MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

Rozpoznání anomálních uživatelů na českém Twitteru

DIPLOMOVÁ PRÁCE

Petr Jirásek

Brno, jaro 2015

Prohlášení

Prohlašuji, že tato diplomová práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Petr Jirásek

Vedoucí práce: doc. RNDr. Lubomír Popelínský, Ph.D.

Klíčová slova

Dolování z dat, detekce odlehlých bodů, rozpoznání anomálních uživatelů, Twitter, klasifikace sentimentu

Poděkování

Rád bych poděkoval vedoucímu práce doc. RNDr. Lubomíru Popelínskému, Ph.D. za jeho odborné vedení a věcné rady. Dále bych rád poděkoval své rodině za podporu při psaní této práce a svým přátelům za poskytnutí zpětné vazby. Zvláště bych rád poděkoval Jiřímu Jiráskovi a Josefu Buryanovi za pomoc při anotaci dat.

Shrnutí

Diplomová práce se zabývá návrhem metody pro rozpoznání anomálních uživatelů na sociální síti Twitter, jejich charakteristikou a vlivem metody odstranění anomálií na výsledky klasifikace. V práci je čtenář seznámen se specifiky sociální sítě Twitter a způsobem dolování získaných dat. Dále práce specifikuje pojem anomálního a normálního uživatele a popisuje existující přístupy v této oblasti. V hlavní části práce je podrobně popsána navržená metoda včetně výsledků testovaných přístupů. Následuje analýza anomálních uživatelů a jejich charakteristika doplněná o příklady. Zvláštní pozornost je pak věnována klasifikaci sentimentu a ověření vlivu navržené metody na výsledky klasifikace.

Obsah

1 **Úvod** 1
2 **Dolování dat z Twitteru a jeho specifika** 3
2.1 Sociální síť Twitter 3
2.1.1 Terminologie 4
2.1.2 Specifika sítě 5
2.1.3 Komunity 6
2.2 Dolování z dat a detekce anomálií 6
2.3 Dolování dat z Twitteru 7
2.3.1 Twitter API 8
2.3.2 Česká datová sada 8
2.3.3 Anglická datová sada 10
3 **Anomální uživatelé a existující metody pro jejich detekci** . . . 11
3.1 Kdo je to anomální a normální uživatel? 11
3.2 Motivace pro hledání anomálních uživatelů 12
3.3 Existující metody 12
3.3.1 Detekce robotických účtů s využitím klasifikace . . 12
3.3.2 Detekce spam účtů s využitím klasifikace 13
3.3.3 Detekce spam účtů seskupením uživatelů 14
3.3.4 Detekce spam účtů metodou shlukování 15
4 **Návrh metody pro rozpoznání anomálních uživatelů** 17
4.1 Informace o uživateli 17
4.1.1 Chování uživatele a informace o jeho účtu 17
4.1.2 Časové informace 21
4.1.3 Obsah příspěvků 21
4.2 Anotace uživatelů 22
4.3 Předzpracování příspěvků 23
4.4 Výběr atributů 24
4.4.1 Chování atributů 24
4.4.2 Výběr pomoci selektivních algoritmů 27
4.4.3 Testování a konečný výběr 29
4.5 Testované přístupy a jejich výsledky 30
4.5.1 Učení bez učitele 31
4.5.2 Učení z klasifikovaných a neklasifikovaných příkladů 34
4.5.3 Učení s učitelem 35
4.6 Shrnutí 37

5 **Charakteristika anomálních uživatelů** 39
5.1 Shlukování anomálních uživatelů 39
5.1.1 Hierarchické shlukování 40
5.1.2 K-means 41
5.2 Charakteristiky uživatelů 42
5.2.1 Robotické účty 42
5.2.2 Spam účty 43
5.2.3 RSS čtečky, informační a propagační účty 43
5.2.4 Jednosměrně tématicky zaměřené účty 44
6 **Klasifikace a vliv anomálních uživatelů** 45
6.1 Klasifikace sentimentu 45
6.2 Učení na základě emotikonů 45
6.3 Trénovací množina 46
6.4 Testovací množina 47
6.5 Výsledky klasifikace do tří tříd 48
6.5.1 Bez odstranění anomálních uživatelů 48
6.5.2 Po odstranění anomálních uživatelů 48
6.6 Výsledky klasifikace do dvou tříd 52
6.6.1 Bez odstranění anomálních uživatelů 52
6.6.2 Po odstranění anomálních uživatelů 53
6.6.3 Využití pro klasifikaci do tří tříd 56
6.7 Shrnutí 57
7 **Závěr** 59
A **Nástroj pro anotaci uživatelů** 67
B **Obsah přiložených souborů** 68

1 Úvod

Poslední roky zažíváme rozmach sociálních sítí a rozsáhlých komunitních služeb, které využívají miliony uživatelů a zanechávají zde svá data, ať už se jedná o osobní informace nebo informace o svém chování v kontextu používané služby. S ohledem na současný trend dolování informací je tak možné data z těchto služeb využít k hlubší analýze a hledat mezi daty skryté závislosti a jevy, které nám mohou pomoci pochopit chování uživatelů a vazby mezi nimi. Příkladem takového chování může být jednání osoby snažící se šířit hromadné nevyžádané zprávy nebo chování osoby vykazující podezřelou aktivitu v kontextu svého okolí nadměrným užíváním zřídka se vyskytujících klíčových slov.

Sociální síť Twitter je jednou z největších sociálních sítí na světě. Odhaduje se, že její počet měsíčně aktivních uživatelů dosahuje 288 miliónů [1]. I mezi uživateli na sociální síti Twitter se nacházejí účty, které se chovají odlišně než ostatní a jedná se o tzv. anomální uživatele. Jejich nalezením můžeme analyzovat, zda se tito uživatelé nechovají proti obecným pravidlům sítě, zda se kupříkladu jedná o tzv. robotické účty nebo pouze o účty, které se v některých ze svých znaků chovají příliš odlišně.

Twitter netvoří pouze uživatelské profily, ale jsou to především příspěvky a obsah, který uživatelé vytváří. Obsah, který je díky jeho velikosti a dostupnosti ve strojově čitelné podobě dále možné využít pro účely klasifikace a jiných metod nejen z oblasti strojového učení. Lze tak prostřednictvím těchto dat kupříkladu učit různé nástroje, které jsou schopny rozpoznat téma příspěvku nebo určit jeho sentiment [2, 3]. Úspěšnost těchto nástrojů je ale zpravidla tak dobrá, jak dobrá jsou data, z kterých se nástroje učí.

Cílem této práce je tedy navrhnout metodu pro rozpoznání anomálních uživatelů, provést jejich charakteristiku a ověřit vliv informace o anomálních uživatelích na výsledky klasifikace sentimentu na sociální síti Twitter.

Práce je rozdělena do několika kapitol. V kapitole 2 se seznámíme se sociální síti Twitter, pojmem dolování z dat a způsobem dolování dat z této sítě. V následující kapitole je věnována pozornost pojmu anomální uživatel a existujícím přístupům pro detekci některých podskupin těchto uživatelů. Ve čtvrté kapitole se práce zabývá návrhem metody pro rozpoznání anomálních uživatelů a jejími výsledky. Poté následuje kapitola zabývající se charakteristikou anomálních uživatelů doplněná o příklady. V šesté kapitole je věnována pozornost klasifikaci sentimentu a vlivu informace o anomálních

uživatelích na výsledky klasifikace při použití navržené metody. Poslední kapitola pak obsahuje závěrečné shrnutí výsledků a možná pokračování této práce.

2 Dolování dat z Twitteru a jeho specifika

V první části se kapitola zabývá základním popisem sociální sítě Twitter a jejich specifik včetně popisu základní terminologie. V další části pak představuje stručný úvod do oblasti dolování z dat a detekci anomálií a poté se dále zaměřuje na dolování dat z českého a anglického Twitteru. Popisuje heuristiky pro získání dat a doplňuje související informace.

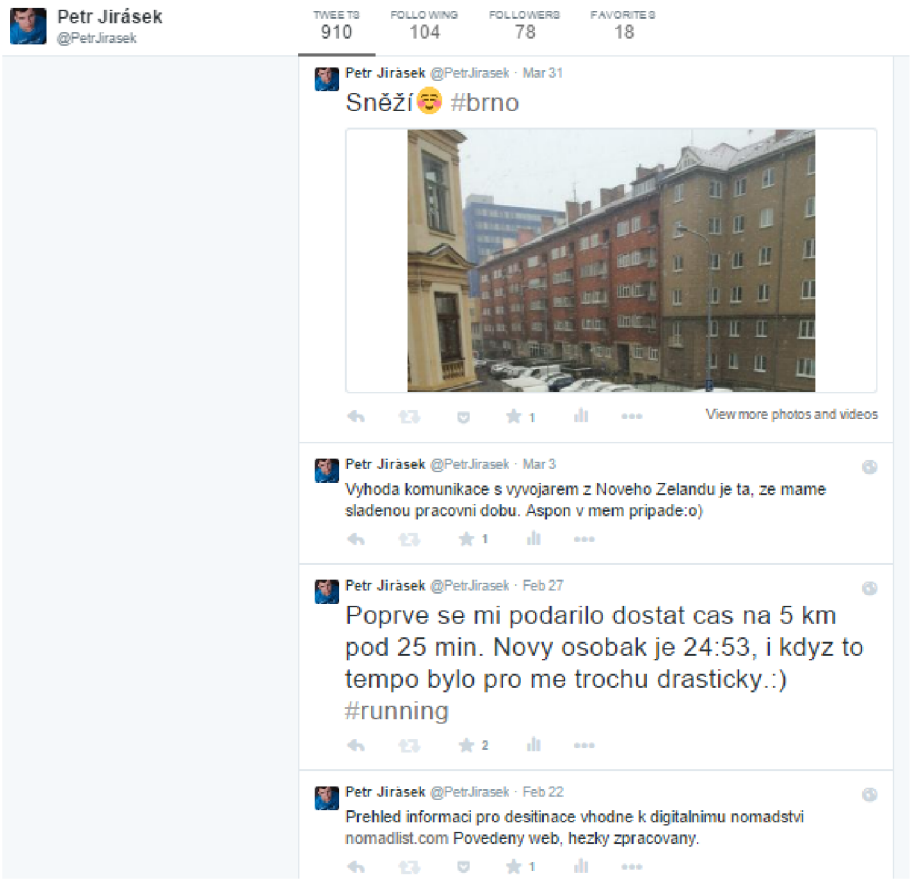
2.1 Sociální síť Twitter

Sociální síť Twitter je jednou z největších sociálních sítí a řadí se svou velikosti k sítím jako Facebook nebo Instagram [4]. Svým charakterem se jedná o mikroblogovací službu, kde uživatelé vkládají příspěvky krátké velikosti (tzv. tweet) s maximální délkou 140 znaků na svou nástěnku a mohou zároveň reagovat na příspěvky ostatních. Právě svým omezením na počet znaků se sociální síť Twitter výrazně odlišuje od ostatních sociálních sítí.

Příspěvky mohou obsahovat kromě textu také odkazy, zmínky uživatelů nebo klíčová slova, což pomáhá k větší interakci v rámci sítě a umožňuje uživatelům vyhledávat zprávy dle jejich oblasti zájmu.

Uživatelé mohou od jiných uživatelů odebírat jejich zprávy a zároveň mají své vlastní odběratele. Dále mohou příspěvky uživatelů, od kterých odebírají obsah, označit jako oblíbené nebo převzít a sdílet svým odběratelům umístěním zprávy na svou nástěnku. Právě díky funkci převzetí příspěvku dochází k rychlému šíření informací v síti, přičemž taková informace může být pravdivá, ale také falešná (viz. případ napadení účtu tiskové agentury [5]).

Obrázek 2.1 zobrazuje ukázku profilu uživatele s příspěvky na sociální síti Twitter.



Obrázek 2.1: Profil uživatele s příspěvky

2.1.1 Terminologie

Nyní si popíšeme základní pojmy používané na sociální síti Twitter.

Odběratel Odběratel je uživatel sledující příspěvky jiného uživatele.

Klíčové slovo Klíčové slovo (tzv. „hashtag“) může být obsažené v příspěvku a zpravidla nese význam. V textu je vždy uvozeno znakem #.

Zmínka Zmínka v příspěvku reprezentuje odkaz na uživatele (tzv. „mention“). V textu je vždy uvozena znakem @.

Médium Médium může být přiloženo k příspěvku. Zpravidla se jedná o fotografie, videa nebo GIF animace.

Převzetí zprávy Převzetí zprávy (tzv. „retweet“) je funkce umožňující převzít příspěvek jiného uživatele a sdílet ho svým odběratelům.

Oblíbená zpráva Uživatel může zprávy jiných uživatelů označit jako oblíbené (označit jako „favorite“).

2.1.2 Specifika sítě

Twitter je oproti jiným sociálním sítím v několika ohledech specifický a to především z pohledu délky příspěvků, jazykového modelu a domény vkládaných příspěvků.

Délka příspěvku Každý příspěvek na sociální síti Twitter může obsahovat maximálně 140 znaků, což může ovlivnit jak způsob, jakým autor příspěvku zprávu koncipuje, tak množství informací, které může obsahovat. Někteří uživatelé se snaží limit na počet znaků obcházet vložením odkazu do konce zprávy směřujícího mimo sociální síť na její pokračování. Dalším způsobem je přiložení obrázku ke zprávě obsahujícího zaznamenaný text rozšiřující sdělení.

Jazykový model Samotné omezení délky příspěvku přímo ovlivňuje také jazykový model. Uživatelé mají tendenci používat různé slangové zkratky, přesmyčky, aby tak nepřekročili daný počet znaků, což ve výsledku přidává na neformálnosti výsledné zprávy. Někteří zcela ignorují diakritiku nebo interpunkční znaménka a výjimkou není vynechání samohlásek u často používaných slov. Do zprávy také uživatelé vkládají emotikony a klíčová slova, zmiňují ostatní uživatele a využívají expresivních výrazů.

Doména Příspěvky se mohou týkat různého tématu a mohou reflektovat aktuální situaci v okolí uživatele nebo ve světě. Na síti lze sledovat trendy, které rezonují napříč celou komunitou. Může se kupříkladu jednat o zprávy

z válečných konfliktů, úmrtí významných osobností nebo informace o jiných mimořádných událostech.

Na druhou stranu, řada uživatelů informuje o novinkách ve svém životě, ve svém okolí nebo jednoduše pouze o tom, co zrovna dělají nebo dělali. Dále zde existuje řada uživatelských účtů firem, které mají své komerční zájmy a zaměřují se na své kampaně a komunikaci se svou cílovou skupinou. Zvláštní skupinou jsou pak robotické účty informující o počasí, o aktuálním dni nebo o tom, kdy zrovna dochází k nějaké události (např. kdy odbíjejí hodiny na Pražském orloji apod.).

2.1.3 Komunity

Twitter je globální sociální síť využívána napříč uživateli celého světa reprezentujícími komunity. Mezi nejpočetnější patří anglická komunita, která zastřešuje uživatele z USA, Velké Británie a z dalších anglicky hovořících zemí a rozšiřují ji také uživatelé, kteří vkládají příspěvky v angličtině, ale jejich národním jazykem angličtina není.

Česká komunita patří mezi ty menší a její velikost je odhadována na 150 000 až 200 000 uživatelů [6, 7], jelikož v rámci české komunity mohou být i uživatelé, kteří převážně přispívají příspěvky v angličtině nebo v jiném cizím jazyce. Je nutné brát také na zřetel, že se v české komunitě mohou vlivem blízkosti jazyka mísit i uživatelé ze slovenské komunity, což dále snižuje možnosti přesného odhadu velikosti. Český Twitter je především tvořen mladšími uživateli a lidmi s vyšším vzděláním z řad sportovců, novinářů, známých osobností, IT pracovníků nebo osob z obchodního prostředí. Také se zde pohybuje řada firem a informačních účtů, kteří se snaží komunikovat se svou cílovou skupinou.

2.2 Dolování z dat a detekce anomálií

Dolování z dat (Data mining) někdy také označováno jako dobývání znalostí z databází (Knowledge Discovery in Databases) je proces netriviální extrakce implicitních, dříve neznámých a potenciálně užitečných informací z dat [8]. Zároveň propojuje poznatky z databází, statistiky a strojového učení. Dolování z dat můžeme použít na řešení mnoha různých problémů. Jedním z takových problémů může být třeba predikce hodnot v časové řadě, kdy známe řadu historických hodnot a chceme předpovídat, jaké další hod-

noty budou následovat. Dalším může být shlukování, kdy hledáme v datech shluky objektů s podobnými rysy. Metod, které lze uplatnit v této oblasti je celá řada a s ohledem na rostoucí množství dat ve všech možných disciplínách roste také jejich uplatnění [8, 9].

