 0,4 % nižší. V kategorii agresivní hate speech vůči skupině se v analýze nacházel pouze jediný případ, proto v této analýze tuto kategorii nebude nijak zohledňovat. Koeficient Kendallovo tau-b stejně jako koeficient Gamma vyšel pro data v kontingenční tabulce -0,104 což značí nízkou zápornou asociaci mezi analyzovanými proměnnými. Tedy čím vyšší je vzdělání tím nižší je hate speech. Pro alespoň střední zápornou asociaci mezi proměnnými by byla potřeba hodnota okolo -0,2, hodnota signifikance pro Kendallovo tau-b vyšla 0,114. Pro inferenční statistiku jsem zvolil neparametrický test Kruskal-Wallis test, který ověřuje nulovou hypotézu. Ta říká, že distribuce hate speech mezi kategoriemi vzdělání je rovná. Signifikance neboli hodnota p vyšla u Kruskal-Wallis test pro zkoumaná data 0,286, díky čemuž se nulová hypotéza přijímá, a tedy na úrovni spolehlivosti 0,05 (95 %) nebyl zjištěn statisticky významný rozdíl mezi distribucí hate speech u jednotlivých kategorií vzdělání. Vyšla zde ale nižší hodnota p než u pohlaví, lze tedy předpokládat, že by přece jen mohly být mezi jednotlivými kategoriemi vzdělání větší rozdíly v distribuci dat, než tomu je u kategorií pohlaví.

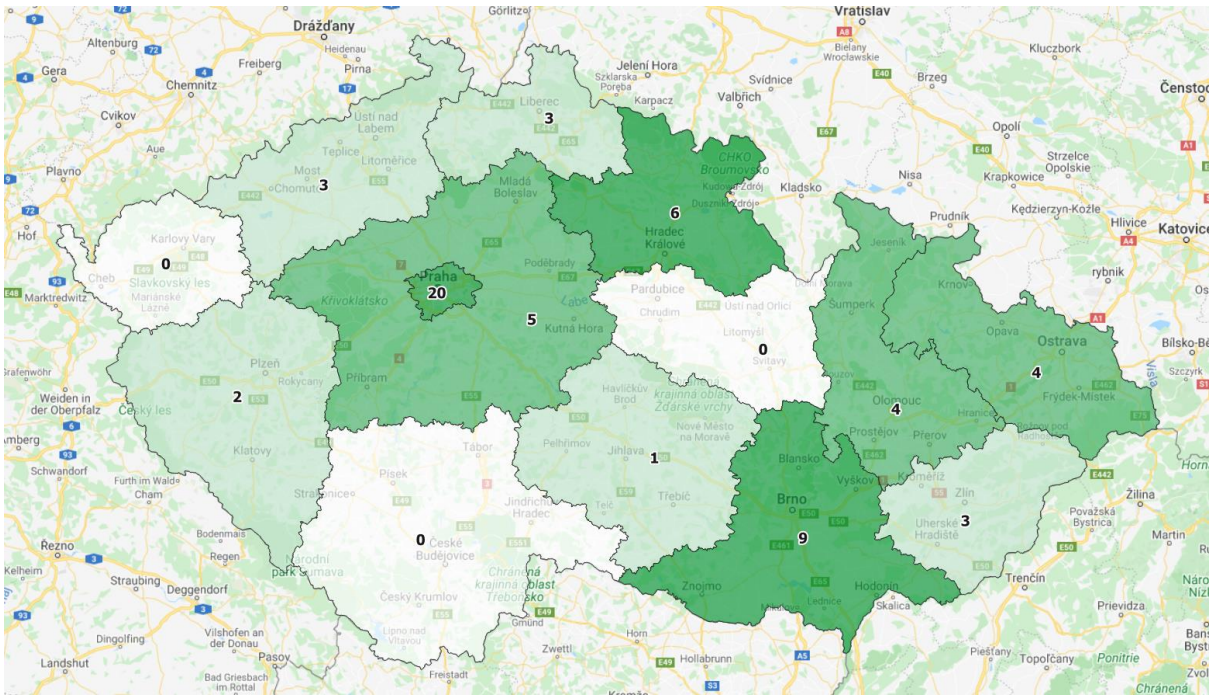
Z analýzy vlivu uvedeného vzdělání na přítomnost hate speech v komentáři vyplývá, že se zvyšujícím se uvedeným vzděláním klesá přítomnost nenávistných projevů v komentářích. Tento fakt se potvrdil jak na úrovni procentuálního zastoupení hate speech v jednotlivých kategoriích, tak i na jednotlivých kategoriích hate speech. Bohužel se ale nepotvrdila žádná vyšší asociace mezi zkoumanými proměnnými.

5.4. Vliv místa bydliště na přítomnost hate speech v komentářích

Tato podkapitola se věnuje analýze vlivu místa bydliště na přítomnost hate speech v komentářích. V analyzovaném datovém souboru se nachází 401 případů, které na svém profilu uvedly místo bydliště. Profily na sociální síti Facebook, které uvedly bydliště a zároveň obsahovaly jeden ze stupňů hate speech bylo 67, ty jsem rozdělil dle krajů v České republice a hodnoty vložil do mapy na obrázku 5.1 Pokud se podíváme na absolutní počty nenávistných komentářů jednoznačně nejvíce má Praha (20 hate speech komentářů) následována Jihomoravským (9 hate speech komentářů) a Královéhradeckým krajem (6 hate speech komentářů). Na mapě není zobrazen počet nenávistných komentářů profilů, které jako místo

bydliště měli uvedeno zahraničí, těchto případů bylo 7. Naopak komentující z Karlovarského, Jihočeského a Pardubického kraje nenapsali žádný komentář, který by obsahoval hate speech.