V této práci se budeme zaměřovat především na metody založené na učení s učitelem, učení z klasifikovaných a neklasifikovaných příkladů a učení bez učitele. Učení s učitelem spočívá v práci s daty, kde jsou přiřazené třídy anotátorem (učitelem). Typicky při klasifikaci máme k dispozici data, kde jsou příklady v trénovací množině označeny třídami. Učení z klasifikovaných a neklasifikovaných příkladů naopak pracuje s daty, kde u první části stále známe přiřazené třídy, ale u druhé části tuto informaci již nemáme. Tento způsob učení se používá typicky v doménách, kde je těžké získat anotovaná data. Učení bez učitele naopak pracuje s daty, kde nemáme přiřazené žádné třídy. Taková data pak můžeme například shlukovat, kdy hledáme objekty s podobnými znaky tvořící shluky nebo můžeme v datech hledat odlehlé objekty označované také jako anomálie [9].

Právě detekce anomálií úzce souvisí s touto prací. Jak už bylo naznačeno, detekce anomálií spočívá v hledání takových instancí v datech, které jsou příliš odlišné nebo vzdálené od ostatních. Anomálie můžeme hledat několika způsoby. Jedním z nich je využití klasifikátoru, kdy trénovací množina je tvořena příklady, které anotátor označí jako odlehlé a těmi, které naopak ne. Klasifikátor se tak může naučit rozpoznávat anomální instance a tento přístup odpovídá kategorii metod učení s učitelem. Můžeme také využít klasifikátory, které jsou schopny pracovat s trénovací množinou, kde část příkladů má přiřazenou třídu a část ne. Pokud ale instance v datech nemají přiřazené žádné třídy, lze použít některou z metod učení bez učitele, kde se typicky počítá odlehlost každé instance od ostatních na základě její vzdálenosti [9].

2.3 Dolování dat z Twitteru

Získání relevantních dat k analýze je důležitým předpokladem pro úspěšné dolování z dat. Twitter nabízí nástroje pro vývojáře umožňující přístup k informacím a pokud chceme například získat vzorek dat z české komunity, musíme použít heuristiku [10, 11].

2.3.1 Twitter API

Twitter API nabízí třetím stranám možnost přistupovat k datům na Twitteru. Mezi dostupná data patří informace o uživatelích, o jejich odběratelích nebo o příspěvcích uživatele. Dále je možné provádět vyhledávání na síti nebo ovládat vlastní účet skrze nástroje rozhraní. Požadavky, které klient zasílá, jsou omezeny maximálním limitem v rámci časového okna. U požadavku na získání informací o uživateli nebo vyhledávání příspěvků se jedná o 180 požadavků v rámci 15 minutového časového intervalu [10].

Z pohledu licenčního ujednání společnosti Twitter nelze získaná data jakýmkoliv způsobem distribuovat třetím stranám nebo libovolně měnit jejich licenci s výjimkou písemného souhlasu společnosti [12].

2.3.2 Česká datová sada

Jelikož Twitter nepodporuje vyhledávání česky psaných příspěvků [11], je tak obtížnější získat data, která by obsahovala relevantní vzorek českého Twitteru. Proto bylo nutné při získávání dat pro účely této práce navrhnout heuristiku, která nám umožní relevantní vzorek získat.

Heuristika využívá předpokladu, že české účty přispívající v českém jazyce s vysokým počtem odběratelů budou většinově sledovat čeští uživatelé. Proto jsou v prvním kroku heuristiky vybráni uživatelé, kteří jsou pravděpodobnými ústředními body české komunity. Mezi tyto uživatele patří zpravidla politici, novináři, známé osobnosti, české informační servery nebo osoby z obchodního a technického prostředí. Tabulka 2.1 uvádí jmenný seznam vybraných účtů rozdělených do skupin.

V dalším kroku heuristiky je pak nutné získat seznamy odběratelů těchto vybraných uživatelských účtů prostřednictvím Twitter API. Každého získaného odběratele si uložíme do datové struktury, kde se každý uživatel bude nacházet nejvýše jednou. Jakmile datovou strukturu naplníme, můžeme dále pokračovat stejným postupem, kdy budeme do struktury vkládat odběratele jednotlivých uživatelů a volitelně také uživatele, které sami sledují. Jakmile získáme dostatečně velkou komunitu, je vhodné provést filtraci uživatelů. Uživatele je vhodné zařadit do množiny českých uživatelů, pokud prvních 20 nejnovějších zpráv uživatele obsahuje alespoň 25 % česky psaných zpráv, ale práh je možné dále upravovat dle vlastního zvážení.

K detekci jazyka zprávy lze využít informaci z metadat přiložených ke zprávě s využitím detekčního algoritmu Twitteru [13]. Bohužel, pro české

zprávy tento postup nefunguje správně. Při práci s daty se ukázalo, že algoritmus pro detekci jazyka z Twitteru označuje české zprávy za slovenské, což může způsobovat problémy. Řešením je využít nástroj pro detekci jazyka a nespoléhat se na metadata přiložená k příspěvku. Při testování několika nástrojů se nejlépe osvědčila knihovna language-detection [14] v programovacím jazyce Java, která je schopna rozpoznat texty jak v českém, slovenském nebo anglickém jazyce.

Skupina	Uživatelské účty
Novináři a sportovní komentátoři	@cermak, @rozanek, @jaromirbosak, @robertzaruba, @eriktabery, @veselovskyma, @jindrichsidlo, @jansedlak, @polesny, @AlexandrMitrofa
Politici	@AndrejBabis, @schwarzenbergk, @Tomio_Okamura, @PavelBelobradek, @P_Fiala, @ODS_CZ, @PetrMachMEP
Zpravodajské a informační servery	@IDNEScz, @novinkycz, @computer_zive, @Lupacz, @Aktualnecz, @DVTVcz, @CT24zive, @sportCT
Ostatní	@DavidGrudl, @RadekHulan, @stepanb, @geekovo, @janrezac, @janovsky,

Tabulka 2.1: Seznam vybraných významných uživatelských účtů českého Twitteru

Jakmile získáme množinu českých uživatelů, můžeme postupně pro každého uživatele načítat informace o jeho profilu a ty ukládat do perzistentního úložiště. Úložiště by mělo být schopné data efektivně prohledávat a exportovat s možností dotazovat se. Pro tyto účely se dostatečně osvědčila relační databáze MySQL [15], ale lze použít i jinou vhodnou databázi.

Sběr dat proběhl v období ledna 2015. Tabulka 2.2 zobrazuje některé údaje o získaných datech.

Počet uživatelů	40 603
Počet uživatelů s alespoň 200 příspěvky	14 897
Počet českých příspěvků	2 113 560

Tabulka 2.2: Informace o české datové sadě

2.3.3 Anglická datová sada

Oproti českému jazyku, Twitter podporuje vyhledávání anglických zpráv prostřednictvím svého pokročilého vyhledávání [11]. Díky tomu při dolování anglického Twitteru dochází pouze ke sledování nových příspěvků v anglickém jazyce a ukládání autorů příspěvků do fronty. Jakmile je fronta dostatečně velká, provedeme filtraci, jelikož musíme ověřit, že uživatel skutečně přispívá v anglickém jazyce a zachycený příspěvek nebyl pouze výjimkou. Pokud prvních 20 nejnovějších zpráv uživatele obsahuje alespoň 25 % anglicky psaných zpráv, pak lze uživatele zařadit do množiny anglických uživatelů. Poté jsou prostřednictvím Twitter API získána data o uživatelích a jejich příspěvky.

Sběr dat proběhl na přelomu ledna a února 2015. Tabulka 2.3 zobrazuje některé údaje o získaných datech.

Počet uživatelů	27 164
Počet uživatelů s alespoň 200 příspěvky	25 634
Počet anglických příspěvků	4 671 938

Tabulka 2.3: Informace o anglické datové sadě

3 Anomální uživatelé a existující metody pro jejich detekci

V první části se kapitola zabývá definicí pojmu anomální uživatel a motivací pro rozpoznání anomálních uživatelů. V druhé části jsou pak představeny existující přístupy pro vyhledávání různých druhů anomálních uživatelů na sociální síti Twitter.

3.1 Kdo je to anomální a normální uživatel?

Abychom měli dostatečnou představu, jaký uživatel je anomální a jak anomální uživatelé vypadají, je potřeba si ujasnit, co pod pojmem anomální uživatel rozumíme. Anomálním uživatelem označujeme uživatele, který se chová extrémně v některých svých atributech chování, existuje zde podezření, že takový účet není spravován člověkem, ale je určitým způsobem automatizován nebo šíří v rámci sítě informačně prázdné zprávy, duplicity nebo naopak šíří nadmíru zpráv založených na základě nějakého vzorce.

Zpravidla se tak jedná o tyto skupiny uživatelů:

- Spam účty a infikované účty
- Robotické účty a automaticky spravované účty
- Informační a propagační účty (RSS čtečky apod.) s nulovou interakcí s okolím
- Svým charakterem jednosměrně tématicky zaměřené a ostatní zvláštní účty

V opačném případě chápeme jako normální uživatele všechny ostatní účty. Jsou to uživatelé, kteří zpravidla vykazují interakci se svým okolím, jejich profil není výrazně jednosměrně orientován a jejich jednání je přirozené. Do této skupiny mimo jiné patří:

- Osobní účty projevující interakci s okolím
- Účty firem, společností a internetových služeb poskytující servis svým uživatelům nebo projevující interakci s okolím
- Informační účty komunikující se svými odběrateli
- Celebrity a významné osobnosti

3.2 Motivace pro hledání anomálních uživatelů

Anomální uživatele na sociální síti Twitter hledáme z několika důvodů. Prvním z nich je snaha o jejich studium a pochopení jejich chování, což nám může pomoci při jejich klasifikaci do jednotlivých podskupin, pakliže je naše motivace tyto podskupiny identifikovat. Dále někteří anomální uživatelé mohou porušovat obecné podmínky užívání služby a jejich rozpoznáním můžeme takové uživatelské účty pozastavit nebo trvale zablokovat.

Sociální síť Twitter se sama snaží některé druhy anomálních uživatelů vyhledávat a blokovat, jelikož v rámci sítě mohou šířit nebezpečné odkazy nebo nevyžádané zprávy obtěžující normální uživatele [12]. Další motivací Twitteru je také snaha zobrazovat placenou reklamu svých inzerentů relevantním uživatelům. Pakliže by taková reklama byla zobrazována účtům, které nejsou spravovány člověkem, mohlo by to zásadně ovlivnit výsledky marketingové kampaně a její finanční návratnost. Dalším důležitým aspektem pro inzerenty je, aby v síti byli aktivní reální uživatelé a ne pouze automatizované účty, jelikož právě reální uživatelé mají potencionální kupní sílu a sám Twitter vnímá tento stav jako riziko [16].

Další motivací je pak využití těchto metod k získání dat s možnosti odfiltrování vlivu anomálních uživatelů. Twitter disponuje obrovským množstvím dat, které různé organizace a nástroje využívají k trénování svých modelů při klasifikaci a jiných činnostech, přičemž vliv dat produkovaných anomálními uživateli může mít negativní vliv na výsledky těchto nástrojů.

3.3 Existující metody

Nyní si představíme některé z metod pro detekci dvou častých typů anomálních uživatelů a to robotů a spam účtů.

3.3.1 Detekce robotických účtů s využitím klasifikace

Cílem metody pro detekci robotických účtů na Twitteru [17] je rozpoznání uživatelů na základě jejich míry automatizovaného chování, přičemž metoda se vždy snaží označit uživatele jako člověka, kyborga nebo robota.

Za účet spravovaný člověkem je považován takový účet, kde se nevyužívají žádné automatizační prostředky. Za tzv. kyborga je považován takový účet, který je spravován jak člověkem, tak automatizačním prostředkem. Typicky se tak jedná o RSS čtečky, kde člověk vybírá konkrétní obsah do

fronty automatického správce účtu nebo o účty, které jednají zcela automaticky, ale do jejichž interakce s okolím se v menší míře zapojuje člověk. Poslední skupinou jsou tzv. roboti, kteří jsou spravováni zcela bez pomoci člověka.

Celý systém pro detekci chování uživatelského účtu se skládá ze tří nezávislých komponent propojených s komponentou provádějící výsledné rozhodnutí o zařazení uživatele do konkrétní třídy. První komponenta je určena pro detekci automatizace a periodického chování daného účtu a na základě sekvence časů vložení jednotlivých příspěvků analyzovaného účtu provádí výpočet hodnotícího koeficientu charakterizujícího, jak periodicky se účet chová. Druhá komponenta analyzuje obsah příspěvků uživatele se zaměřením se na detekci nevyžádaných příspěvků pomoci klasifikátoru vycházejícího z Bayesovské klasifikační metody [18]. Poslední komponenta pak extrahuje informace o uživatelském účtu. Mezi tyto informace patří počet příspěvků, počet závadných URL adres, kde závadná adresa je rozpoznána vyhledáním v databázi závadných adres, nebo informace o počtu sledovaných uživatelů a odběratelů a jejich poměr.

Informace ze všech komponent jsou dále odeslány komponentě provádějící výsledné rozhodnutí o přiřazení třídy. Zde se pro rozhodnutí využívá metody LDA[1] [17].

Metoda dosahuje úspěšnosti (neboli celkové správnosti [8]) 90,5 % při použití trénovací množiny o velikosti 3 000 uživatelů, kde každá ze tři tříd má rovnoměrné zastoupení. Jako testovací množina byla autory pro vyhodnocení použita množina o stejné velikosti se stejným podílem jednotlivých tříd. Uživatelé zařazeni do testovací množiny nejsou součástí trénovací množiny [17].

3.3.2 Detekce spam účtů s využitím klasifikace

Detekce spam účtů s využitím klasifikace je přístup založený na jednoduché myšlence, kdy se na základě trénovací množiny složené z normálních a závadných účtů naučíme klasifikátor, který poté můžeme použit k samotné detekci. Důležité je u metody určit, jaké informace o účtu zařadíme do trénovací množiny a jaký typ klasifikátoru použijeme.

Metoda představená v práci [19] vychází zcela z této myšlenky. O každém uživatelském účtu si uchovává informace o počtu odběratelů, počtu

1. Linear Discriminant Analysis

sledovaných uživatelů, jejich poměru, počtu URL odkazů, klíčových slov a zmínek v příspěvcích, o počtu odpovědí a převzetí příspěvků a také bere v potaz aktivitu během dne, která spočívá v evidenci počtu příspěvků spadajících do konkrétní hodiny dne. Informace jsou vždy odvozeny ze 100 posledních příspěvků uživatele. Naopak metoda nebere příliš v potaz slovník obsažený v příspěvcích uživatele a zaměřuje se především na metriky uživatelského účtu.

Autory byly použity učící algoritmy NaiveBayes, SMO, RandomForest a IBk implementované v programu Weka [20]. Při vyhodnocení byla využita křížová validace na množině obsahující 742 normálních uživatelů a 258 spam účtů. Nejlepších výsledků dosahuje klasifikátor RandomForest, kde jeho přesnost dosahuje 95,7 % [19].

3.3.3 Detekce spam účtů seskupením uživatelů

Metoda hledání spam účtu pomoci seskupení uživatelů v sociálních kampaních [21] předpokládá, že více závadných účtů se sklonem šířit nevyžádané nebo závadné zprávy je ovládáno jediným subjektem za účelem šíření propagační kampaně s co největším zásahem, co do počtu oslovených uživatelů v síti. Nevyžádané kampaně tak mají charakter komerčních kampaní šířících prodejní odkazy propagující různé produkty nebo směřující na závadné webové stránky obsahující nebezpečný software, nabídku podezřelých produktů apod.

Oproti předešlým metodám, které se zpravidla zaměřují na analýzu samotného uživatele, je tato metoda založena na seskupování uživatelů v sociálních kampaních. Sociální kampaň je vektor tří složek, kde první složka obsahuje URL odkaz, druhá složka množinu zpráv obsahujících ve svém těle URL odkaz z první složky vektoru a třetí složka množinu všech uživatelů, kteří jsou autory těchto zpráv. Metoda tedy zařazuje do jedné sociální kampaně takové uživatele, kteří publikují zprávu obsahující stejný URL odkaz a URL odkaz je zde chápán jako jedinečný identifikátor kampaně.

Pro každou kampaň jsou pak odvozeny jednotlivé atributy a celý problém se tak redukuje na klasifikaci kampaně na normální a závadnou. Mezi informace, které jsou odvozeny z kampaně, patří kupříkladu faktor rozmanitosti uživatelské množiny kampaně, který se vypočítá jako podíl celkového počtu zpráv v kampani s počtem uživatelů. Dalšími informacemi pak jsou počet závadných slov obsažených ve zprávách, které jsou detekovány na základě předdefinovaného listu, průměrný počet použitých klíčových slov

a zmínění uživatelů v příspěvcích, koeficient určující, s jakou mírou náhodnosti jsou zprávy vkládány, počet přesměrování z URL odkazu kampaně k cílovému odkazu nebo také celkový počet zpráv a další.

Autoři metody pro vyhodnocení vyzkoušeli několik klasifikačních metod. Mezi ně zařadili RandomForest, DecisionTable, SMO, RandomTree a další metody implementované v programu Weka. Trénovací množina obsahovala 744 závadných a 580 normálních kampaní. Kampaním byla třída přiřazena manuálně na základě automatického předvýběru. Ten spočíval kupříkladu ve vyhledávání potencionálně závadných kampaní dle počtu výskytů URL odkazu kampaně v databázích spravujících seznamy obsahující závadné odkazy a v dalších kritériích.

Pro vyhodnocení byla použita metoda křížové validace s nastavením testovacího vzorku o velikosti 10 %. Nejlepší úspěšnosti 94,5 % dosahuje klasifikační metoda RandomForest [21].

3.3.4 Detekce spam účtů metodou shlukování

Metoda detekce spam účtů shlukováním [22] spočívá v rozpoznání uživatelských účtů generujících velké množství příspěvků v kombinaci s nevyžádaným nebo závadným obsahem s důrazem na jejich šíření v síti. Pro detekci není využíván přístup založený na tradičních klasifikátorech tak, jak je tomu kupříkladu u metody v podkapitole 3.3.2, ale využívá se přístupu shlukování v kombinaci s hledáním odlehlých bodů.

O každém uživateli si metoda eviduje informace o jeho profilu a příspěvcích. Celkově se tak jedná o 107 atributů, přičemž prvních 12 atributů se zaměřuje na vlastnosti uživatele. Mezi tyto vlastnosti patří počet sledovaných uživatelů účtem a počet odběratelů, celkový počet příspěvků a počet převzatých příspěvků, stáří účtu, počet URL odkazů a klíčových slov a další. Druhá skupina 95 atributů jsou atributy reprezentující počty výskytů jednotlivých znaků v příspěvcích uživatele z ASCII tabulky standardní americké klávesnice.

Pro samotnou detekci pak metoda uvažuje přístupy založené na algoritmech DenStream a StreamKM++ [22]. V obou případech se jedná o algoritmy určené ke shlukování, přičemž DenStream je inkrementální verze algoritmu DBSCAN [23], který hledá shluky na základě hustoty instancí v prostoru a je možné u tohoto algoritmu nastavit, jaký minimální počet prvků musí každý nalezený shluk obsahovat a jak od sebe mohou být prvky shluku maximálně vzdálené. Naopak StreamKM++ je inkrementální verze

algoritmu k-means++ [24], který hledá předdefinovaný počet shluků a umísťuje iniciální centroidy v prostoru tak, aby byly od sebe dostatečně vzdálené.

V práci se provádí modifikace algoritmů DenStream a StreamKM++ a ta spočívá především v rozšíření schopnosti algoritmů klasifikovat příchozí instance jako normální nebo závadné. U obou algoritmů se v první fázi vytvoří model shluků nad trénovací množinou o velikosti 300 normálních uživatelů. Poté jsou příchozí instance algoritmy zařazovány do nejbližších shluků. Jestliže vzdálenost instance od nejbližšího shluku přesahuje definovaný práh a je tedy odlehlou, pak je taková instance považována za závadnou.

Při ověřování výsledků byla autory použita trénovací množina o velikosti 1 483 normálních účtů a 104 závadných účtů. Testovací množina disponovala počtem 1 548 normálních účtů a 104 závadných účtů. Při vyhodnocení nejlepších výsledků dosahovala kombinace obou algoritmů, kdy je nejprve spuštěn StreamKM+ a výsledný model ohodnocených instancí je pak předán algoritmu DenStream, který považuje normálně ohodnocené instance za normální. Kombinace algoritmů při takovém nastavení dosahuje úspěšnosti 98 % [22].

4 Návrh metody pro rozpoznání anomálních uživatelů

V rámci kapitoly návrhu metody pro detekci anomálních uživatelů jsou na začátku představeny informace, které jsou o uživateli uchovávány. Dále následuje vysvětlení metodiky pro anotaci uživatelů a výběr podmnožiny atributů, které jsou nakonec z informací o uživateli použity. Dále následují informace o předzpracování textových dat. V závěru kapitoly jsou představeny jednotlivé testované přístupy s popisem použitých dat a jejich výsledky, přičemž celou kapitolu zakončuje shrnutí.

4.1 Informace o uživateli

Základním předpokladem pro úspěšný návrh metody pro detekci anomálních uživatelů je získání vhodných informací o uživateli. Proto si představíme informace, které byly na základě analýzy uživatelského profilu uvažovány při návrhu metody. Informace lze rozdělit do tří skupin a to na skupinu atributů zabývajících se chováním uživatele, dále skupiny atributů odvozených z jeho časové aktivity a na závěr skupiny atributů ze slovníku uživatele.

Pokud bylo obdobné použití navrženého atributu nalezeno v jiných pracích, jsou v popisu atributu uvedeny odkazy na zdroje.

4.1.1 Chování uživatele a informace o jeho účtu

První skupina atributů se zaměřuje na chování uživatele. Patří zde informace o tom, kolik odběratelů uživatel má, kolik jiných uživatelů sleduje, jak často vkládá do svých příspěvků URL odkazy, klíčová slova nebo zmínky jiných uživatelů a další atributy.

Počet odběratelů Počet odběratelů je atribut udávající počet uživatelů, kteří daného uživatele sledují. Dá se předpokládat, že celebrity budou mít vysoký počet odběratelů, běžní uživatele menší počet a nezajímavé účty téměř žádného. Na druhou stranu, tato informace může být zavádějící, jelikož síť závadných účtu může být sobě sama vzájemnými odběrateli nebo zde může být snaha ze strany účtu využít principu „sleduj zpátky“[1]. Prin-

1. Často označováno jako „follow back“

cip využívá předpokladu, kdy sledovanému uživateli je ze strany Twitteru odeslána notifikace o novém odběrateli a tento uživatel tuto vazbu opětuje. Poté může uživatel vazbu zrušit, Twitter o této změně uživatele neinformuje a účet tak získá odběratele na základě tohoto principu. V důsledku toho nemusí být atribut tak vypovídající. Atribut je často používán i v jiných pracích a byl použit v [17, 19, 22].

Počet sledovaných Počet sledovaných udává informaci o počtu uživatelů, který účet sleduje. Celebrity tedy sledují zpravidla menší počet uživatelů, naopak některé účty za účelem využití principu „sleduj zpátky“ sledují větší počet uživatelů. Použit v [17, 19, 22].

Reputace Reputace je odvozená metrika, kterou lze spočítat vzorcem 4.1. Nabývá hodnoty 1 tehdy, pakliže uživatel má pouze odběratele, ale nesleduje žádné ostatní uživatele, což se může blížit chování celebrit, které mají vysoký počet odběratelů. Naopak metrika se blíží k hodnotě 0, pakliže uživatel sleduje větší počet uživatelů, než sledují jeho samého. Použit v [19, 22].

$$\frac{\textit{počet odběratelů}}{\textit{počet odběratelů} + \textit{počet sledovaných}} \tag{4.1}$$

Počet příspěvků Počet příspěvků udává celkový počet zpráv napsaných uživatelem za celou jeho historii existence. V úvahu připadá také použití relativní varianty atributu, která by se spočítala jako podíl se stářím účtu uživatele. Je ale nutné brát na zřetel, že uživatelé mohou ze své historie jednotlivé příspěvky mazat, proto se zdá být lepší vycházet z časových informací, konkrétně z intervalu mezi posledními vkládanými příspěvky, než z informace o stáří uživatele. Dále je nutné také uvažovat, že účet se během své existence postupně vyvíjí a má různé trendy chování, proto obecně průměrovat atributy hodnotou stáří uživatele nemusí být vhodné. Použit v [22].

Počet oblíbených Uživatelé mohou označit příspěvky jiných uživatelů jako oblíbené. Počet oblíbených tedy udává celkový počet oblíbených příspěvků uživatele. Použit v [22].

Počet pozitivních příspěvků Počet pozitivních příspěvků reprezentuje počet zpráv obsahujících pozitivní emotikon. Seznam pozitivních emotikonů je obsažen v tabulce 4.1.

Třída	Emotikony
Pozitivní	:), :-), :o), :-o), :D, :-D, XD, ;), ;-), ;o), ;-o), ;D, ;-D
Negativní	:(, :-(, :o(, :-o(, :,-(, :'-(, ;(, :-(, ;o(, ;-o(, ;,-(, ;'-(

Tabulka 4.1: Emotikony dle sentimentu

Počet negativních příspěvků Počet negativních příspěvků uvádí počet zpráv obsahujících negativní emotikon. Seznam negativních emotikonů je obsažen v tabulce 4.1.

Počet neutrálních příspěvků Počet neutrálních příspěvků uvádí počet zpráv neobsahujících pozitivní nebo negativní emotikon. Seznam pozitivních a negativních emotikonů je obsažen v tabulce 4.1.

Počet zmínek uživatelů Každý uživatel může ve své zprávě zmínit jiného uživatele použitím symbolu @ a uvedením jeho uživatelského jména. Počet zmínek uživatelů uvádí celkový počet takových zmínění. Použit v [19].

Počet klíčových slov Každý uživatel může ve své zprávě používat klíčová slova použitím symbolu # a uvedením názvu. Počet použitých klíčových slov uvádí celkový počet použití tohoto nástroje. Použit v [19].

Počet URL odkazů Každý uživatel může ve své zprávě odkazovat prostřednictvím URL odkazů. Počet URL odkazů uvádí celkový počet odkazů obsažených v příspěvcích uživatele. Použit v [17, 19, 22].

Počet médií Každý uživatel může do svých zpráv vkládat multimediální obsah, tj. obrázky, videa a další. Počet médií uvádí celkový počet multimediálních objektů obsažených ve zprávách uživatele.

Počet převzatých příspěvků Počet převzatých příspěvků reprezentuje počet zpráv, které uživatel převzal od jiného uživatele, aby je šířil svým odběratelům. Použit v [19, 22].

Počet odpovědí Na příspěvky jiných uživatelů lze reagovat. Počet odpovědí udává počet příspěvků, které byly reakcí na zprávy jiných uživatelů. Použit v [19].

Počet oblíbených příspěvků Udává počet příspěvků uživatele, které byly označené alespoň jedním odběratelem jako oblíbené.

Celkový počet převzetí Udává celkový počet odběratelů, kteří převzali příspěvky uživatele.

Celková oblíbenost příspěvků Udává celkový počet odběratelů, kteří zařadili příspěvky uživatele mezi své oblíbené.

Počet otazníků Udává celkový počet použití znaku otazníku ve zprávách uživatele.

Počet vykřičníků Udává celkový počet použití znaku vykřičníku ve zprávách uživatele.

Počet unikátních slov Udává celkový počet unikátních výskytů slov obsažených v příspěvcích uživatele.

Počet unikátních filtrovaných slov Udává celkový počet unikátních výskytů slov obsažených v příspěvcích uživatele, přičemž nejsou započítaná slova obsahující pouze číslice.

Směrodatná odchylka délky příspěvků Udává směrodatnou odchylku mezi délkami zpráv uživatele.

Směrodatná odchylka zaokrouhlené délky příspěvků Udává směrodatnou odchylku mezi délkami zpráv uživatele, kde délka zprávy je zaokrouhlena na desítky.

4.1.2 Časové informace

Druhá skupina atributů se zaměřuje na informace, jak často a v jaké době uživatel přispívá. Zde především patří informace o distribucí příspěvků během denního cyklu nebo jak se mění intervaly, v jakých uživatel nové příspěvky vkládá.

Intervaly mezi příspěvky Intervaly mezi příspěvky umožňují sledovat několik odvozených statistik, mezi které patří minimum, maximum, průměr, rozptyl a směrodatná odchylka. Minimum nám může naznačit, v jakém minimálním intervalu uživatel vkládá své příspěvky. Průměr naopak může naznačit, s jakou frekvencí bude uživatel aktivní a směrodatná odchylka nám může odhalit, jak se liší jednotlivé intervaly mezi vkládanými příspěvky. Pokud je směrodatná odchylka nízká, může to znamenat, že uživatel přispívá v pravidelných intervalech nebo automatizovaně. Naopak pokud je odchylka vysoká, může to naznačovat, že přispívání není pravidelné.

Aktivita během dne Skupina atributů vycházejících z denní aktivity uživatele využívá intuitivního předpokladu rozdílnosti aktivity člověka a robota během dne. Uveďme si příklad. Normální uživatel zpravidla vkládá příspěvky dle jeho denního programu. Může přispívat od rána do večera, ale během noci, kdy zpravidla spí, pravděpodobně přispívat nebude. Naopak robotické účty se intuitivně takto nechovají. Mohou přispívat v pravidelných intervalech a nejsou omezeni denní dobou. Pakliže přispívají v hodinových intervalech, je jejich aktivita rovnoměrná a bude zaznamenána i v nočních a ranních hodinách. Pokud přispívají v denních intervalech, jsou všechny zprávy soustředěny pouze v jedné hodině dne.

Aktivita během dne je tedy skupina atributů, kde každý atribut reprezentuje počet příspěvků v konkrétní denní hodinu a zároveň nad touto aktivitou jsou počítány statistiky minima, maxima, průměru, rozptylu a směrodatné odchylky.

Myšlenka použití tohoto atributu vychází z práce [19], zde ale nejsou dále počítány další statistiky jako rozptyl apod.

4.1.3 Obsah příspěvků

Poslední skupina atributů se zabývá obsahem příspěvků uživatele a jeho slovníkem.

Dokument uživatele Dokument uživatele obsahuje posledních 200 zpráv uživatele. Zprávy jsou před vložením do dokumentu uživatele předzpracovány (viz. podkapitola 4.3) a poté je nad celým dokumentem provedena tokenizace[2], přičemž každé užité slovo se stává atributem, jehož hodnota reprezentuje počet výskytů slova.

4.2 Anotace uživatelů

Abychom mohli provádět vyhodnocování úspěšnosti různých přístupů při návrhu metody pro detekci anomálních uživatelů, byla provedena anotace některých uživatelů na základě požadavků plynoucích z tvorby trénovacích, testovacích množin nebo vyhodnocení výstupu konkrétních algoritmů.

Anotace uživatelů proběhla na základě hodnocení dvou anotátorů pomoci vlastního nástroje popsaného v příloze A. Pokud při anotaci došlo u uživatele ke shodě anotátorů, pak byla uživateli výsledná třída přiřazena. Pokud došlo k jejich neshodě, o třídě rozhodl třetí anotátor.

Tabulka 4.2 ukazuje shodu anotátorů při anotaci. Celková čistá shoda je 95,5 %, přičemž Cohenovo kappa [25] dosahuje hodnoty 0,907, což značí velmi dobrou shodu a relevantnost anotace. Cohenovo kappa nabývá hodnoty jedna při úplném souhlasu mezi anotátory, naopak hodnoty nula nabývá právě když shoda mezi anotátory odpovídá náhodné shodě. Výjimečně může koeficient nabývat i záporné hodnoty a v tom případě je shoda mezi anotátory nižší než očekávaná. Hodnota koeficientu vyšší než 0,6 je považována za dobrou shodu a hodnota vyšší než 0,8 za velmi dobrou [25].

Tabulka 6.3 ukazuje shodu anotátorů pro anglické uživatele. Zde je čistá shoda 94 % a Cohenovo kappa je rovno hodnotě 0,881 a lze tedy také provedenou anotaci považovat za velmi dobrou.

		Anotátor B	
		Normální	Anomální
Anotátor A	Normální	154	12
	Anomální	6	226

Tabulka 4.2: Anotace českých uživatelů

2. Rozčlenění textu na samostatné jednotky (zpravidla slova, dvojice po sobě následujících slov apod.)

		Anotátor B	
		Normální	Anomální
Anotátor A	Normální	138	11
	Anomální	8	164

Tabulka 4.3: Anotace anglických uživatelů

4.3 Předzpracování příspěvků

Jelikož jazyk používaný na sociální síti Twitter má svá specifika, tak u příspěvků uživatele je prováděno předzpracování, jehož cílem je redukce prostoru atributů odvozených z obsahu příspěvků. Nad každým příspěvkem, z kterého se pak skládá dokument zpráv uživatele, se provádí následující metody předzpracování. Některé z použitých metod byly dříve použity i při klasifikaci sentimentu na sociální síti Twitter [26].