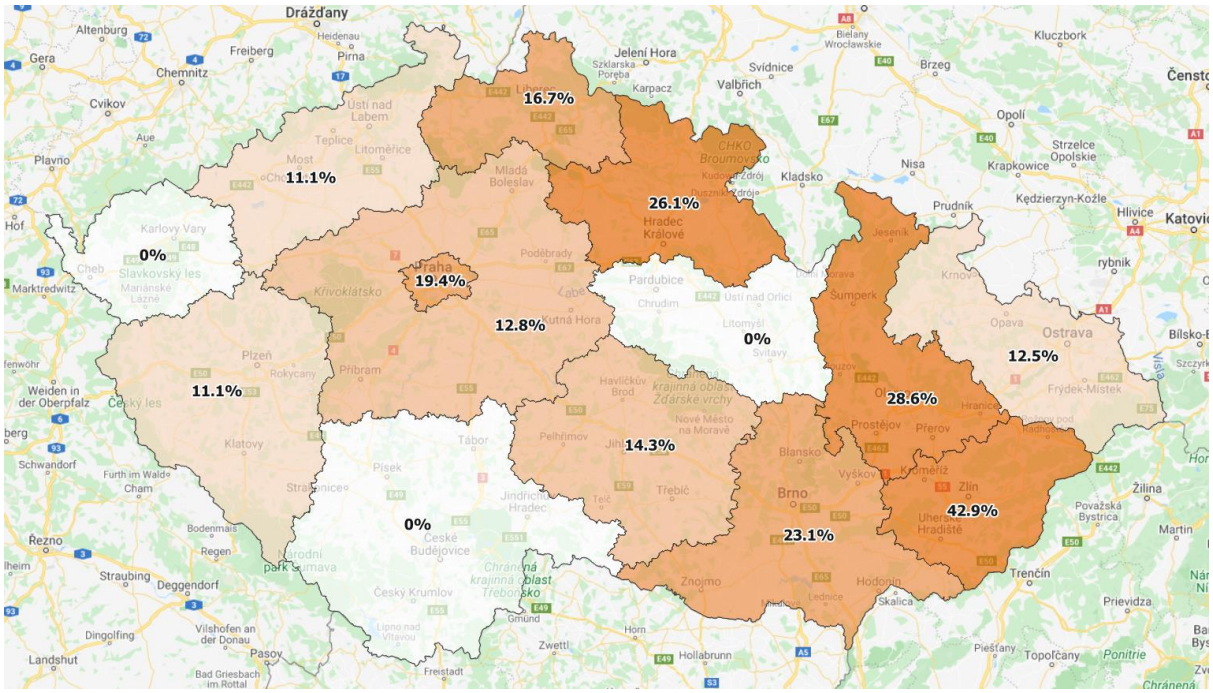
Obrázek 5.1: Mapa počtu hate speech komentářů v krajích ČR



Zdroj: vlastní (zpracováno v QGIS)

Absolutní počet hate speech případů nám ale nevysvětlí vztah mezi místem bydliště a hate speech, jelikož komentáře z jednotlivých krajů nejsou proporčně zastoupeny. Proto jsem zjistil procentuální zastoupení hate speech komentářů z celkového počtu komentářů v jednotlivých krajích a zanesl je do mapy na obrázku 4.2. Největší procento hate speech komentářů má Zlínský kraj, kde 42,9 % komentářů obsahuje hate speech. Poměrně vysoké procento hate speech komentářů má také Olomoucký a Královehradecký kraj a to 28,6 % respektive 26,1 %. Vyšší procento hate speech komentářů bylo zaznamenáno také u Jihomoravského kraje, Prahy a Libereckého kraje. Bylo by vhodné tyto výsledky ověřit na souboru komentářů, který obsahuje robustnější záznamy o místě bydliště jednotlivých uživatelů, jelikož v některých krajích bylo zaznamenáno pouze 7 komentářů, což je velmi malý vzorek na určení vztahu mezi místem bydliště autora komentáře a přítomností hate speech v komentáři. Výsledky této analýzy jsou tedy spíše orientační a mohou sloužit jako podklad pro další výzkum, nelze z nich ale vyvodit žádný jednoznačný vliv místa bydliště na přítomnost hate speech v komentářích.

Obrázek 5.2: Mapa procentuálního zastoupení hate speech komentářů v krajích ČR



Zdroj: vlastní (zpracováno v QGIS)

5.5. Vliv zpravodajského serveru na přítomnost hate speech v komentářích

V následující části analýzy se věnuji vlivu vybraných facebookových stránek zpravodajských serverů na přítomnost hate speech v komentářích. Vybrány byly stránky zpravodajských serverů iDnes, Novinky.cz, Aktuálně, ČT24 a Parlamentní listy, způsob výběru zpravodajských serverů je popsán v kapitole 3.1. Pro analyzování vlivu zpravodajských serverů jsem zvolil kontingenční tabulky a sílu vztahu mezi proměnnými otestuji za pomoci koeficientů Phi a Cramerovo V. Dále se kapitola věnuje analýze vlivu typu zpravodajského příspěvku na hate speech příspěvky, pod kterými byly sbírány komentáře jsou rozděleny na kategorie domácí zpravodajství a zahraniční zpravodajství.

V datovém souboru nejsou komentáře z jednotlivých zpravodajských serverů zastoupeny rovnoměrně, to především proto, že facebookové stránky zpravodajských serverů se liší v intenzitě sdílených zpráv a počtu komentářů pod jednotlivými příspěvky. Z toho důvodu se zaměřuji pouze na procentuální zastoupení jednotlivých kategorií hate speech u zpravodajských serverů a absolutní počty případů nejsou v této analýze zohledněny. Pokud se podíváme pouze na procentuální zastoupení hate speech v komentářích v tabulce 5.4 a nebudeme brát v potaz jednotlivé kategorie hate speech tak z dat vyplývá, že nejvíce hate

speech v komentářích je na facebookové stránce Parlamentních listů, kde 23,6 % komentářů obsahovalo hate speech, poměrně hodně hate speech komentářů s hate speech bylo i na facebookové stránce serveru Novinky a to 21,7 %. Menší procento hate speech měli pak servery iDnes a Aktuálně, kde komentáře obsahovaly hate speech pouze v 17,1 % respektive 18 %. Nejmenší procento hate speech obsahovaly komentáře na facebookové stránce zpravodajství ČT24, kde bylo pouze 14,6 % komentářů obsahující hate speech.

Tabulka 5.4: Kontingenční tabulka hate speech – zpravodajský server

		ČT24	**Aktuálně**	**iDnes**	**Novinky**	**Parlamentní listy**	**Celkem**
Neobsahuje hate speech	Počet	152	132	136	123	107	650
	Očekávaný počet	144,6	130,8	133,3	127,6	113,8	650
	% Zpravodajství	85,40 %	82,00 %	82,90 %	78,30 %	76,40 %	81,30 %
Hate speech vůči osobě	Počet	11	21	19	13	16	80
	Očekávaný počet	17,8	16,1	16,4	15,7	14	80
	% Zpravodajství	6,20 %	13,00 %	11,60 %	8,30 %	11,40 %	10,00 %
Hate speech vůči skupině	Počet	11	8	9	21	14	63
	Očekávaný počet	14	12,7	12,9	12,4	11	63
	% Zpravodajství	6,20 %	5,00 %	5,50 %	13,40 %	10,00 %	7,90 %
Agresivní hate speech vůči skupině	Počet	4	0	0	0	3	7
	Očekávaný počet	1,6	1,4	1,4	1,4	1,2	7
	% Zpravodajství	2,20 %	0,00 %	0,00 %	0,00 %	2,10 %	0,90 %
	Počet	178	161	164	157	140	800
	% Zpravodajství	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %

Zdroj: vlastní

Intenzita hate speech se také liší u jednotlivých facebookových stránek zpravodajských serverů. V kategorii hate speech vůči osobě měly hned tři zpravodajské servery vyšší procento případů, než byla očekávaná hodnota jsou jimi Aktuálně, iDnes a Parlamentní listy. Nejvyšší procento komentářů v této kategorii má server Aktuálně a to o 3 % více než je očekávaná

hodnota, naopak ze sledovaných zpravodajských serverů zde má nejmenší procento ČT24, a to o 3,8 % méně oproti očekávané hodnotě. V kategorii hate speech vůči skupině měli vyšší procentuální zastoupení pouze servery Novinky a Parlamentní listy a to o 5,5 % respektive 2,1 % oproti očekávaným hodnotám. Výrazně menší procento v této hate speech kategorii měli servery Aktuálně a iDnes a to o 2,9 % respektive 2,4 % méně oproti očekávané hodnotě. Kategorie Agresivní hate speech vůči skupině měla pouze 7 případů z toho 4 případy byly zaznamenány na facebookové stránce ČT24 a 3 případy na stránce Parlamentních listů. V případě ČT24 se jednalo o 1,3 % více hate speech komentářů v této kategorii, než byla očekávaná hodnota a u Parlamentních listů to bylo o 1,2 % více.

Koeficient Phi vyšel pro provedenou analýzu 0,185 a Cramerovo V 0,107 což značí střední až silnou asociaci mezi zkoumanými proměnnými. Hodnota signifikance pro Phi i Cramerovo V vyšla 0,007. Dále jsem data testoval neparametrickým Kruskal-Wallis testem, který ověřil nulovou hypotézu. Ta říká, že distribuce hate speech u vybraných zpravodajských serverů je rovná. Hodnota p vyšla pro zkoumaná data 0,209, díky čemuž se nulová hypotéza přijímá. Na úrovni spolehlivosti 0,05 (95 %) tedy nebyl zjištěn statisticky významný rozdíl mezi distribucí dat u vybraných zpravodajských serverů, případný rozdíl v distribuci hate speech může být způsoben chybou výběru.

Z provedené analýzy je patrné, že více hate speech se objevilo na facebookových stránkách zpravodajských serverů Parlamentní listy a Novinky. U serveru novinky bylo výrazně vyšší zastoupení hate speech pouze ve třetí kategorii hate speech vůči osobě, naopak Parlamentní listy měly relativně vysoké zastoupení ve všech kategoriích hate speech. U serverů Aktuálně a iDnes byl zaznamenaný menší počet nenávistných komentářů, které navíc v drtivé většině spadaly do druhé kategorie hate speech vůči osobě. Nejméně nenávistných komentářů bylo na facebookových stránkách zpravodajského serveru ČT24, kde pouze 14,6 % komentářů obsahovalo v hate speech a hodnoty v jednotlivých kategoriích hate speech byly až na vyjímku ve čtvrté kategorii agresivní hate speech vůči skupině na velmi nízké úrovni.

5.6. Vliv vulgarit v komentáři na přítomnost hate speech

Část komentářů v analyzovaném datovém souboru obsahuje vulgární výrazy. Z celkového počtu 800 případů obsahuje komentář vulgarismy pouze v 45 případech, což tvoří 5,6 % z celkového počtu komentářů. Analýzu vlivu vulgarit v komentáři na hate speech provedu za pomoci kontingenční tabulky a koeficientu Phi a Cramerovo V. Z kontingenční tabulky 5.5

vyplývá, že případy v kategorii vulgární komentář v 42,2 % případů obsahovaly také hate speech. Naopak komentáře v kategorii komentář bez vulgarit, obsahovaly hate speech jen v 17,4 % případů. Pokud se zaměříme na jednotlivé kategorie hate speech tak u kategorie hate speech vůči osobě bylo z celkového počtu vulgárních komentářů 28,9 % což je o 18,9 % více oproti očekávané hodnotě, naopak u komentáře bez vulgarity byla hodnota o 1,1 % nižší oproti hodnotě očekávané. U kategorie hate speech vůči skupině bylo zastoupení komentářů bez vulgarit pouze 7,7 % naopak komentáře obsahující vulgární výrazy měly v této kategorii 11,1 %, což je o 3,2 % více než byla očekávaná hodnota. Kategorie agresivní hate speech vůči skupině obsahovala jen 7 případů, ale i zde je patrný trend, že vulgární komentáře častěji obsahují nenávistné projevy. Vulgární komentáře v této kategorii měli o 1,3 % vyšší zastoupení, než bylo očekáváno a komentáře bez vulgarit měli o 0,1 % nižší zastoupení oproti očekávanému zastoupení v této kategorii.

Tabulka 5.5: Kontingenční tabulka hate speech – vulgarita komentáře

		Komentář bez vulgarismů	**Vulgární komentář**	**Celkem**
Neobsahuje hate speech	Počet	624	26	650
	Očekávaný počet	613,4	36,6	650
	% Komentářů	82,60 %	57,80 %	81,30 %
Hate speech vůči osobě	Počet	67	13	80
	Očekávaný počet	75,5	4,5	80
	% Komentářů	8,90 %	28,90 %	10,00 %
Hate speech vůči skupině	Počet	58	5	63
	Očekávaný počet	59,5	3,5	63
	% Komentářů	7,70 %	11,10 %	7,90 %
Agresivní hate speech vůči skupině	Počet	6	1	7
	Očekávaný počet	6,6	0,4	7
	% Komentářů	0,80 %	2,20 %	0,90 %
Celkem	Počet	755	45	800
	% Komentářů	100,00 %	100,00 %	100,00 %

Zdroj: vlastní

Koeficient Phi a Cramerovo V vyšel na zkoumaných datech shodně 0,165, tato hodnota značí silný vztah mezi proměnnými hate speech a vulgarita komentáře. Signifikance u této

analýzy vyšla méně než 0,01. Vhodný test pro zkoumání inferenční statistiky u vulgárnosti komentářů je Mann-Whitney U test, který ověřuje nulovou hypotézu, tedy že distribuce hate speech je u kategorií komentář bez vulgarismů a vulgární komentář identická. Pro ověření nulové hypotézy je u Mann-Whitney U testu důležitá hodnota p, ta vyšla méně než 0,001. Na úrovni spolehlivosti 0,05 (95 %) je rozdíl mezi zkoumanými průměry jednotlivých kategorií statisticky významný a rozdíl mezi kategoriemi není způsoben náhodným výběrem vzorků.