- odstranění zmínek uživatelů nahrazením za značku *_USER_*
- odstranění klíčových slov nahrazením za značku *_HASHTAG_*
- odstranění URL odkazů nahrazením za značku *_URL_*
- odstranění odkazů směřujících na přiložená média příspěvku nahrazením za značku *_MEDIA_*
- odstranění diakritiky
- odstranění emotikonů
- odstranění všech znaků a slov obsahujících znaky, které nespadají do intervalu *a-z*, *A-Z* a *0-9*
- převod velkých písmen na malé

Tabulka 4.4 ukazuje, jaký vliv má předzpracování na počet atributů. Hodnoty byly získány jak na české, tak anglické datové sadě. Obě sady obsahovaly 2 000 000 náhodně vybraných příspěvků, aby bylo možné provést srovnání mezi jazyky. Můžeme si povšimnout, že dochází k velké redukci. Jednou z možných příčin většího počtu atributů u české sady po předzpracování je skutečnost, že český jazyk má bohatě rozvinuté skloňování a časování, což zvyšuje velikost prostoru. Další příčinou je také časté vkládání cizojazyčných slov do česky psaných příspěvků.

	Česká sada	Anglická sada
Před předzpracováním	2 918 922	2 963 370
Po předzpracování	631 138	280 369

Tabulka 4.4: Počet sloupců matice atributů pro 2 000 000 příspěvků

4.4 Výběr atributů

Před samotným testováním úspěšnosti jednotlivých přístupů pro rozpoznání anomálních uživatelů je vhodné provést konečný výběr podmnožiny z navržených atributů, jelikož některé mohou k rozlišení anomálních a normálních uživatelů svou informací přispívat více a některé méně nebo vůbec. K tomu, abychom mohli ověřit výběr atributů a případně zúžili konečnou množinu atributů, použijeme tři přístupy. Prvním je využití rozložení hodnot atributů v datech a jejich analýza, druhým přístupem je využití selektivních algoritmů v programu Weka a třetí je manuální výběr založený na pozorování výsledků algoritmů při prováděných experimentech.

4.4.1 Chování atributů

Analýza rozložení hodnot jednotlivých atributů nám umožňuje sledovat chování každého z nich v datech, z čehož lze usuzovat, který atribut nám může poskytnout informace vedoucí k nalezení anomálního uživatele a který naopak ne. Mezi vhodné atributy tak mohou patřit ty, ve kterých anomální uživatelé nabývají odlehlých hodnot oproti normální nebo většinové skupině uživatelů. Díky tomu je pak takový atribut potencionálním kandidátem, který může nést informaci o odlehlosti uživatele. Příkladem tak může být atribut, jehož rozložení hodnot odpovídá normálnímu rozdělení, u kterého lze vysledovat odlehlé hodnoty.

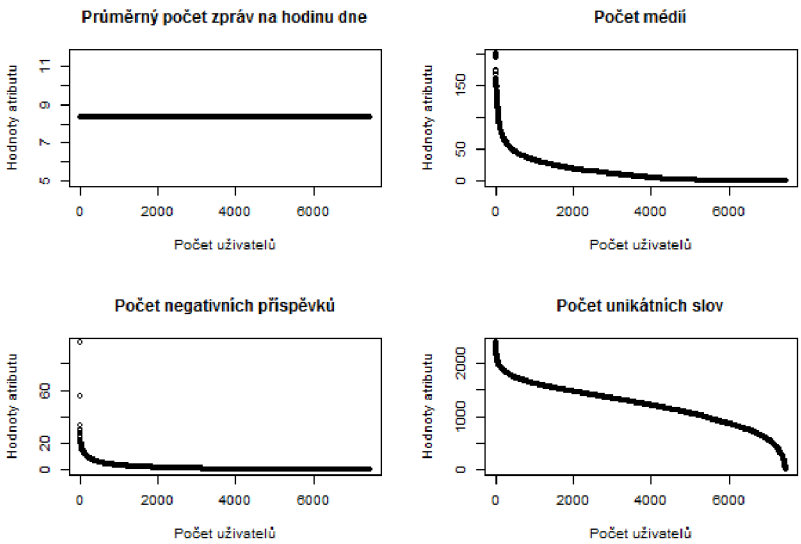
Naopak vlastnosti uživatelů, které jsou napříč vzorkem uživatelů neměnné a nesou spíše rysy rovnoměrného rozdělení, jsou pro hledání anomálních uživatelů spíše nezajímavé, jelikož takový atribut pravděpodobně nenese informaci, která by normálního uživatele odlišila od anomálního. Kupříkladu na obrázku 4.1 můžeme vidět, že průměrný počet zpráv na hodinu dne není rozhodně dobrý atribut. Naopak atributy počet médií, počet negativních příspěvků nebo počet unikátních slov mohou nést informace o tom, že uživatel je normální nebo anomální. Obrázek 4.2 pak zobrazuje

další skupinu atributů, které mohou být vhodné pro rozpoznání anomálních uživatelů.

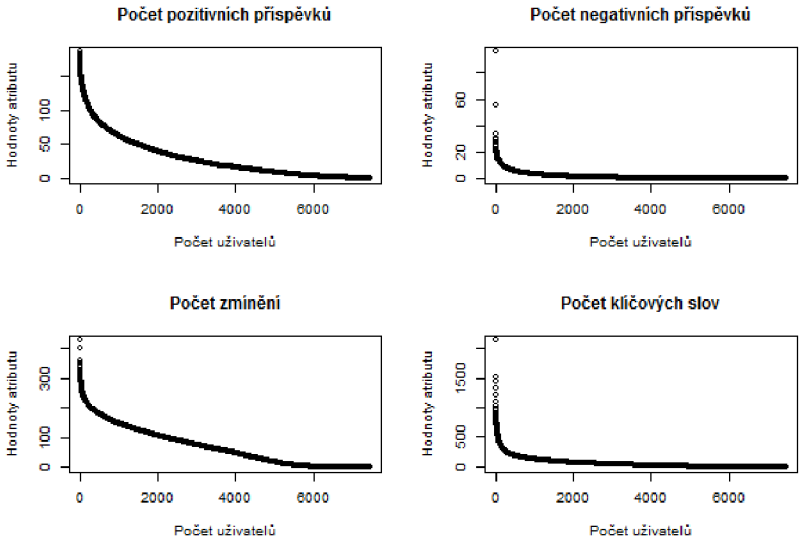
Nemusíme ale používat jen jeden způsob pro zobrazení chování atributů. Obdobným způsobem pomoci histogramů na obrázku 4.3 se můžeme podívat i na další atributy a na obrázku 4.4 naopak můžeme vidět krabicový diagram k zobrazení charakteristik atributů, přičemž můžeme vidět, jakých hodnot dosahuje majoritní skupina uživatelů a sledovat odlehlé hodnoty atributu.

Využitím metody analýzy chování atributů byla předvybrána řada atributů, které pak byly brány v potaz i s ohledem na výsledky dalších postupů při výběru atributů. Mezi tyto atributy patří kupříkladu počet pozitivních příspěvků, počet negativních příspěvků, maximum během denní aktivity nebo její rozptyl. Naopak, díky této metodě byly také nalezeny atributy, které není vhodné dále uvažovat. Jedná se například o průměrný počet zpráv na hodinu dne nebo reputaci.

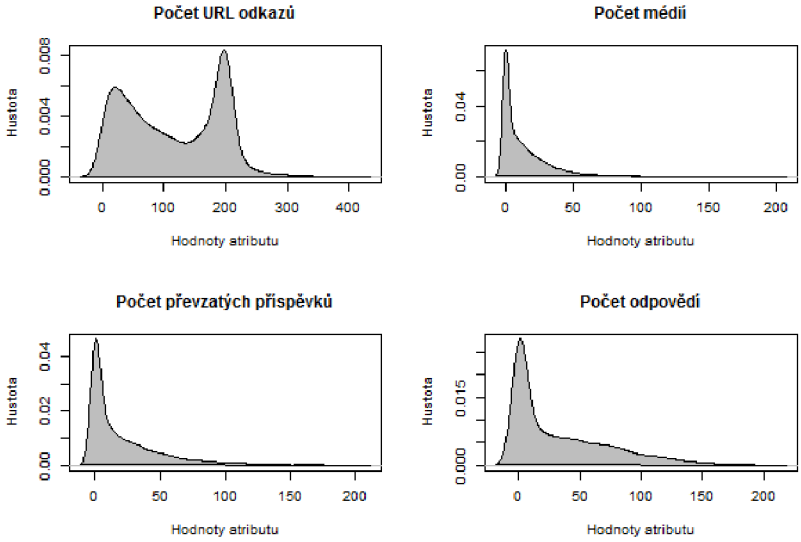
Při analýze chování atributů není explicitně nutné využívat pouze zobrazení využívající rozložení hodnot, histogramy nebo krabicové digramy, ale lze použít také další metody. Vhodné je třeba zmínit přístup paralelních souřadnic [27].



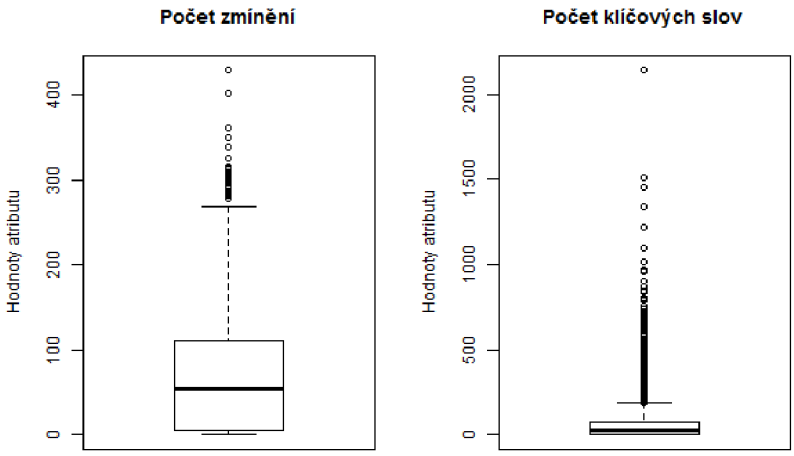
Obrázek 4.1: Distribuce hodnot vybraných atributů v českých datech



Obrázek 4.2: Distribuce hodnot vybraných atributů v českých datech



Obrázek 4.3: Histogramy vybraných atributů v českých datech



Obrázek 4.4: Krabicové diagramy vybraných atributů v českých datech

4.4.2 Výběr pomoci selektivních algoritmů

Další metodou, jak vybrat vhodné atributy, je využití algoritmů pro jejich výběr. Algoritmy zpravidla vrátí na výstup podmnožinu vlastností uživatele nebo tyto vlastnosti ohodnotí dle jejich přínosu pro rozpoznání anomálních a normálních uživatelů. V rámci testování byla použita trénovací množina 100 normálních a 100 anomálních uživatelů náhodně vybraných z množiny anotovaných uživatelů pro každý jazyk. V množině byly obsaženy všechny atributy představené v kapitole 4.1 s výjimkou atributů z dokumentu uživatele, jelikož snahou bylo najít atributy, které mohou být rozhodující pro určení anomality uživatele bez ohledu na jeho dokument. Pro výběr byly použity algoritmy CfsSubsetEval a GainRatioAttributeEval s výchozím nastavením implementované v programu Weka.

Tabulka 4.5 zobrazuje vybrané atributy každým algoritmem u české množiny a tabulka 4.6 zobrazuje atributy pro anglickou množinu. Výstupem algoritmu GainRatioAttributeEval byl ohodnocený seznam všech atributů, ale v tabulce je vždy uvedeno pouze prvních 10 nejlépe ohodnocených. Naopak algoritmus CfsSubsetEval na výstup vrátil podmnožinu vybraných atributů, která je v tabulkách vypsána vždy celá. Navíc u každého atributu vráceného algoritmem CfsSubsetEval, který se nevyskytuje v se-

znamu pro GainRatioAttributeEval, je v závorce doplněno pořadí atributu přiřazené tímto algoritmem v celkovém seznamu.

CfsSubsetEval	GainRatioAttributeEval
Počet klíčových slov	Počet URL odkazů
Počet URL odkazů	Směrodatná odchylka délky příspěvků
Počet odpovědí (13.)	Směrodatná odchylka zaokrouhlené délky příspěvků
Počet otazníků	Směrodatná odchylka denní aktivity
Počet unikátních filtrovaných slov (11.)	Rozptyl denní aktivity
Směrodatná odchylka délky příspěvků	Počet převzatých příspěvků
Průměrný interval mezi příspěvky (30.)	Počet klíčových slov
Rozptyl denní aktivity	Maximum během denní aktivity
	Počet otazníků
	Aktivita během 19. hodiny dne

Tabulka 4.5: Výběr atributů v české množině pomoci selektivních algoritmů

Můžeme si všimnout, že u obou množin dominují atributy týkající se počtu klíčových slov, zmínění a URL odkazů. Dále se zde také objevují atributy referující o časové aktivitě během dne nebo atributy zabývající se délkou příspěvků a počtu unikátních slov v uživatelově slovníku. Důležitou roli se zdá mít i atribut počtu převzatých příspěvků nebo počtu odpovědí.

Zajímavé je si také povšimnout některých rozdílů mezi českou a anglickou množinou, kde kupříkladu atribut celkové oblíbenosti příspěvku se vyskytuje ve výběru pro anglickou množinu, ale ne pro českou. Příčinou je pravděpodobně skutečnost, že používání funkce oblíbené na sociální síti Twitter je v anglickém prostředí mnohem rozšířenější a může mít lepší informační hodnotu, než v českém prostředí, kde nástroj oblíbených příspěvků není v takové míře používán i s ohledem na absolutní velikost uživatelské základny.

CfsSubsetEval	GainRatioAttributeEval
Počet sledovaných (23.)	Počet unikátních filtrovaných slov
Počet klíčových slov	Počet unikátních slov
Počet převzatých příspěvků	Počet převzatých příspěvků
Celková oblíbenost příspěvků	Počet klíčových slov
Počet unikátních filtrovaných slov	Počet médií
Aktivita během 5. hodiny dne (37.)	Celkový počet převzatých příspěvků
Aktivita během 15. hodiny dne (34.)	Počet URL odkazů
Aktivita během 20. hodiny dne (14.)	Počet zmínění uživatelů
	Celková oblíbenost příspěvků
	Počet odpovědí

Tabulka 4.6: Výběr atributů v anglické množině pomoci selektivních algoritmů

4.4.3 Testování a konečný výběr

Při výběru atributů také můžeme využít přístup, kdy během testování jednotlivých metod pro rozpoznání anomálního uživatele můžeme jednotlivě odebírat a přidávat atributy a sledováním získaných výsledků odhadnout jejich konkrétní vliv.

Při testování se ukázalo, že je vhodné použít informace o počtu pozitivních a negativních příspěvků, jelikož právě normální uživatelé mají většinou tendenci používat emoce v příspěvcích oproti automatizovaným účtům. Naopak informace o počtu odběratelů, sledovaných uživatelů nebo reputaci spíše vedou k označení populárních účtů jako anomálních. Vlastnosti o počtu otazníků nebo vykřičníků v příspěvcích uživatele mohou někdy přinést kladné výsledky, ale také vedou k označení některých skupin normálních uživatelů za anomální.

Výběr konečného seznamu atributů proběhl s využitím přístupů popsaných v kapitolách 4.4.1 a 4.4.2, kde získané informace z těchto analýz byly dobrými vodítky při prováděných experimentech, které sledovaly vliv na výsledky metod při testování po přidání nebo odebrání jednotlivých atri-

butů. Experimenty byly prováděny jak na českých i anglických datech tak, aby výsledná skupina atributů byla dostatečně robustní a nezávislá vůči některým odlišnostem české a anglické komunity, jako je třeba míra používání nástroje pro označení oblíbeného příspěvku apod.

Tabulka 4.7 zobrazuje konečný seznam atributů. Jedná se celkem o 48 atributů, které jsou doplněny o atributy odvozené z dokumentu uživatele.

Počet pozitivních příspěvků
Počet negativních příspěvků
Počet zmínek
Počet klíčových slov
Počet URL odkazů
Počet médií
Počet převzatých příspěvků
Počet odpovědí
Unikátní počet filtrovaných slov
Směrodatná odchylka délky příspěvku
Směrodatná odchylka zaokrouhlené délky příspěvku
Maximální denní aktivita
Směrodatná odchylka denní aktivity
Atributy aktivity během 0. až 23. hodiny dne
Atributy z dokumentu uživatele

Tabulka 4.7: Seznam vybraných atributů

4.5 Testované přístupy a jejich výsledky

Vybrané testované metody pro detekci anomálních uživatelů lze zařadit do tří skupin. Do skupiny metod založených na učení bez učitele, učení z klasifikovaných a neklasifikovaných příkladů a na učení s učitelem. Pro každou skupinu metod byla vytvořena množina instancí pro vyhodnocení testovaných metod, které jsou blíže popsány v jednotlivých podkapitolách, přičemž atributy u každého uživatele odpovídají vybraným atributům v podkapitole 4.4.3. Detailní nastavení jednotlivých metod včetně výstupů je součásti přílohy B .

4.5.1 Učení bez učitele

Metody založené na učení bez učitele nevyžadují roztřídění instancí v trénovací množině do jednotlivých tříd. Základní myšlenka metody spočívá v měření vzájemné podobnosti jednotlivých instancí a výpočtu koeficientu, který ohodnocuje odlehlost dané instance od ostatních.