Z provedené analýzy vyplývá, že je zde jasný vliv přítomnosti vulgárních výrazů v komentáři na to, zda bude komentář obsahovat hate speech. To se potvrdilo jednak vyšším procentuálním zastoupení hate speech komentářů v kategorii vulgárních komentářů a také při detailnější zkoumání vlivu v rámci jednotlivých kategorií hate speech, kde v každé z hate speech kategorií měla kategorie vulgárních komentářů větší procentuální zastoupení oproti kategorie komentář bez vulgarismů.

5.7. Shrnutí analytické kapitoly

Kapitola se věnovala analyzování hate speech v komentářích na facebookových stránkách zpravodajských serverů a vlivu vybraných proměnných na přítomnost hate speech v komentáři. Analýza přináší unikátní závěry především co se týče analyzování autorů facebookových komentářů a zjištění jistých charakteristik pro osoby častěji píšící hate speech v komentářích. Analýza byla provedena na náhodně vybraném vzorku 800 komentářů pod vybranými zpravodajskými servery, jedná se o unikátní vzorek především kvůli vysokému počtu sledovaných proměnných, avšak analýza byla limitována poměrně malým počtem případů, proto by bylo vhodné do budoucna analýzu zopakovat na větším vzorku komentářů.

Nejprve byla zkoumána samotná přítomnost komentářů obsahující hate speech v komentářové sekci pod zpravodajskými příspěvky na Facebooku. Zde se ukázalo, že vybrané případy komentářů obsahují výrazně vyšší procento hate speech oproti podobným studiím hate speech. To je do jisté míry dáno zvolenou škálou pro měření hate speech a také velmi konfliktním prostředím politického a zahraničního zpravodajství, ze kterého převážná část případů pocházela.

Druhou zkoumanou problematikou byl vliv vybraných proměnných na přítomnost hate speech v komentářích. U analýzy vlivu pohlaví bylo zjištěno, že nejvíce hate speech komentářů píši takové účty, u kterých nelze určit pohlaví (stránky, profily bez fotek a s divnými jmény). Pokud se podíváme na porovnání nenávistných projevů v komentářích u mužů a žen, tak z toho

o něco lépe vycházejí ženy, které měly o 2 % nižší přítomnost hate speech. Pokud se podíváme na vliv vzdělání je zde poměrně jasný směr a to, že s vyšším vzděláním klesá procento nenávistných komentářů. To se potvrdilo jak na procentuálním zastoupení hate speech u jednotlivých stupňů vzdělání, tak i u jednotlivých hate speech kategorií. U vlivu bydliště na přítomnost hate speech v komentářích se nepodařilo najít žádný vztah, to bylo způsobeno především vysokým počtem kategorií (14 krajů + zahraničí) a poměrně malým počtem případů, které měly uvedeno bydliště a zároveň psali nenávistné projevy.

Další analyzovanou proměnnou byl typ zpravodajského serveru, zde bylo zjištěno že komentáře pod příspěvky veřejnoprávního zpravodajského serveru ČT24 obsahovaly nejméně hate speech. Naopak komentáře pod facebookovými příspěvky zpravodajského serveru Parlamentní listy, který je v České republice považován za jeden z webů, které šíří tzv. fake news, obsahovaly největší procento hate speech. Při analyzování proměnné přítomnost vulgarit se prokázal jasný vliv na přítomnost hate speech, tedy pokud komentář obsahuje vulgární výrazy, obsahuje mnohem častěji hate speech než komentář, který neobsahuje vulgarismy.

V analýze jsem se věnoval pouze nejdůležitějším proměnným z datového souboru, do budoucna je možné dále analyzovat proměnné typ zpravodajského příspěvku, přítomnost emoji, délku komentáře a počet reakcí na komentář, fotografie a podpora politické strany, které byly sesbírány, ale z důvodu rozsahu diplomové práce a u některých případů malého počtu případů nebyly do analýzy zahrnuty.

Závěr

Cílem diplomové práce bylo analyzovat komentáře na sociální síti Facebook pod příspěvky týkající se politiky, zpravodajství a zahraničních událostí. Pro analýzu byla vybrána pětice zpravodajských serverů, a to iDnes, Aktuálně, Novinky, ČT24 a Parlamentní listy. Diplomová práce měla stanoveny dva cíle. Prvním cílem bylo zjistit, zda se nachází pod příspěvky zpravodajských serverů na Facebooku komentáře obsahující hate speech a pokud ano, tak v jaké míře v porovnání s ostatními studiemi. Druhým cílem bylo zjistit základní charakteristiky hate speech komentářů a jejich autorů.

Pro určení míry hate speech existuje množství kategorizací, v práci jsem se rozhodl pro určení intenzity hate speech v komentářích použít vlastní škálu, která vychází ze škály autorů Miro-Llinares a Rodrigez-Salsa (2016), kteří určují míru hate speech na základě dopadu na oběť a doplnil o dimenzi, která určuje, zda je cílem hate speech jednotlivec nebo skupina osob, která sdílí některou z chráněných charakteristik.

Na základě dosavadního výzkumu a zaznamenaných dat bylo stanoveno několik hypotéz, které mohou v rámci problematiky hate speech komentářů na českých sociálních sítích vysvětlit samotnou přítomnost hate speech a prozradit základní charakteristiky autorů nenávistných komentářů. První hypotéza zkoumá, zda *zastoupení hate speech v komentářích pod příspěvky zpravodajských serverů bude v porovnání s jinými výzkumy stejné nebo vyšší (H1).* Hypotézy H2 až H6 testují vliv vybraných proměnných, jakými jsou pohlaví, vzdělání, místo bydliště, zpravodajský server a vulgárnost komentáře na přítomnost hate speech v komentářích pod příspěvky zpravodajských serverů na sociální síti Facebook. Hypotézy byly ověřovány vybranými kvantitativními metodami na vlastním vzorku 800 případů náhodně sesbíraných komentářů pod příspěvky na facebookových stránkách sledovaných zpravodajských serverů. Kromě samotného obsahu komentáře bylo u každého případu zaznamenáno dalších 13 proměnných týkající se komentáře a autora, které je možné využít v budoucích výzkumech zabývajících se hate speech na českých sociálních sítích.