Pro účely detekce anomálních uživatelů bylo testováno využití dvou metod LOF[3] [28] a OR[4] implementovaných v balíčku DMwR [29] v jazyce R.

Pro vyhodnocení úspěšnosti metod založených na učení bez učitele byla vytvořena množina 6 000 náhodně vybraných uživatelů pro český a anglický Twitter. Hodnota parametru nejbližšího okolí byla u LOF nastavena na hodnotu 50, jelikož při experimentech dosahovala lepších výsledků než nižší hodnoty a naopak vyšší hodnoty vedly k prodloužení doby potřebné pro výpočet bez výrazného vlivu na výsledek. Výsledné ohodnocené instance algoritmem byly sestupně seřazeny dle přiřazeného koeficientu a vyhodnocení bylo podrobeno 100 prvních instancí, které dle grafu dosahují nejvyššího koeficientu (zhruba odpovídá koeficientu 2 a více). Graf pro českou množinu se nachází na obrázku 4.5 a pro anglickou na obrázku 4.6. Tokenizace při předzpracování byla provedena na unigramy a počet tokenů byl vždy omezen na 2 500 nejfrekventovanějších pro každou třídu.

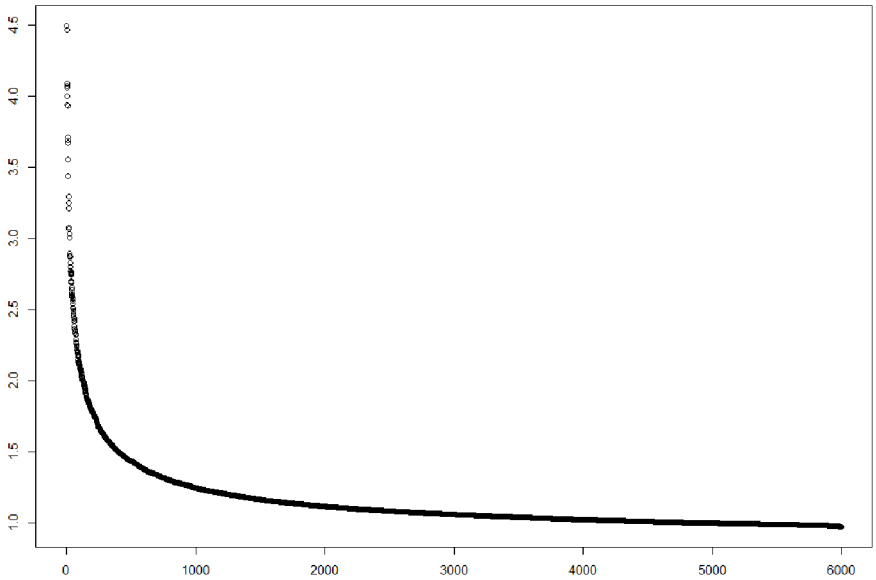
U OR se postupovalo obdobným způsobem. Na základě sestupného pořadí ohodnocených instancí bylo vyhodnocení podrobeno také 100 prvních uživatelů, kteří dosahovali nejvyššího koeficientu (zhruba odpovídá koeficientu 0,8 a více). Graf pro českou množinu lze vidět na obrázku 4.7 a pro anglickou množinu na obrázku 4.8.

Výsledná úspěšnost obou metod pro obě množiny je zobrazena v tabulce 4.8, kde můžeme vidět, že LOF dosahuje úspěšnosti až 98 % a OR až 92 %. Úspěšností v tomto případě rozumíme, jaký podíl instancí podrobených vyhodnocení bylo anotátory označeno jako anomální.

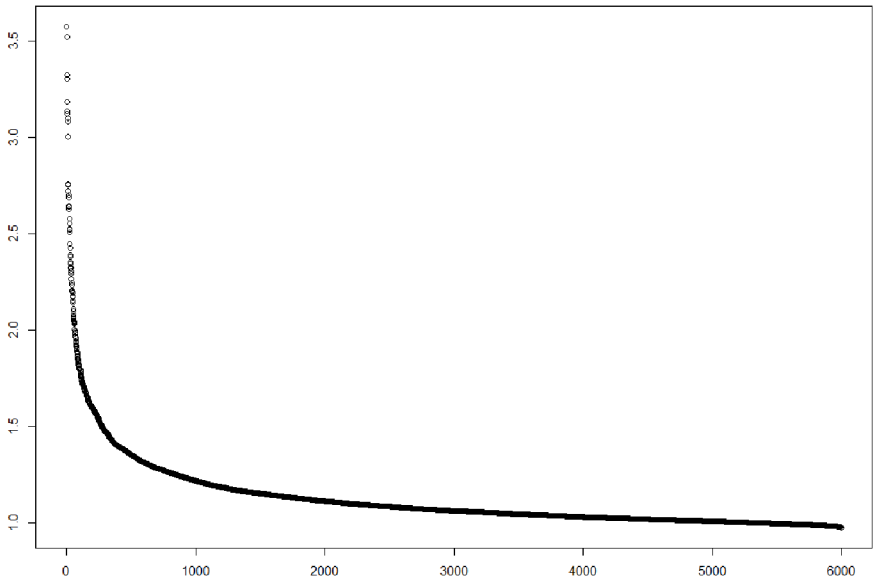
	Česká množina	Anglická množina
LOF	98 %	97 %
OR	92 %	89 %

Tabulka 4.8: Výsledky metod LOF a OR pro českou a anglickou množinu

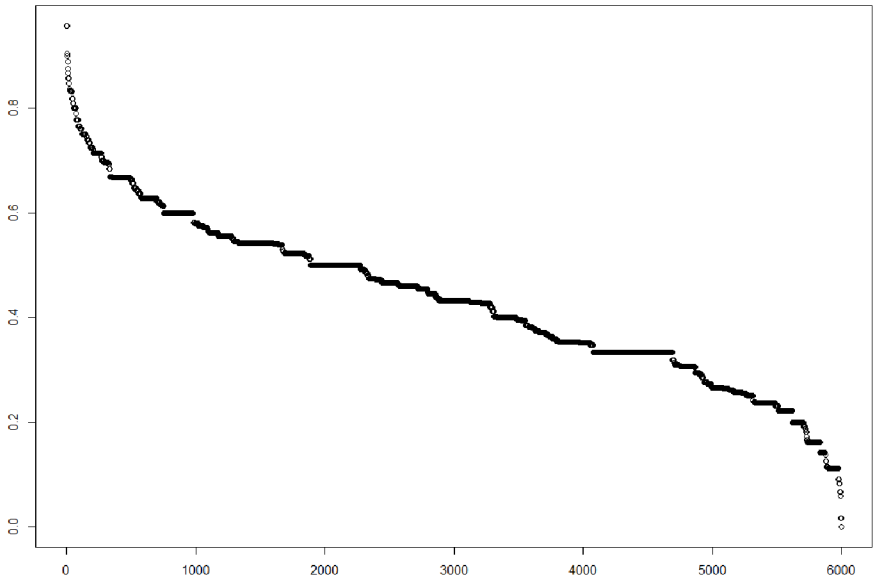
3. Local outlier factor
4. Outlier ranking



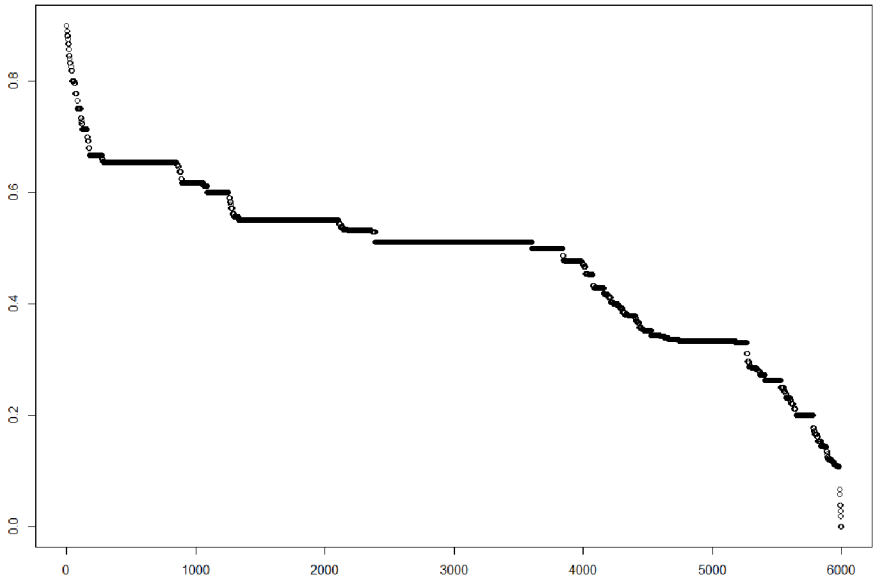
Obrázek 4.5: Distribuce koeficientu odlehlosti u LOF pro českou množinu



Obrázek 4.6: Distribuce koeficientu odlehlosti u LOF pro anglickou množinu



Obrázek 4.7: Distribuce koeficientu odlehlosti u OR pro českou množinu



Obrázek 4.8: Distribuce koeficientu odlehlosti u OR pro anglickou množinu

4.5.2 Učení z klasifikovaných a neklasifikovaných příkladů

Přístup učení z klasifikovaných a neklasifikovaných příkladů je založen na principu, kdy se klasifikátor učí na trénovací množině obsahující malý počet příkladů s přiřazenou třídou a příklady s nepřiřazenou třídou a poté klasifikuje jednotlivé instance do daných tříd.

Pro účely našeho testování byly testovány 3 metody obsažené v rozšiřujícím balíčku CollectiveClassification [30] pro program Weka a to YATSI[5] [31] v kombinaci s NaiveBayesMultinomial (dále YATSI-NBM), YATSI v kombinaci s RandomForest [32] (dále YATSI-RF) a CollectiveForest [31] (dále CF).

YATSI je metoda založená na principu, kdy se použitý klasifikátor (v našem případě NBM nebo RF) naučí na příkladech s přiřazenou třídou z trénovací množiny model, poté klasifikuje instance s nepřiřazenou třídou z trénovací množiny a použitím metody hledání nejbližších sousedů instancím přiřadí výslednou třídu. CF naopak využívá klasifikátor RandomForest tak, že se nejprve naučí model na příkladech s přiřazenou třídou a pak klasifikuje instance bez přiřazené třídy. Následně si vytvoří novou trénovací množinu, do které přidá ty instance, kde si byl algoritmus při predikci instancí bez přiřazených tříd nejvíce jistý svou správností. Celý proces se pak opakuje, kdy klasifikátor s novou trénovací množinou postupně klasifikuje zbývající příklady bez přiřazené třídy [31].

Pro vyhodnocení byla vytvořena trénovací množina obsahující 30 anomálních a 30 normálních uživatelů, neoznačkovaná množina obsahující 120 normálních a 120 anomálních uživatelů, přičemž ta samá množina byla poté použita jako testovací, jelikož implementace algoritmů YATSI a CF v programu Weka vyžadují, aby testované instance byly obsaženy v trénovacích datech. Při testování byly použity různé způsoby tokenizace a počet tokenů byl vždy omezen na 3 000 nejfrekventovanějších pro každou třídu.

Vyhodnocení výsledků pro českou množinu je obsaženo v tabulce 4.9. Můžeme vidět, že nejlepších výsledků dosahuje metoda CF, která dosahuje úspěšnosti 84,6 %. Naopak CF pro anglickou množinu dle výsledků v tabulce 4.10 dosahuje úspěšnosti až 90,8 %. V obou případech baseline[6] nabývá hodnoty 50 %.

5. Yet Another Two Stage Idea
6. Klasifikace do majoritní třídy

4.5.3 Učení s učitelem

Skupina metod založená na přístupu učení s učitelem vychází z principu, kdy se klasifikátor učí na trénovací množině obsahující příklady s přiřazenou třídou a poté klasifikuje jednotlivé instance do daných tříd.

Při testování bylo použito 5 metod implementovaných v programu Weka a to NaiveBayesMultinomial (dále NBM), Support Vector Machine [33] (dále SVM), Random Forest (dále RF), J48 a metoda majoritně hlasujících klasifikátorů složených z NBM, RF a J48 (dále Vote).

Dále byla pro vyhodnocení vytvořena trénovací množina obsahující 100 anomálních a 100 normálních uživatelů. Pro vyhodnocení byla také vytvořena testovací množina o velikosti 50 anomálních a 50 normálních uživatelů. Při testování byly použity různé způsoby tokenizace a počet tokenů byl vždy omezen na 3 000 nejfrekventovanějších pro každou třídu. Baseline je rovna hodnotě 50 %.

Vyhodnocení výsledků pro českou množinu je obsaženo v tabulce 4.11. Můžeme vidět, že nejlepších výsledků dosahuje majoritně hlasující klasifikátor, který dosahuje úspěšnosti 89 %. Podobné úspěšnosti dosahuje také klasifikátor RF a J48.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	YATSI-NBM	59,6 %	74,0 %	52,4 %
	YATSI-RF	57,5 %	74,5 %	48,6 %
	CF	84,6 %	84,7 %	84,6 %
Bigramy	YATSI-NBM	52,5 %	56,1 %	44,3 %
	YATSI-RF	50,4 %	58,4 %	35,0 %
	CF	83,8 %	83,8 %	83,7 %
Uni + big	YATSI-NBM	69,6 %	73,9 %	68,1 %
	YATSI-RF	62,1 %	76,9 %	56,0 %
	CF	84,6 %	84,6 %	84,6 %

Tabulka 4.9: Výsledky metod učení z klasifikovaných a neklasifikovaných příkladů pro českou množinu

U výsledků pro anglickou množinu v tabulce 4.12 lze vidět podobný trend, kdy majoritně hlasující klasifikátor dosahuje v nejlepším případě úspěšnosti až 94 % a také klasifikátory RF a J48 dosahují obdobných výsledků.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	YATSI-NBM	65,0 %	77,0 %	60,6 %
	YATSI-RF	70,4 %	71,9 %	69,9 %
	CF	87,5 %	87,7 %	87,5 %
Bigramy	YATSI-NBM	59,6 %	71,3 %	53,1 %
	YATSI-RF	54,6 %	76,2 %	42,8 %
	CF	90,0 %	90,0 %	90,0 %
Uni + big	YATSI-NBM	60,4 %	77,9 %	53,1 %
	YATSI-RF	88,3 %	88,4 %	88,3 %
	CF	90,8 %	90,8 %	90,8 %

Tabulka 4.10: Výsledky metod učení z klasifikovaných a neklasifikovaných příkladů pro anglickou množinu

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	80,0 %	85,7 %	79,2 %
	SVM	68,0 %	80,5 %	64,3 %
	RF	88,0 %	88,2 %	88,0 %
	J48	85,0 %	85,4 %	85,0 %
	Vote	89,0 %	89,8 %	88,9 %
Bigramy	NBM	76,0 %	83,8 %	74,5 %
	SVM	82,0 %	86,8 %	81,4 %
	RF	83,0 %	83,3 %	83,0 %
	J48	88,0 %	88,1 %	88,0 %
	Vote	89,0 %	89,8 %	88,9 %
Uni + big	NBM	78,0 %	84,7 %	76,9 %
	SVM	66,0 %	79,8 %	61,6 %
	RF	86,0 %	86,5 %	85,9 %
	J48	87,0 %	87,1 %	87,0 %
	Vote	89,0 %	90,3 %	88,9 %

Tabulka 4.11: Výsledky metod učení s učitelem pro českou množinu

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	77,0 %	82,8 %	75,9 %
	SVM	56,0 %	76,6 %	45,4 %
	RF	92,0 %	92,1 %	92,0 %
	J48	92,0 %	92,0 %	92,0 %
	Vote	92,0 %	92,1 %	92,0 %
Bigramy	NBM	70,0 %	81,3 %	67,0 %
	SVM	57,0 %	76,9 %	47,2 %
	RF	88,0 %	88,0 %	88,0 %
	J48	92,0 %	92,0 %	92,0 %
	Vote	89,0 %	89,8 %	88,9 %
Uni + big	NBM	75,0 %	83,3 %	73,3 %
	SVM	56,0 %	76,6 %	45,4 %
	RF	94,0 %	94,0 %	94,0 %
	J48	93,0 %	93,0 %	93,0 %
	Vote	94,0 %	94,0 %	94,0 %

Tabulka 4.12: Výsledky metod učení s učitelem pro anglickou množinu

4.6 Shrnutí

Při testování dosahovala nejlepší úspěšnosti až 98 % metoda LOF a dobrých výsledků dosahovala také klasifikace s využitím metody RandomForest a to až 94 %. Při porovnávání obou přístupů je nutné uvažovat také jejich rozdílnost. Klasifikační metody se musí vypořádat jak s klasifikací normálních a anomálních uživatelů, tak také musí být schopny rozhodnout pro každou testovanou instanci. Naopak metoda využívající LOF nám vrátí pouze seznam uživatelů seřazených sestupně dle přiřazeného skóre, kde předpokládáme, že největší hustota anomálních uživatelů se nachází na začátku seznamu a dále postupně klesá.

Dalším aspektem, který je nutné brát v úvahu při hodnocení vhodnosti zvoleného přístupu, jsou nároky na reálné nasazení v praxi. V případě klasifikace se klasifikátor naučí požadovaný model a klasifikace poté probíhá velmi rychle. Naopak u LOF je nutné, aby metoda provedla analýzu vždy na celé množině a není tudíž vhodná pro rozpoznání anomálních uživatelů v případě požadavku na detekci v reálném čase. Na druhou stranu, pokud by byla použita inkrementální implementace algoritmu LOF, bylo by možné

toto omezení překonat [34].

LOF je tedy vhodné použít tehdy, pakliže máme množinu uživatelů a potřebujeme získat „nejvíce“ anomální uživatele (uživatelé ohodnoceni nejvyšším koeficientem odlehlosti). Typicky se tak jedná o případy, kdy chceme očistit trénovací množinu apod. Naopak klasifikaci je vhodné použít tehdy, pokud je kladen důraz na rychlé rozpoznání uživatele a je požadováno přiřazení třídy každé testované instanci.

5 Charakteristika anomálních uživatelů

Zdá se, že anomální uživatelé nejsou zpravidla pouze jednotlivé odlehlé instance, ale shluky, které lze charakterizovat chováním v některých jejich atributech. Proto v této kapitole provedeme analýzu anomálních uživatelů pomoci shlukové analýzy. Na základě pozorování pak vytvoříme charakteristiku anomálních uživatelů popisující jednotlivé tendence těchto skupin s uvedením konkrétních příkladů na Twitteru.

5.1 Shlukování anomálních uživatelů

Na základě pozorování anomálních uživatelů a jejich hodnot atributů se ukazuje, že řada uživatelů má v některých svých atributech podobné tendence chování a je možné jednotlivé typy anomálních uživatelů charakterizovat. Jednou z metod, kterou lze použít pro rozdělení anomálních uživatelů do jednotlivých skupin, je použití metody shlukování. Metoda sama o sobě nedá informaci o tom, o jakou skupinu uživatelů se jedná, ale bližší analýzou těchto účtů můžeme provést jejich studium a pokusit se skupiny uživatelů charakterizovat.

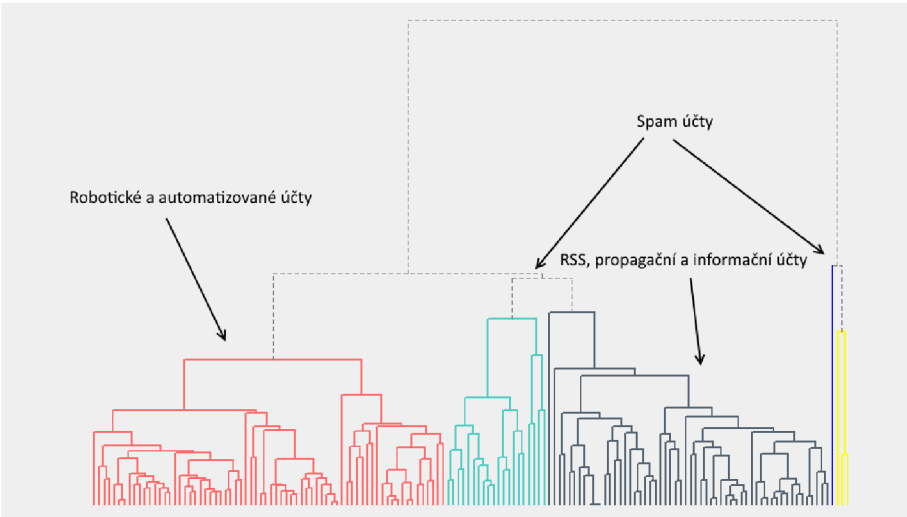
U shlukování je zpravidla důležité vědět, kolik shluků hledáme. Zde budeme vycházet z již existujících metod [35, 17, 21], které hledají skupiny anomálních účtů jako jsou spam účty, robotické účty nebo informační účty včetně RSS čteček. Dále z pozorování víme, že existuje zvláštní skupina jednosměrně tématicky zaměřených účtů, které jsou odlišné především svým slovníkem. V případě, že bylo nutné při analýze zvolit parametr počtu hledaných shluků, byla vybrána tato hodnota z intervalu 3 až 5.

Při analýze byla využita metoda hierarchického shlukování a metoda hledání nejbližších sousedů. Pomoci obou metod byla provedena shluková analýza nad 150 anomálními uživateli jak v české, tak v anglické datové sadě, přičemž množina atributů neobsahovala atributy odvozené z dokumentu příspěvků uživatele, jelikož naší motivací je pokusit se charakterizovat jednotlivé skupiny pouze na základě chování uživatelů a jejich aktivitě. S využitím poznatků získaných při pozorování byly poté vyvozeny charakteristiky chování anomálních uživatelů.

5.1.1 Hierarchické shlukování

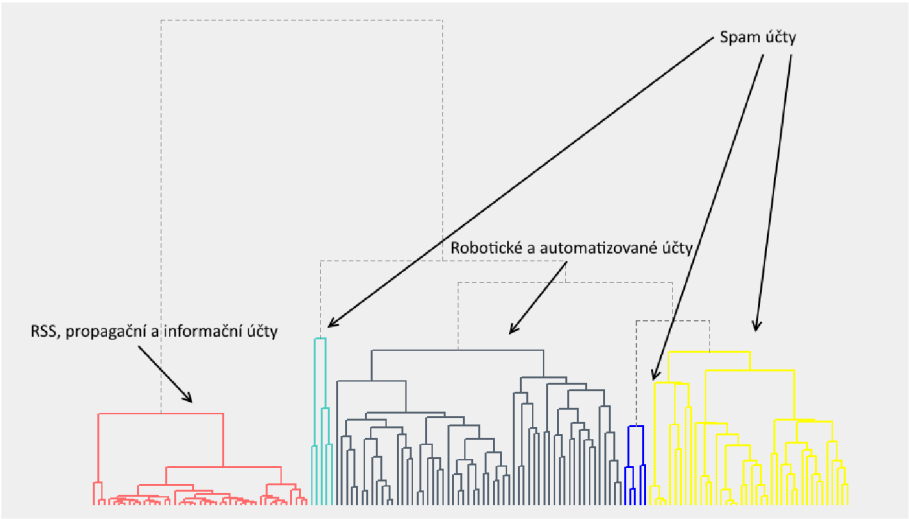
Analýza byla provedena pomocí implementace v jazyce R [36], kde byly testovány různé aglomerační přístupy. Jako vhodný přístup se ukázalo použít výchozí nastavení, kde vzdálenost mezi body se spočítala s využitím Euklidovské vzdálenosti a způsob výpočtu vzdálenosti shluků byl nastaven na průměr. Na základě analýzy byl také vytvořen dendrogram pro českou a anglickou množinu anomálních uživatelů, kde byla provedena hlubší analýza pro 5 nejvýše postavených shluků.

U české množiny na obrázku 5.1 můžeme vidět dominanci dvou shluků. Červený shluk představuje účty s určitou mírou automatizace. Účty v tomto shluku vykazují vysokou aktivitu pouze v jedinou hodinu dne, nevykazují interakci s okolím a jejich slovník neobsahuje velký počet unikátních slov oproti jiným uživatelům. Zároveň některé účty vykazují nízké odchylky v délce vkládaných příspěvků. Naopak zelený, modrý a žlutý shluk obsahuje uživatele, které lze charakterizovat jako spam účty. Shluky se pak liší pouze zvolenou strategií pro šíření svých zpráv. Uživatelé v zeleném shluku vykazují vysoký počet použití klíčových slov v kombinaci s URL odkazy, naopak žlutý a modrý shluk obsahuje mnohem agresivnější uživatele co do počtu použití klíčových slov. Černý shluk pak zastupuje uživatelské účty, které lze charakterizovat jako informační účty a RSS čtečky s téměř žádnou interakcí s okolím, ale s vysokým počtem užití URL odkazů v příspěvku.



Obrázek 5.1: Dendrogram pro českou množinu

Naopak v anglické množině na obrázku 5.2 dominují 3 větší shluky a dva menší. Červený shluk reprezentuje skupinu uživatelů, které lze označit jako informační účty a RSS čtečky. Jejich hlavními znaky je vysoký počet použitých URL odkazů a naopak téměř nulová interakce s okolím. Slovník, který tyto účty používají, je rozmanitý. Zelený, modrý a žlutý shluk lze označit jako spam účty, přičemž každý shluk odlišuje strategie, jakou tyto účty používají. Dominantní žlutý shluk obsahuje účty, které dosahují vysokého počtu použitých klíčových slov a odkazů. Naopak zelený shluk svou strategii doplňuje i o vysoký počet zmínění. Modrý shluk pak používá obdobnou strategii jako žlutý shluk, ale z pohledu použití klíčových slov je mnohem agresivnější. Všechny tři shluky také používají nízký počet unikátních slov ve svém slovníku. Poslední černý shluk lze označit jako shluk robotických účtů v kombinaci s automatizovanými propagačními účty. Slovník tohoto shluku není příliš rozmanitý, řada účtů má vysokou aktivitu pouze v jedinou hodinu dne.



Obrázek 5.2: Dendrogram pro anglickou množinu

5.1.2 K-means

Analýza byla provedena pomoci implementace algoritmu hledání nejbližších sousedů k-means v jazyce R s výchozím nastavením [37], kde vzdálenost mezi body se spočítala s využitím Euklidovské vzdálenosti a maxi-

mální počet iterací byl nastaven na hodnotu 1 000. Parametr počtu hledaných shluků byl roven hodnotě 4.

Při pozorování bylo rozdělení uživatelů zařazených do jednotlivých shluků obdobné s rozdělením u předchozí metody. Snížení počtu hledaných shluků vedlo ke sloučení shluků spam účtů, přičemž byly stále zachovány shluky představující robotické účty a informačně propagační účty kombinované s RSS čtečkami. U české množiny došlo ke sloučení všech tří shluků obsahujících spam účty do jednoho shluku, naopak informačně propagační účty se rozdělily do dvou shluků, kde rozdělujícím atributem byla rozmanitost použitého slovníku jednotlivých účtů.

5.2 Charakteristiky uživatelů

Na základě analýzy pomoci shlukování anomálních uživatelů v kapitole 5.1 bylo provedeno pozorování jednotlivých uživatelů a jejich zařazení v rámci shluku. Díky této zkušenosti bylo možné lépe sledovat tendence v chování u jednotlivých atributů v daných skupinách a také provést jejich charakteristiku. Při analýze byly charakterizovány tři hlavní skupiny anomálií a to robotické účty, spam účty a informační, propagační účty včetně RSS čteček vykazující nulovou interakci s okolím. Zvláštní odlehlou skupinou jsou pak jednosměrně tématicky zaměřené účty.

5.2.1 Robotické účty

Robotické účty vkládají své příspěvky v určité periodě. Perioda může být různě dlouhá, zpravidla se ale jedná o periody hodinové nebo denní. Roboti často přispívají v jedinou hodinu během dne a tudíž distribuce rozložení příspěvků oproti normálnímu uživateli je velmi odlišná. Řada účtů také nedosahuje vysoké rozmanitosti používaného slovníku, jestliže účet informuje pouze o změně nějaké hodnoty (účty o aktuální naměřené teplotě apod.), protože hodnota je vkládána stále do stejné šablony zprávy. Od toho se také odvíjí nízké odchylky v délce vkládaných příspěvků.

Robotické účty například informují o počasí, o TV programu nebo aktuálním dni (jaký je den, kdo má svátek, kurzy měn) apod.

> **@PocasiPraha**: po: 1/6 °C, Možnost deště. út: 3/7 °C, Zataženo. st: 4/9 °C, Možnost deště. čt: 2/7 °C, Možnost deště. [38]

> **@TV_program_CT24**: 20:00: Zprávy, 20:05: Hyde Park, 21:00: Zprávy | http://vtelevizi.cz [39]

5.2.2 Spam účty

Spam účty zpravidla vykazují ve své aktivitě vysoký počet použití klíčových slov ve svých příspěvcích doprovázený vyšším počtem URL odkazů, jelikož díky použití klíčových slov se rozšiřuje potencionální okruh uživatelů, kteří mohou daný příspěvek sledovat. Výjimkou také není použití vyššího počtu zmínění jiných uživatelských účtů, což automaticky vede k zasláni notifikace zmíněným uživatelům a šancí spam účtu na vyšší přesměrování přes jeho URL odkazy. Samotný obsah příspěvků je často podřízen strategii získání co největší audience a počtu odběratelů.

Zpravidla se tak jedná o propagační účty s agresivní marketingovou politikou, závadné účty odkazující na závadný nebo nevyžádaný komerční obsah. Účty také někdy vyzývají ostatní uživatele, aby spam účet sledovali a na oplátku nabízí, že odběratele budou sledovat zpět. Rozšiřují si tak dosah v síti a na první pohled se mohou tvářit jako významné účty s vysokým počtem odběratelů. Na druhou stranu, samotné závadné účty se sledují někdy i mezi sebou, aby si uměle navyšovaly statistiky sledovanosti a působily důvěryhodněji.

> **@JulezPooh**: #VnAmuzik @jakki2004 @VicesAdulation @RT2EAT @Marc1TT @TwitWhizz [40]

> **@FotoHanus**: Soccer league Pražský přebor #prebor #press #fotbalpraha #canon #1div #300mm #pro #report #praha... https://instagram.com/p/1Yk2bRN1zH/ [41]

5.2.3 RSS čtečky, informační a propagační účty

Účty tohoto typu zpravidla vykazují vysoký počet URL odkazů, nižší rozmanitost použitého slovníku nebo nízké výkyvy v délce příspěvků. Zároveň tyto účty nevykazují téměř žádnou interakci s okolím prostřednictvím reakcí na jiné uživatele nebo sdílením jejich příspěvků. Někdy jsou URL odkazy nahrazeny doprovodnými obrázky nebo videi k příspěvkům, které jsou motivovány danou propagační strategií. Příspěvky informačních účtů a RSS čteček také málokdy vykazují použití emotikonů ve svých zprávách.

Jedná se tak o účty některých informačních služeb a webových aplikací, které automaticky přispívají aktualizacemi o změnách, nových příspěvcích

nebo nových produktech. Výjimkou také nejsou uživatelé, kteří aktivně používají sociální síť Facebook, ale příspěvky z této sítě si nechávají automaticky v rámci některé služby přeposílat na svůj Twitter účet, přičemž jej aktivně vůbec nepoužívají a účet tak upadá do podoby pouhé RSS čtečky příspěvků z jiné sociální sítě. Mimo to, takové příspěvky nejsou kompletní, jelikož s omezením délky zprávy na 140 znaků dochází k uříznutí příspěvku a jejího pokračování mimo síť Twitter.

> @**webfaircz**: Nové vlákno: Koupím adult, gambling http://goo.gl/fb/Wuaah [42]
>
> @**EngineersDay**: New Job: Mechanical & Electrical Engineers, Company: (Dublin, OH, United States) http://ift.tt/1IpHBla [43]

5.2.4 Jednosměrně tématicky zaměřené účty

Jednosměrně tématicky zaměřené účty jsou speciální skupinou účtů, které svým charakterem nelze vždy zcela jednoznačně zařadit do některé z předchozích skupin, přitom svým chováním můžou nést některé z jejich rysů. Klíčový u těchto účtů je především jejich slovník, který tyto účty odhaluje.

Mohou to být úzce zaměřené účty informující o zemětřesení nebo dopravě v konkrétním městě, účty vkládající neustále stejný obsah, počítající počet dnů do konce světa nebo citující pasáže z bible apod. Existují také svým zaměřením zvláštní robotické účty, které informují o tom, kdy zrovna odbíjí hodiny na Pražském orloji nebo britském Big Benu.

> @**big_ben_clock**: BONG BONG BONG BONG BONG BONG BONG BONG BONG BONG BONG BONG [44]
>
> @**vorloj**: bim bam bim bam bim bam bim bam bim bam bim [45]

6 Klasifikace a vliv anomálních uživatelů

Klasifikace je důležitou metodou, která se využívá na sociální síti Twitter při řešení řady problémů. Používá se nejen při detekci robotických účtů [17], ale také při klasifikaci sentimentu [26] nebo rozpoznání politické preference voliče [46]. Výsledky klasifikace jsou ale často závislé na kvalitě dat, na kterých dochází k učení, tudíž je vhodné vždy data pročistit. Jednou z metod může být zohlednění vlivu anomálních uživatelů a filtrace jejich obsahu.