Jednotlivé hypotézy byly ověřeny za pomoci statistických metod a popsány v kapitole 5. Zastoupení hate speech v komentáři bylo komparováno s podobnými studiemi hate speech, převážně se však studie věnovaly sociální síti Twitter. Pro analyzování vlivu jednotlivých proměnných na přítomnost hate speech v komentářích byly použity kontingenční tabulky. Ve zkoumaném datovém souboru bylo jako hate speech označeno 18,75 % komentářů, což je podstatně více než v podobných studiích na sociální síti Twitter, kde se poměr hate speech komentářů a příspěvků pohyboval od 0,8 % do 11 %, což potvrzuje první hypotézu, tedy že ve

zkoumaném datovém souboru je více hate speech než v podobných studiích. Vyšší procento hate speech bylo zaznamenáno pouze u studie hate speech na českém Facebooku, která byla velmi úzce zaměřena na konfliktní témata migrace a islámu. Vyšší poměr hate speech komentářů v analyzovaném datovém souboru může být částečně způsoben poměrně přísnou škálou, která jako hate speech označovala i komentáře útočí na jednotlivce. Některé studie útoky na jednotlivce hodnotí pouze jako znak ofenzivního projevu, ale nezařazují je mezi projevy nenávistné. Dále je nutné zmínit, že je velmi obtížné porovnávat míru hate speech u jednotlivých studií, které se liší jednak druhem sociální sítě, ale také způsobem sběru dat a odlišnými tématy příspěvků a komentářů.

Analýza vlivu pohlaví na přítomnost hate speech v komentáři testovala hypotézu č. 2, tedy že *komentáře mužů obsahují častěji hate speech než komentáře žen.* Z analýzy vyplynulo, že nejčastěji jsou autory hate speech komentářů takové účty, u kterých nelze určit pohlaví. Také se potvrdilo, že komentáře mužů častěji obsahovaly hate speech, a to o celé 2 % více oproti komentářům žen. Vliv vzdělání na přítomnost hate speech byl testován hypotézou č. 3, tedy že *komentáře vzdělanějších lidí budou obsahovat méně hate speech.* Tato hypotéza se potvrdila, u lidí uvádějících jako vzdělání VOŠ a VŠ byly hate speech komentáře pouze v 14,3 % případů, naopak u lidí uvádějící jako své vzdělání ZŠ nebo fiktivní školy jakými jsou například Vysoká škola života byla hate speech v 30,8 % případů. U lidí uvádějících jako své vzdělání SOU bylo 25 % komentářů obsahujících hate speech. Při analýze *místa bydliště na přítomnost hate speech v komentářích* (H4) se neprokázala žádná souvislost, nejvyšší procento hate speech komentářů bylo u lidí pocházejících ze Zlínského, Olomouckého a Královéhradeckého kraje. Vzhledem k tomu, že nebyla nalezena žádná souvislost mezi místem bydliště a četností hate speech komentářů nešlo k analýze hate speech využít sociodemografická data ze sčítání lidu a data výsledků minulých voleb. Při analyzování vlivu jednotlivých zpravodajských serverů na přítomnost hate speech komentářů se potvrdila hypotéza č. 5, tedy že *příspěvky pod renomovanými zpravodajskými servery budou méně často hate speech komentáře*. Nejméně hate speech bylo zaznamenáno pod příspěvky ČT24, což je zpravodajský kanál veřejnoprávní televize. Naopak nejvíce hate speech bylo zaznamenáno pod příspěvky zpravodajského serveru Parlamentní listy, který je považován za zdroj některých fake news a je nejméně důvěryhodný z vybraných zpravodajských serverů. Dále byl testován *vliv vulgarismů v komentáři na přítomnost hate speech* (H6), zde se hypotéza potvrdila. Komentáře obsahující vulgarismy v 42,2 % případů obsahovaly také hate speech, naopak komentáře bez vulgarismů obsahovaly hate speech jen v 17,4 % případů. Pouze hypotéza *vliv vulgarismů v komentáři na přítomnost hate speech* (H6) se prokázala jako statisticky významná na úrovni spolehlivosti 0,05 (95 %),

tedy existuje statisticky významný rozdíl v přítomnosti hate speech u komentářů obsahující vulgarismy a u komentářů bez vulgarismů a není ovlivněn chybou výběru. U hypotéz H2, H3 a H5 jsou patrné vztahy mezi zkoumanými proměnnými a mírou hate speech v komentářích, tyto vztahy ale nebyly statisticky významné a mohou být zatíženy chybou výběru. Bylo by proto vhodné provést další výzkum na rozsáhlejším datovém souboru.

Přínosem diplomové práce k dosavadnímu výzkumu hate speech na sociálních sítích v České republice je ověření předpokladů o osobnostních charakteristikách autorů hate speech komentářů a zjištění, jak často se objevuje hate speech v komentářích na sociální síti Facebook. Podařilo se zjistit, že i běžná diskuze o politice a světovém dění obsahuje poměrně často hate speech komentáře. Dále bylo zjištěno, že autoři hate speech komentářů jsou častěji muži s nižším vzděláním, kteří pravděpodobně píší hate speech komentáře pod příspěvky méně renomovaných zpravodajských serverů a zároveň je pravděpodobné že budou tyto příspěvky obsahovat vulgarismy.

Diplomová práce zkoumá problematiku na poměrně malém vzorku 800 případů, které jsou náhodně sesbírané pouze z vybraných zpravodajských serverů. Bylo by proto vhodné v dalších výzkumech rozšířit jednak počet případů, tak i počet míst na sociálních sítích kde budou komentáře sbírány. Co do počtu míst na sociálních sítích se zde nabízí jednak další facebookové stránky například politických stran, politiků a známých osobností, ale také soukromé stránky a skupiny extremistických hnutí, kam se mohla část hate speech komentářů přesunout po tom, co se začali autoři některých nenávistných komentářů trestně stíhat. Dalším specifikem práce je zvolená škála pro kategorizaci hate speech, která na jednu stranu dokáže postihnout více útočných a nenávistných projevů než ostatní kategorizace, je zde však problém s následným porovnáním výsledků s ostatními výzkumy. Pro další výzkum je možné použít datový soubor vytvořený v rámci diplomové práce, který obsahuje dalších 7 proměnných, které z důvodu rozsahu diplomové nebyly analyzovány a mohli by tak přinést další zajímavé poznatky o hate speech komentářích a jejich autorech.

Seznam pramenů a literatury

Antifa.cz. 2020. „Kdo jsme?" (online). (cit. 15. 5. 2020) Dostupné z: <https://www.antifa.cz/content/kdo-jsme>

Aro, Jessikka. 2016. „The cyberspace war:propaganda and trollingas warfare tools." *European View* 15 (1): 121–132.