V této kapitole se zaměříme na problém klasifikace sentimentu, kde se zaměříme na klasifikaci příspěvků do dvou (pozitivní a negativní) a do tří (pozitivní, neutrální a negativní) tříd. Dále popíšeme tvorbu trénovací a testovací množiny, způsob předzpracování a dosažené výsledky. Také použijeme navrženou metodu pro detekci anomálních uživatelů a ověříme, zda odstranění anomálních uživatelů z trénovací množiny má vliv na výsledky klasifikace a může tak být relevantní technikou pro zvýšení kvality učících dat.

6.1 Klasifikace sentimentu

Analýza sentimentu se v posledních letech stává čím dál populárnější oblastí pro studium. Může se kupříkladu využít v oblasti predikce cen akcií na finančních trzích díky sledování zpráv na sociálních sítích [47] nebo při marketingové analýze ohlasů zákazníků na nový produkt vstupující na trh nebo jejich recenzí [48]. Stává se čím dál populárnější také na sociální síti Twitter a to především kvůli specifickému jazykovému modelu a omezením, které se zde projevují [49].

Klasifikace sentimentu spočívá v detekci pozitivních nebo negativních emocí v příspěvků uživatele. Pokud příspěvek není pozitivní ani negativní, pak je považován za neutrální. Rozpoznání třídy sentimentu není jednoduchý problém, jelikož jeden příspěvek může obsahovat i více emocí zároveň, projevy ironie nebo sarkazmu, což určení třídy sentimentu ztěžuje [50].

6.2 Učení na základě emotikonů

Uživatelé na sociálních sítích s oblibou vkládají do svých příspěvků emotikony označující jejich subjektivní sentiment. Díky dostupnosti velkého

množství příspěvků je tak možné takto označené příspěvky sdružovat do datových sad a místo lidské práce anotátorů použít právě emotikony jako zdroj informace o sentimentu ve zprávě. Na druhou stranu, je třeba si uvědomit, že informace od uživatelů nemusí být vždy zcela správné a v některých případech nemusí emotikon korespondovat se skutečným sentimentem zprávy.

Metoda analýzy sentimentu učením na základě emotikonů byla použita již při analýze sentimentu příspěvků na anglickém Twitteru [26], kde dosahovala úspěšnosti 83 % při klasifikaci do dvou tříd (pozitivní a negativní) nad testovací množinou o velikosti 359 příspěvků. Trénovací množina měla 800 000 příspěvků z každé třídy.

6.3 Trénovací množina

Trénovací množina pro účely klasifikace sentimentu v této práci byla vybudována s využitím metody učení na základě emotikonů. Aby každá třída měla rovné zastoupení v trénovací množině, bylo náhodně vybráno ze získaných dat 18 000 příspěvků z každé třídy pro českou množinu a 8 000 z každé třídy pro anglickou, přičemž rozdílnost velikosti množin se odvíjí od počtu získaných příspěvků obsahujících emotikony pro každý jazyk. Třídy byly detekovány podle obsažených emotikonů v těle zprávy. Konkrétní emotikony pro každou třídu jsou uvedeny v tabulce 6.1.

Třída	Emotikony
Pozitivní	:), :-), :o), :-o), :D, :-D, XD, ;), ;-), ;o), ;-o), ;D, ;-D
Negativní	:(, :-(, :o(, :-o(, :,-(, :'-(, ;(, :-(, ;o(, ;-o(, ;,-(, ;'-(
Neutrální	Neobsahuje pozitivní ani negativní emotikon.

Tabulka 6.1: Třídy sentimentu dle emotikonů

V případě, že zpráva obsahuje emotikon ze skupiny pozitivních a negativních emotikonů zároveň, pak taková zpráva nebyla zařazena do trénovací množiny. Předzpracování probíhá totožným způsobem, jako v podkapitole 4.3.

6.4 Testovací množina

Testovací množina byla vytvořena z náhodně vybraných příspěvků[1], kterým byla dále přiřazena třída sentimentu na základě anotace dvou osob. Tabulky 6.2 a 6.3 ukazují, jaká panovala shoda mezi anotátory při přiřazování jednotlivých tříd. Cohenovo kappa pro českou testovací množinu je rovno hodnotě 0,749 a pro anglickou testovací množinu 0,740, což splňuje požadavky na důvěryhodnost anotace. Čistá shoda anotátorů pro českou množinu je 83,28 % a pro anglickou 82,68 %.

		Anotátor B		
		Pozitivní	Neutrální	Negativní
Anotátor A	Pozitivní	161	27	9
	Neutrální	23	159	14
	Negativní	10	14	163

Tabulka 6.2: Shoda anotátorů pro českou testovací množinu sentimentu

		Anotátor B		
		Pozitivní	Neutrální	Negativní
Anotátor A	Pozitivní	129	17	7
	Neutrální	21	130	11
	Negativní	6	17	118

Tabulka 6.3: Shoda anotátorů pro anglickou testovací množinu sentimentu

V případě, že se anotátoři při přiřazení třídy příspěvku lišili, rozhodla o přiřazení jedné z tříd třetí osoba. Tabulka 6.4 ukazuje rozdělení příspěvků do jednotlivých tříd v testovací množině.

Třída	Česká množina	Anglická množina
Pozitivní	181	145
Neutrální	209	164
Negativní	190	147

Tabulka 6.4: Rozdělení tříd v testovací množině při klasifikaci sentimentu

1. V tomto případě výběr probíhal ze všech příspěvků, bez ohledu na to, zda příspěvek obsahoval emotikon

6.5 Výsledky klasifikace do tří tříd

Nyní si představíme výsledky klasifikace sentimentu na pozitivní, neutrální a negativní sentiment. V první části rozebereme výsledky využívající trénovací množinu včetně příspěvků anomálních uživatelů, poté příspěvky těchto uživatelů odstraníme a budeme sledovat, jaký vliv na výsledky klasifikace bude zaznamenán. Měření provedeme vždy jak pro českou, tak anglickou trénovací množinu, abychom eliminovali vedlejší vlivy spojené s vlastnostmi konkrétního jazyka.

Při každém měření jsou výsledky uvedeny pro klasifikátory NBM, SVM a RF. Zároveň byla vždy provedena tokenizace textu na unigramy, bigramy a na kombinaci obsahující jak unigramy, tak bigramy. Počet tokenů byl vždy omezen na 3 000 nejfrekventovanějších pro každou třídu. K vyhodnocení byl použit program Weka. Detailní informace o nastavení jednotlivých klasifikátorů a tokenizaci textu lze nalézt v příloze B.

6.5.1 Bez odstranění anomálních uživatelů

Nejprve si ukážeme výsledky klasifikace sentimentu do tří tříd se zachováním příspěvků anomálních uživatelů v trénovací množině. Baseline pro českou množinu dosahuje hodnoty 35,4 % a pro anglickou 36 %.

Tabulka 6.5 zobrazuje výsledky klasifikace sentimentu do tří tříd pro českou množinu a tabulka 6.6 pro anglickou. V obou případech si můžeme povšimnout, že nejlepších výsledků obou množin dosahuje tokenizace na unigramy v kombinaci s klasifikátorem NBM, kde je úspěšnost vyšší než 62 % a u anglické množiny dokonce vyšší než 64 %. Naopak tokenizace na bigramy dosahuje nejhorších výsledků, kde jednou z příčin může být řidší matice atributů obsahujících informace o výskytu jednotlivých tokenů.

6.5.2 Po odstranění anomálních uživatelů

Nyní využijeme navrženou metodu pro rozpoznání anomálních uživatelů aplikací metody LOF, která dosahovala při experimentech v podkapitole 4.5.1 lepších výsledků, než metoda OR. Zároveň LOF použijeme se stejným nastavením, jako v podkapitole 4.5.1. Ke každému příspěvku v trénovací množině je vždy evidován uživatel, který je jeho autorem. Díky tomu získáme množinu uživatelů, kteří jsou autory příspěvků naší trénovací množiny pro klasifikaci sentimentu. Poté uživatele, kteří disponují v historii ale-

spoň 200 příspěvky, podrobíme metodě pro detekci anomálních uživatelů. Uživatelé s nižším počtem příspěvků nebereme v úvahu, jelikož stáří jejich účtu není dostatečné na to, aby se u nich projevil trend chování.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	62,9 %	65,6 %	61,4 %
	SVM	49,8 %	47,4 %	43,6 %
	RF	53,4 %	53,7 %	52,0 %
Bigramy	NBM	44,5 %	43,6 %	43,2 %
	SVM	35,5 %	25,6 %	25,6 %
	RF	47,1 %	47,1 %	44,2 %
Uni + big	NBM	61,6 %	63,8 %	59,7 %
	SVM	50,7 %	48,8 %	44,0 %
	RF	50,0 %	50,1 %	48,5 %

Tabulka 6.5: Klasifikace sentimentu v české datové sadě na pozitivní, neutrální a negativní příspěvky bez odstranění anomálních uživatelů

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	64,5 %	65,2 %	64,2 %
	SVM	55,9 %	57,3 %	56,1 %
	RF	57,7 %	58,7 %	57,9 %
Bigramy	NBM	51,1 %	54,5 %	51,2 %
	SVM	46,5 %	48,2 %	42,3 %
	RF	52,0 %	54,0 %	51,9 %
Uni + big	NBM	61,2 %	63,2 %	61,5 %
	SVM	56,4 %	57,6 %	56,5 %
	RF	56,8 %	58,0 %	56,9 %

Tabulka 6.6: Klasifikace sentimentu v anglické datové sadě na pozitivní, neutrální a negativní příspěvky bez odstranění anomálních uživatelů

Po aplikaci metody získáme graf, který zobrazuje distribuci koeficientu odlehlosti, přičemž ke každému uživateli známe jemu přiřazenou hodnotu. Z chování metody víme, že anomální uživatelé dosahují nejvyššího koeficientu odlehlosti, který se pohybuje kolem hodnoty 2 a více. Příspěvky anomálních uživatelů odstraníme z trénovací množiny.

Pro českou množinu lze graf distribuce koeficientu odlehlosti vidět na obrázku 6.1 a pro anglickou na obrázku 6.2. V obou případech na základě distribuce hodnot vybereme 200 uživatelů s nejvyšší hodnotou odlehlosti a jejich příspěvky odstraníme z trénovacích množin. Tabulka 6.7 zobrazuje, o jaký počet příspěvků se sníží velikost trénovacích množin.

Po odstranění příspěvků provedeme vyhodnocení totožným způsobem, jako bylo provedeno bez odstranění anomálních uživatelů. Pro lepší přehlednost srovnání budou nyní v tabulkách výsledků zaznamenány změny v naměřených výsledcích oproti předchozímu měření bez použití metody.

Stav	Česká množina	Anglická množina
Před odstraněním	48 000	24 000
Po odstranění	47 539	23 378

Tabulka 6.7: Změna velikosti trénovacích množin klasifikace sentimentu do dvou tříd po odstranění anomálních uživatelů

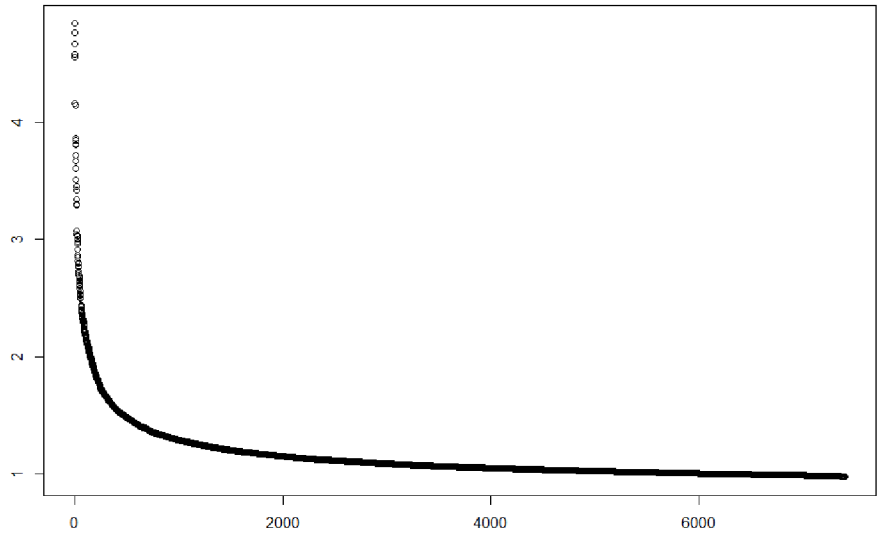
Tabulka 6.8 zobrazuje vliv na klasifikaci sentimentu pro českou množinu a tabulka 6.9 pro anglickou. Můžeme si povšimnout, že jak u české, tak anglické množiny došlo v obou případech ke zlepšení u klasifikátoru NBM při tokenizaci na unigramy. U české množiny došlo také ke zlepšení u kombinace unigramů a bigramů pro klasifikátor RF a u anglické množiny naopak při tokenizaci na bigramy u klasifikátoru NBM.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	0,7 %	0,9 %	0,7 %
	SVM	0	0	0
	RF	0	0,2 %	0,4 %
Bigramy	NBM	-0,5 %	-0,6 %	-0,6 %
	SVM	-1,5 %	12,9 %	-2,3 %
	RF	-0,5 %	-1,0 %	-0,7 %
Uni + big	NBM	-0,7 %	-0,9 %	-0,7 %
	SVM	-0,4 %	-0,5 %	-0,5 %
	RF	1,7 %	1,8 %	1,6 %

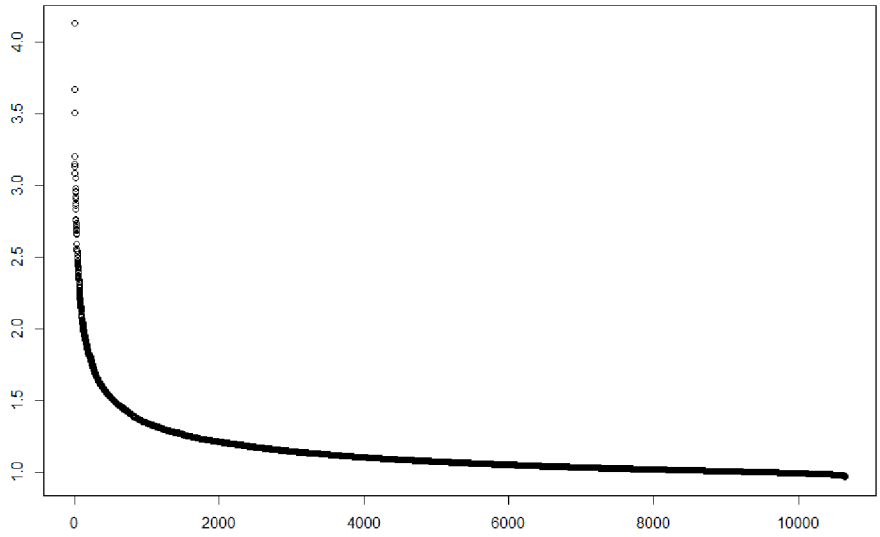
Tabulka 6.8: Vliv na klasifikaci sentimentu v české datové sadě na pozitivní, neutrální a negativní příspěvky po odstranění anomálních uživatelů

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	1,3 %	0,8 %	1,2 %
	SVM	0	-0,4 %	0
	RF	-1,1 %	-0,8 %	-1,0 %
Bigramy	NBM	2,0 %	1,8 %	2,0 %
	SVM	-4,0 %	2,7 %	-7,1 %
	RF	-1,1 %	-1,3 %	-1,0 %
Uni + big	NBM	-0,2 %	-0,8 %	-0,3 %
	SVM	0	-0,1 %	0
	RF	-1,5 %	-1,5 %	-1,4 %

Tabulka 6.9: Vliv na klasifikaci sentimentu v anglické datové sadě na pozitivní, neutrální a negativní příspěvky po odstranění anomálních uživatelů



Obrázek 6.1: Distribuce koeficientu odlehlosti pro českou množinu autorů příspěvků v trénovací množině pro klasifikaci do tří tříd



Obrázek 6.2: Distribuce koeficientu odlehlosti pro anglickou množinu autorů příspěvků v trénovací množině pro klasifikaci do tří tříd

6.6 Výsledky klasifikace do dvou tříd

Nyní si představíme výsledky klasifikace sentimentu na pozitivní a negativní sentiment. Postup při získávání výsledků je shodný s postupem popsaným v kapitole 6.5 při klasifikaci do tří tříd. Jediným rozdílem je odstranění neutrální třídy příspěvků z trénovací a testovací množiny pro každý jazyk.