Aronovich, Alex. 2018. „Detect fake profiles - understanding phishing" *cybintsolutions.com,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.cybintsolutions.com/detect-fake-profiles-phishing/>

Bastl, Martin, Miroslav Mareš, Josef Smolík, Petra Vejvodová. 2011. *Krajní pravice a krajní levice v ČR*. Praha: Grada.

Boeckmann, Robert, Carolyn Turpin-Petrosino. 2002. „Understanding the Harm of Hate Crime." *Journal of Social Issue* 58 (2): 207–225.

British Institute of Human Rights. 2012. „Mapping study on projects against hate speech online" *Council of Europe,* (online). (cit. 22. 3. 2020) Dostupné z: <https://rm.coe.int/16807023b4>

Burnap, Pete, Matthew L. Williams. 2015. „Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making: Machine Classification of Cyber Hate Speech." *Policy and Internet* 7 (2): 223–242.

Cakl, Ondřej. 2019. „Nenávist šířená po internetu." *Transparency International,* 16. 12. 2019 (online). (cit. 15. 5. 2020) Dostupné z: <https://www.transparency.cz/nenavist-sirena-po-internetu/>

Cambridge Dictionary. „Emoji." *dictionary.cambridge.org,* (online). (cit. 17. 5. 2020) Dostupné z: <https://dictionary.cambridge.org/dictionary/english/emoji>

Cambridge Dictionary. „Fake News definition." *dictionary.cambridge.org,* (online). (cit. 17. 5. 2020) Dostupné z: <https://dictionary.cambridge.org/dictionary/english/fake-news>

Cambridge Dictionary. „Hate speech definition." *dictionary.cambridge.org,* (online). (cit. 17. 5. 2020) Dostupné z: <https://dictionary.cambridge.org/us/dictionary/english/hate-speech>

Committee of Ministers. 1997. „Recommendation No. R (97) 20 of the Committee of Ministers to Member States on Hate Speech." *Council of Europe*, (online). (cit. 15. 5. 2020) Dostupné z: <https://rm.coe.int/1680505d5b>

„Committee of Ministers." *Council of Europe*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.coe.int/en/web/no-hate-campaign/committee-of-ministers1>

„Council of Europe's work on hate speech." *Council of Europe*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.coe.int/en/web/no-hate-campaign/coe-work-on-hate-speech>

České noviny. 2019. „Teplický soud řeší nenávistné komentáře pod fotkou prvňáků." *ceskenoviny.cz*, (online). (cit. 22. 3. 2020) Dostupné z: <https://www.ceskenoviny.cz/zpravy/teplicky-soud-resi-nenavistne-komentare-pod-fotkou-prvnaku/1749421>

Český helsinský výbor. 2012. „Zpráva Českého helsinského výboru o stavu lidských práv za rok 2012." *helcom.cz*, (online). (cit. 22. 3. 2020) Dostupné z: <http://www.helcom.cz/dokumenty/uploads/2013/09/ZLP_2012.pdf>

ČT24. „Praktiky Parlamentních listů podle rozsudku soudu: Démonizace, manipulace a apel na strach" *ct24.cz*, 20. 12. 2019 (online). (cit. 15. 5. 2020) Dostupné z: <https://ct24.ceskatelevize.cz/media/3010385-praktiky-parlamenich-listu-podle-rozsudku-soudu-demonizace-manipulace-a-apel-na-strach>

Davidson, Thomas, Dana Warmsley, Michael Macy, Ingmar Weber. 2017. „Automated hate speech detection and the problem of offensive language." *ICWSM-17*, (online). (cit. 15. 5. 2020) Dostupné z: <https://arxiv.org/pdf/1703.04009.pdf>

Developer Twitter, 2020. „Publish and manage Tweets, and analyze Tweet data." twitter.com, (online). (cit. 18. 5. 2020) Dostupné z: <https://developer.twitter.com/en/products/tweets>

ElSherief, Mai, Kulkarni Vivek, Nguyen Dana, Wang William, Belding Elizabeth. 2018. „Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media." *ICWSM-18*, (online). (cit. 15. 5. 2020) Dostupné z: <https://arxiv.org/pdf/1804.04257.pdf>

European Commission against Racism and Intolerance. 2016. „ECRI General policy recommendation NO. 15 on combating hate speech." *Council of Europe*, (online). (cit. 15. 5.

2020) Dostupné z: <https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01>

„Evropská úmluva o ochraně lidských práv.“ *Evropský soud pro lidská práva,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.echr.coe.int/Documents/Convention_CES.pdf>

Evropský soud pro lidská práva: Rozhodnutí ve věci Handyside proti Spojenému království ze dne 7. prosince 1976. Stížnost č. 5493/72, § 49.

Facebook Community Standards. „Objectionable content - Hate speech.” *Facebook.com,* (online). (cit. 14. 4. 2020) Dostupné z: <https://www.facebook.com/communitystandards/objectionable_content>

Fendrych, Martin. 2016. „Hate speech proti imigrantům: Nenávist je na internetu šířena ve jménu Dobra.” *Aktuálně.cz*, 31. 3. 2016 (online). (cit. 22. 3. 2020) Dostupné z: <https://nazory.aktualne.cz/komentare/hate-speech-nenavist-sirena-ve-jmenu-dobra/r~8f6a3848f72e11e5af7e0025900fea04/>

Gambäck, Björn, Sikdar Utpal. 2017. „Using Convolutional Neural Networks to Classify Hate-Speech.“ in *Proceedings of the First Workshop on Abusive Language Online*: 85–90.

Herczeg, Jiří. 2008. *Trestné činy z nenávisti.* Praha: Wolters Kluwer.

Hoax.cz. 2020. „Co je to hoax.” (online). (cit. 18. 5. 2020) Dostupné z: <https://www.hoax.cz/hoax/co-je-to-hoax>

Holgate, Eric, Isabela Cachola, Daniel Preotiuc-Pietro, Junyi Jessy Li. 2018. „Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions.” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*: 4405–4414.

Hrdina, Matouš, Hana Daňková, Liudmila Kopecka. 2017. „Projevy nenávisti v online prostoru a na sociálních sítích.” *Člověk v tísni, o.p.s.,* (online). (cit. 22. 3. 2020) Dostupné z: <https://www.clovekvtisni.cz/media/publications/553/file/1459365027-hate-speech-zaverecnazprava-final-verze.pdf>

Chaffey, Dave. 2020. „Global social media research summary 2020.” *Smart Insights,* 17. 4. 2020 (online). (cit. 15. 5. 2020) Dostupné z: <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>

Jäger, Petr, Pavel Molek. 2007. *Svoboda projevu: Demokracie, rovnost a svoboda slova*. Praha: Auditorium.

Jelínek, Milan, Jarmil Vepřek. 2017. „VULGARISMUS.“ in: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.) *Nový encyklopedický slovník češtiny,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.czechency.org/slovnik/VULGARISMUS>

Kopecký, Josef. 2017. „Nejvíc nenávisti je k Romům, zjistil projekt HateFree. Dostane bratříčka.” *iDnes.cz,* 7. 3. 2017 (online). (cit. 22. 3. 2020) Dostupné z: <https://www.idnes.cz/zpravy/domaci/kampan-hatefree-bude-mit-nastupce.A170307_152400_domaci_kop>

Kurzy.cz. „AGROFERT, a.s.“ *Rejstřík friem Kurzy.cz* (online). (cit. 15. 5. 2020) Dostupné z: <https://rejstrik-firem.kurzy.cz/26185610/agrofert-as/>

Ma, Alexandra, Ben Gilbert. 2019. „Facebook understood how dangerous the Trump-linked data firm Cambridge Analytica could be much earlier than it previously said. Here's everything that's happened up until now.” *Businessinsider.com,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3#is-russia-involved-7>

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder. 2019. „Hate speech detection: Challenges and solutions.” *PLoS ONE* 14 (8).

Mareš, Miroslav. 2011. „Problematika Hate Crime“ *mvcr.cz,* (online). (cit. 15. 5. 2020) Dostupné z: <www.mvcr.cz/soubor/problematika-hate-crime.aspx>

„Mezinárodní pakt o občanských a politických právech ze dne 23. března 1976, zveřejněn ve Sbírce zákonů pod č. 120/176 Sb.“ *Beck-online,* (online) (cit. 15. 5. 2020) Dostupné z: <https://www.beck-online.cz/bo/chapterview-document.seam?documentId=onrf6mjzg43f6mjsgaxgi2brfuya>

Michl, Petr. 2019. „Infografika: Sociální sítě v Česku v roce 2019.” *Focus Agency*, 8. 11. 2019 (online). (cit. 15. 5. 2020) Dostupné z: <https://www.focus-age.cz/m-journal/aktuality/infografika--socialni-site-v-cesku-v-roce-2019__s288x14828.html>

Miró-Llinares, Fernando, J.J. Rodriguez-Sala. 2016. „Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy.“ *International Journal of Design & Nature and Ecodynamics* 11 (3): 406–415.

Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. „A Measurement Study of Hate Speech in Social Media.“ in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*: 85–94.

Moulisová, Marcela. 2008. „Rasově motivovaná trestná činnost.“ in Josef Zapletal et al. (eds.) *Aktuální problémy kriminologie*. Praha: Policejní akademie ČR.

Nahodil, Tomáš. 2019. „Pavel Zeman: Nenávistné příspěvky na internetu jsou nemocí naší doby.” *Česká justice,* 16. 10. 2019 (online). (cit. 15. 5. 2020) Dostupné z: <https://www.ceska-justice.cz/2019/10/pavel-zeman-nenavistne-prispevky-internetu-jsou-nemoci-nasi-doby/>

Netrvalová, Sabina. 2017. *Moderování a regulace čtenářských diskuzí na českých zpravodajských serverech.* Praha. Diplomová Práce. Univerzita Karlova. Fakulta sociálních věd.

NFNZ. „Hodnocení médií.“ *nfnz.cz*, (online). (cit. 15. 5. 2020) Dostupné z: <https://rating.nfnz.cz>

Njagi, Dennis, Z. Zuping, Damien Hanyurwimfura, Jun Long. 2015. „A Lexicon-based Approach for Hate Speech Detection.” *International Journal of Multimedia and Ubiquitous Engineering* 10(4): 215–230.

Nobota, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang. 2016. „Abusive language detection in online user content.“ *International Conference on World Wide Web, WWW '16*, (online). (cit. 15. 5. 2020) Dostupné z: <http://www.yichang-cs.com/yahoo/WWW16_Abusivedetection.pdf>

Objevit.cz. 2015. „K čemu jsou dobré hashtagy?” *Objevit.cz*, 13. 10. 2015 (online). (cit. 17. 5. 2020) Dostupné z: <https://www.objevit.cz/k-cemu-jsou-dobre-hashtagy-t154352>

Oxford Constituonal Law. „Hate speech.” *oxcon.ouplaw.com,* (online). (cit. 17. 5. 2020) Dostupné z: <https://oxcon.ouplaw.com/view/10.1093/law-mpeccol/law-mpeccol-e130>

Park, Jo Ho, Pascale Fung. 2017. „One-step and two-step classification for abusive language detection on Twitter.“ in *Proceedings of the First Workshop on Abusive Language Online*: 41–45.

Policie české republiky. „Trestné činy z nenávisti“ *policie.cz*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.policie.cz/clanek/trestne-ciny-z-nenavisti.aspx >

Polston, Vince. 2018. „How to Spot Fake Facebook Profile“ *malwarefox.com*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.malwarefox.com/spot-fake-facebook-profile/>

Province of Ontario Ministry of Attorney General. „Hate Crime and Discrimination“ *attorneygeneral.jus.gov.on.ca*, (online). (cit. 15. 5. 2020) Dostupné z: <http://www.attorneygeneral.jus.gov.on.ca/english/crim/cpm/2005/HateCrimeDiscrimination.pdf>

Romanov, Aleksei, Alexander Semenov, Oleksiy Mazhelis, Jari Veijalainen. 2017. „Detection of Fake Profiles in Social Media - Literature Review.“ in *Proceedings of the 13th International Conference on Web Information Systems and Technologies*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.scitepress.org/Papers/2017/63621/63621.pdf>

Rosenfeld, Michel. 2001. „Hate Speech in Constitutional Jurisprudence: A Comparative Analysis.” in *Cardozo Law School Legal Studies Research Paper No. 41*.

Rosentiel, Tom. 2008. „Where Men and Women Differ in Following the News.” *Pew Research Center*, (online). (cit. 22. 3. 2020) Dostupné z: <https://www.pewresearch.org/2008/02/06/where-men-and-women-differ-in-following-the-news/>

Sharma, Sanjana, Saksham Agrawal, Manish Shrivastava. 2018. „Degree based classification of harmful speech using twitter data.“ in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*: 106–112.

Schmidt, Anna, Michael Wiegand. 2017. „A Survey on Hate Speech Detection using Natural Language Processing.“ in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*: 1–10.

Similar Web. „Website Traffic Statistics & Analytics“ *SimilarWeb.com,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.similarweb.com>

Sun, Key. 2006. „The legal definition of hate crime and the hate offender's distorted cognitions.“ *Issues in mental health nursing* 27 (6): 597–604.

Twitter Help Center. „Hateful conduct policy.” *twitter.com,* (online). (cit. 22. 6. 2020) Dostupné z: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

„Ústavní zákon č. 2/1993 Sb., Listina základních práv a svobod.“ *psp.cz,* (online). (cit. 15. 5. 2020) Dostupné z: <http://www.psp.cz/docs/laws/listina.html>

Viktora, Antonín. 2010. „Soud s předáky českých neonacistů trval jen pár minut.” iDnes.cz, 15. 7. 2010 (online). (cit. 15. 5. 2020) Dostupné z: <https://www.idnes.cz/zpravy/cerna-kronika/soud-s-predaky-ceskych-neonacistu-trval-jen-par-minut.A100715_1416950_krimi_cen>

Vomelová, Jana, Eva Nehudková. 2018. „Nenávistné projevy ve veřejném prostoru.” *Ombudsman veřejný ochránce práv,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.ochrance.cz/fileadmin/user_upload/projekt_ESF/00_2018_VA/KULATE_STOLY/03_14_Rok_2017_v_oblasti_boje_proti_diskriminaci/03_14_DS_Nenavistne_projevy_ve_verejnem_prostoru_PREZENTACE.pdf>

Vozková, Kristýna. 2019. „Řazení komentářů na Facebooku má nová pravidla“ *focus-age.cz*, (online). (cit. 15. 5. 2020) Dostupné z: <https://www.focus-age.cz/m-journal/aktuality/razeni-komentaru-na-facebooku-ma-nova-pravidla__s288x14532.html>

Výborný, Štěpán. 2012. *Svoboda slova versus nenávistné projevy na internetu.* Brno. Rigorózní Práce. Masarykova univerzita. Právnická fakulta.

Výborný, Štěpán. 2013. *Nenávistný internet versus právo.* Praha: Wolters Kluwer.

Výzkum veřejného ochránce práv. 2020. „Nenávistné projevy na internetu a rozhodování českých soudů.” *Ombudsman veřejný ochránce práv,* (online). (cit. 15. 5. 2020) Dostupné z: <https://www.ochrance.cz/fileadmin/user_upload/DISKRIMINACE/Vyzkum/47-2019-DIS-vyzkum_nenavist.pdf>

Waseem, Zeerak, Dirk Hovy. 2016. „Hateful Symbols or Hateful People?Predictive Features for Hate Speech Detection on Twitter.” in *Proceedings of the NAACL Student Research Workshop*: 88–93.

Zákon č. 40/2009 Sb., trestní zákoník, ve znění pozdějších předpisů. in *Zákony pro lidi,* 1. 6. 2020 (online). (cit. 16. 6. 2020) Dostupné z: <https://www.zakonyprolidi.cz/cs/2009-40>

Zavoral, Petr. 2015. „Analýza: Valí se na nás hordy verbeže, negrů a primitivů.” *hatefree.cz,* 29. 7. 2015 (online). (cit. 22. 3. 2020) Dostupné z: <https://www.hatefree.cz/blo/analyzy/1049-analyza-verbez>

Ziqi, Zhang, Lui Lei. 2019. „Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter.“ *Semantic Web* 10 (5): 925–945.

Abstrakt

Diplomová práce se zabývá hate speech v komentářích na facebookových stránkách českých zpravodajských serverů. Cílem práce je zjistit v jaké míře se v komentářích vyskytuje hate speech a kdo jsou uživatelé, kteří hate speech komentáře píší. Za tímto účelem byl vytvořen unikátní datový soubor 800 náhodně sesbíraných komentářů v období od 30. dubna 2019 do 19. ledna 2020, komentáře byly sesbírány na facebookových stránkách zpravodajských serverů iDnes, Aktuálně, Novinky, ČT24 a Parlamentní listy. Zjištěním je, že analyzovaný soubor komentářů obsahuje podstatně vyšší procento hate speech, než bylo zjištěno v podobných zahraničních studiích. Analýza dále prokázala, že existují charakteristiky, které mohou predikovat vyšší šanci na to, že v komentářích bude obsažena hate speech. Analýza prokázala, že účty, u kterých není známo pohlaví píší hate speech častěji než účty s rozpoznatelným pohlavím, také se prokázalo, že muži jsou častěji autory hate speech komentářů než ženy. Prokázal se také vliv vzdělání na přítomnost hate speech v komentářích, autoři s nižším vzděláním častěji píší hate speech komentáře, a naopak uživatelé s vyšším vzděláním píší hate speech komentáře spíše výjimečně. Dále bylo zjištěno, že komentáře na stránkách renomovaných zpravodajských serverů, jakými je například ČT24 obsahují méně hate speech, a naopak komentáře pod méně renomovanými servery obsahují více hate speech. Dále má velmi silný vliv na přítomnost hate speech vulgárnost komentáře, kde se prokázalo, že vulgární komentář mnohem častěji obsahuje také hate speech.

Klíčová slova: nenávistné projevy, svoboda slova, hate speech, sociální sítě, zpravodajské servery, Facebook

Abstract

The diploma thesis deals with hate speech in comments on the Facebook pages of Czech news servers. The aim of this work is to find out to what extent there is hate speech in the comments and who are the users who write hate speech comments. For this purpose, a unique data set of 800 randomly collected comments was created in the period from 30 April 2019 to 19 January 2020, the comments were collected on the Facebook pages of the news servers iDnes, Aktuálně, Novinky, ČT24 and Parlamentní listy. The finding is that the analyzed set of comments contains a significantly higher percentage of hate speech than was found in similar foreign studies. The analysis further showed that there are characteristics that may predict a higher chance that hate speech will be included in the comments. The analysis showed that accounts for which no gender is known write hate speech more often than accounts with a recognizable gender, and it has also been shown that men are more likely to write hate speech than women. The influence of education on the presence of hate speech in comments has also been proven, authors with lower education write hate speech comments more often and, conversely, users with higher education write hate speech comments more rarely. Furthermore, it was found that comments on the pages of reputable news servers such as ČT24 contain less hate speech and, conversely, comments under less reputable servers contain more hate speech. Furthermore, the vulgarity of commentary has a very strong influence on the presence of hate speech, where it has been shown that vulgar commentary also contains hate speech much more often.

Keywords: hate speech, freedom of speech, social networks, news servers, Facebook

Seznam grafů

Seznam tabulek

Seznam obrázků