6.6.1 Bez odstranění anomálních uživatelů

Tabulka 6.10 zobrazuje výsledky klasifikace sentimentu do dvou tříd pro českou množinu. Výsledky pro anglickou množinu lze nalézt v tabulce 6.11. V obou případech si můžeme povšimnout, že nejlepších výsledků obou množin dosahuje tokenizace na unigramy v kombinaci s metodou NBM, kde je úspěšnost vyšší než 80 %. Naopak tokenizace na bigramy dosahuje nejhorších výsledků. Baseline pro českou množinu dosahuje hodnoty 51,2 % a pro anglickou 50,3 %.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	80,6 %	81,7 %	80,5 %
	SVM	70,4 %	70,5 %	70,2 %
	RF	75,5 %	76,1 %	75,4 %
Bigramy	NBM	62,0 %	62,0 %	62,0 %
	SVM	51,2 %	26,2 %	34,7 %
	RF	63,3 %	64,7 %	62,7 %
Uni + big	NBM	78,7 %	80,4 %	78,5 %
	SVM	70,4 %	70,5 %	70,2 %
	RF	75,2 %	76,1 %	75,1 %

Tabulka 6.10: Klasifikace sentimentu v české datové sadě na pozitivní a negativní příspěvky bez odstranění anomálních uživatelů

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	82,5 %	83,7 %	82,4 %
	SVM	70,9 %	74,1 %	69,9 %
	RF	80,1 %	82,6 %	79,8 %
Bigramy	NBM	75,3 %	79,1 %	74,6 %
	SVM	60,3 %	66,3 %	56,4 %
	RF	72,6 %	73,6 %	72,3 %
Uni + big	NBM	80,1 %	82,4 %	79,8 %
	SVM	72,3 %	75,4 %	71,4 %
	RF	78,8 %	79,7 %	78,6 %

Tabulka 6.11: Klasifikace sentimentu v anglické datové sadě na pozitivní a negativní příspěvky bez odstranění anomálních uživatelů

6.6.2 Po odstranění anomálních uživatelů

Nyní opět využijeme navrženou metodu pro rozpoznání anomálních uživatelů stejným způsobem, jako tomu bylo při klasifikaci sentimentu do tří tříd s odstraněním anomálních uživatelů.

Pro českou množinu lze graf distribuce koeficientu odlehlosti vidět na obrázku 6.3 a pro anglickou na obrázku 6.4. V obou případech na základě distribuce hodnot vybereme 100 uživatelů s nejvyšší hodnotou odlehlosti

a jejich příspěvky odstraníme z trénovacích množin. Tabulka 6.12 zobrazuje, o jaký počet příspěvků se sníží velikost trénovacích množin.

Stav	Česká množina	Anglická množina
Před odstraněním	32 000	16 000
Po odstranění	31 713	15 520

Tabulka 6.12: Změna velikosti trénovacích množin klasifikace sentimentu do dvou tříd po odstranění anomálních uživatelů

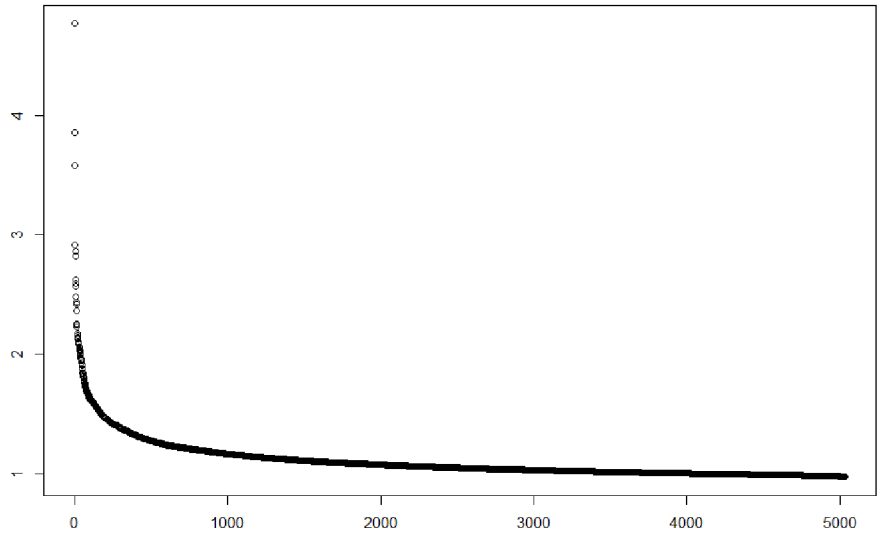
Tabulka 6.13 zobrazuje vliv na klasifikaci sentimentu pro českou množinu a tabulka 6.14 pro anglickou. Můžeme si povšimnout, že u české i anglické množiny došlo v obou případech ke zlepšení u klasifikátorů NBM a RF při tokenizaci na unigramy, přičemž se nejedná o jediná zlepšení, která byla zaznamenána.

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	1,3 %	1,6 %	1,3 %
	SVM	0,2 %	0,2 %	0,3 %
	RF	1,3 %	1,1 %	1,4 %
Bigramy	NBM	-0,5 %	-0,5 %	-0,5 %
	SVM	0	0	0
	RF	1,4 %	1,0 %	1,6 %
Uni + big	NBM	0	0	0
	SVM	0	-0,1 %	0,1 %
	RF	0,5 %	0,2 %	0,6 %

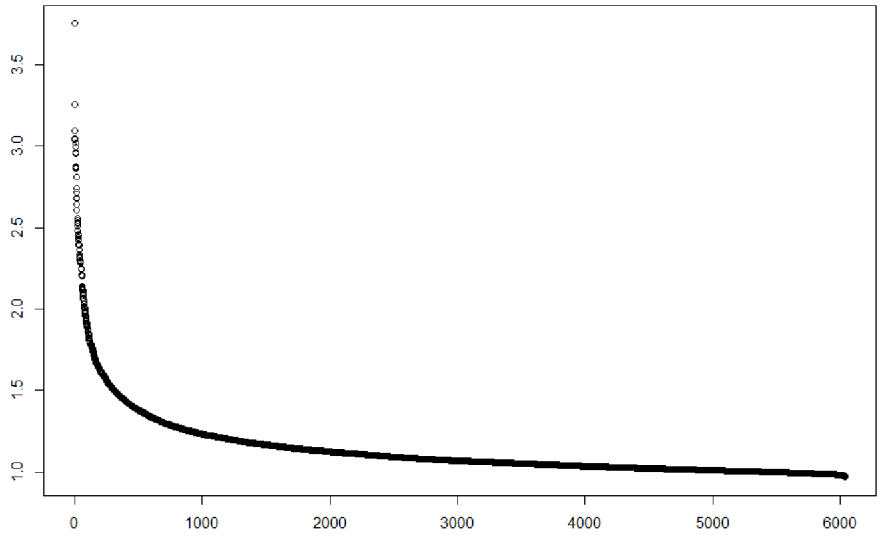
Tabulka 6.13: Vliv na klasifikace sentimentu v české datové sadě na pozitivní a negativní příspěvky po odstranění anomálních uživatelů

Tokenizace	Metoda	Úspěšnost	Přesnost	F-míra
Unigramy	NBM	2,1 %	1,8 %	2,1 %
	SVM	0,3 %	-0,5 %	0,6 %
	RF	1,4 %	0,3 %	1,5 %
Bigramy	NBM	0	-1,9 %	0,3 %
	SVM	2,4 %	-3,6 %	6,2 %
	RF	1,4 %	1,2 %	1,5 %
Uni + big	NBM	1,7 %	1,7 %	1,8 %
	SVM	0,3 %	0,3 %	0,4 %
	RF	0	0,4 %	0

Tabulka 6.14: Vliv na klasifikaci sentimentu v anglické datové sadě na pozitivní a negativní příspěvky po odstranění anomálních uživatelů



Obrázek 6.3: Distribuce koeficientu odlehlosti pro českou množinu autorů příspěvků v trénovací množině pro klasifikaci do dvou tříd



Obrázek 6.4: Distribuce koeficientu odlehlosti pro anglickou množinu autorů příspěvků v trénovací množině pro klasifikaci do dvou tříd

6.6.3 Využití pro klasifikaci do tří tříd

Abychom mohli využít klasifikátor na pozitivní a negativní příspěvky při praktickém použití, je nezbytné vyřešit, jakým způsobem zajistíme, aby příspěvky, které bude klasifikátor klasifikovat, byly skutečně buď pozitivní nebo negativní. Proto byla pro každý jazyk otestována kaskáda klasifikátorů.

První klasifikátor v kaskádě se naučí na trénovací množině obsahující neutrální a sentimentální třídu, přičemž sentimentální třída označuje příspěvky, které jsou pozitivní nebo negativní. Jako trénovací množina je použita množina uvedená v 6.3 a testovací množina uvedená v 6.4. Cílem klasifikátoru je dále propustit jen ty příspěvky, které označí jako sentimentální. Příspěvky, které jsou označeny jako sentimentální, dále klasifikuje klasifikátor do dvou tříd, který byl již v této kapitole popsán a dosahuje mnohem lepší úspěšnosti, než samotný klasifikátor do tří tříd. První klasifikátor využívá kombinaci NBM a tokenizaci na unigramy, která při experimentech dosahovala nejlepších výsledků.

U české množiny první klasifikátor v kaskádě dosahoval úspěšnosti 65 % při baseline 64 %. Druhý klasifikátor pak dosahoval úspěšnosti na příspěv-

cích označených jako sentimentální 67,2 % při baseline 55,1 %. Matici záměn pro celou kaskádu zobrazuje tabulka 6.15. Úspěšnost celé kaskády dosahuje hodnoty 58,6 % při baseline 36 %.

		Klasifikován jako:		
		Pozitivní	Neutrální	Negativní
Třída	Pozitivní	56	111	14
	Neutrální	31	156	22
	Negativní	23	39	128

Tabulka 6.15: Matice záměn pro kaskádu klasifikátorů pro českou množinu

U anglické množiny první klasifikátor dosahoval úspěšnosti 68 % při baseline 64 %. Druhý klasifikátor pak dosahoval úspěšnosti 66,8 % při baseline 38,7 %. Matici záměn pro celou kaskádu zobrazuje tabulka 6.16. Úspěšnost celé kaskády dosahuje hodnoty 62,7 % při baseline 36 %.

Získané výsledky naznačují, že hlavní problém při využití klasifikátoru do dvou tříd při reálném nasazení vychází právě z úspěšnosti, s jakou jsme schopni klasifikovat neutrální a sentimentální příspěvky. Jedním z hledisek, které je nutné brát na zřetel, je skutečnost, že třída neutrálních příspěvků v trénovací množině obsahuje příspěvky bez emotikonů, což nezaručuje absenci pozitivních a negativních příspěvků.

		Klasifikován jako:		
		Pozitivní	Neutrální	Negativní
Třída	Pozitivní	93	41	13
	Neutrální	21	91	52
	Negativní	11	32	102

Tabulka 6.16: Matice záměn pro kaskádu klasifikátorů pro anglickou množinu

6.7 Shrnutí

Klasifikace sentimentu na sociální síti Twitter při klasifikaci na pozitivní, neutrální a negativní sentiment dle provedených experimentů dosahuje pro českou množinu úspěšnosti 62,9 % a pro anglickou 64,5 %. V případě klasifikace pouze na pozitivní a negativní sentiment dosahuje nejlepší český

klasifikátor úspěšnosti 80,6 % a anglický 82,5 %. Je třeba si ale uvědomit, že klasifikátor do dvou tříd v případě praktického použití nemůže ignorovat skutečnost, že v testovaných instancích se budou vyskytovat i neutrální zprávy, což lze řešit klasifikátory v kaskádě, kde první klasifikátor propustí jen ty zprávy, které budou buď pozitivní nebo negativní. Při testování se ale ukázalo, že taková kaskáda dosahuje nižší úspěšnosti, než samostatný klasifikátor do tří tříd sentimentu a pro dosažení lepších výsledků je nezbytné zvýšit úspěšnost prvního klasifikátoru v kaskádě.

Dále bylo zjištěno, že při aplikaci metody pro detekci anomálních uživatelů můžeme odstraněním minoritní části příspěvků z trénovacích množin zvýšit úspěšnost výsledných klasifikátorů. U klasifikace do tří tříd došlo ke zlepšení u české množiny o 0,7 % na 63,6 %. U anglické došlo ke zlepšení o 1,7 % na 65,8 %. U klasifikátoru do dvou tříd došlo u české množiny ke zlepšení o 1,3 % na 81,9 % a u anglické o 2,1 % na 84,6 %.

Při porovnání s jinými pracemi zabývajícími se klasifikací sentimentu dosahuje získaný klasifikátor pro anglický Twitter s využitím metody pro detekci anomálních uživatelů vyšší úspěšnosti, než práce zaměřující se na klasifikaci sentimentu pomoci emotikonů na Twitteru, kde autoři dosáhli úspěšnosti 83 % [26]. Naopak česká práce [48] zaměřená na klasifikaci do dvou a do tří tříd s využitím příspěvků ze sociální sítě Facebook dosahuje lepších výsledků, které jsou ovlivněny hlavně rozdílným přístupem, jelikož práce nevyužívala emotikony k přiřazení tříd příspěvkům v trénovacích množinách, ale příspěvky byly ručně anotovány, přičemž informace o obsažených emotikonech v příspěvku byly rozšiřujícím atributem. Při klasifikaci do tří tříd práce dosáhla F-míry 69 % a při klasifikaci do dvou tříd F-míry 90 %.

7 Závěr

V rámci této práce byla navržena metoda pro rozpoznání anomálních uživatelů, kde metoda LOF dosahuje úspěšnosti až 98 % a metoda využívající klasifikaci s pomocí algoritmu RandomForest dosahuje úspěšnosti až 94 %. Při návrhu metody byly testovány i další přístupy jako učení z klasifikovaných a neklasifikovaných příkladů a dále byla navržena řada nových atributů vhodných pro detekci anomálních uživatelů. Mezi tyto atributy kupříkladu patří počet pozitivních příspěvků, počet médií, počet unikátních slov, směrodatná odchylka délky příspěvků uživatele nebo také maximální denní aktivita a další.

Zároveň byla provedena charakteristika anomálních uživatelů pomocí metody shlukování a jednotlivé podskupiny anomálních uživatelů byly analyzovány a blíže popsány. Ukázalo se, že jak na českém, tak anglickém Twitteru lze rozdělit anomální uživatele do stejných podskupin, které jsou vyznačeny podobným chováním v několika znacích uživatele. Byly tak odhaleny hlavní tři třídy anomálních účtů, mezi které patří spam účty, robotické účty a propagační a informační účty s nulovou interakcí s okolím. Mimo tyto třídy byla charakterizována také skupina účtů, které svým charakterem nelze vždy zcela jednoznačně zařadit do některé z předchozích skupin, přitom svým chováním můžou nést některé z jejich rysů.

Při klasifikaci sentimentu na pozitivní, neutrální a negativní příspěvky bylo dosaženo úspěšnosti až 65,8 % a při klasifikaci na pozitivní a negativní příspěvky úspěšnosti až 84,6 %, přičemž pro český Twitter byla vždy úspěšnost mírně nižší. Také byl navržen postup, jakým způsobem lze využít metodu pro rozpoznání anomálních uživatelů ke zlepšení výsledků klasifikace, což se ukázalo být relevantním postupem. Aplikovaná metoda v některých případech po odstranění příspěvků anomálních uživatelů z trénovací množiny zvýšila úspěšnost klasifikace v nejlepším případě až o 2,1 %. Metoda by tak mohla pomoci v řadě aplikací, kde jsou využívány příspěvky uživatelů k řešení různých klasifikačních problémů.

Všechny provedené experimenty byly vždy provedeny jak s využitím dat z českého, tak anglického Twitteru, což pomohlo ověřit uplatnitelnost navržené metody bez závislosti na zvoleném jazyce a zároveň sledovat rozdílnost výsledků klasifikace sentimentu pro oba jazyky. Také byly navrženy heuristiky pro získání vzorku dat české a anglické komunity a vznikl jednoduchý nástroj pro anotaci uživatelů na sociální síti Twitter.

Práce by se v budoucnu mohla rozšiřovat v několika směrech. Mezi tyto směry patří:

- otestování dalších algoritmů a jejich úspěšnosti při detekci anomálních uživatelů
- návrh dalších atributů
- doplnění metody o schopnost rozpoznat, do jaké třídy anomální uživatel patří
- nalezení a analýza odlehlých anomálních uživatelů v rámci své podskupiny s využitím příslušných metod
- ověřit vliv navržené metody na výsledky dalších klasifikačních problémů na sociální síti Twitter, jako je klasifikace tématu příspěvku a další
- zlepšit úspěšnost klasifikátoru, který v rámci kaskády při klasifikaci sentimentu rozděluje příspěvky na neutrální a příspěvky obsahující sentiment
- testování dalších metod předzpracování příspěvků uživatele, jako je lemmatizace a další

Literatura

[1] About Twitter. [online]. 2015. [cit. 2015-04-11]. Dostupné z: https://about.twitter.com/company

[2] KOULOUMPIS, E.; WILSON, T.; MOORE, J.: Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, ročník 11, 2011: s. 538–541.

[3] SRIRAM, B.; FUHRY, D.; DEMIR, E.; aj.: Short text classification in twitter to improve information filtering. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010: s. 841–842.

[4] DUGGAN, M.: Social Media Update 2014. [online]. 2015. [cit. 2015-04-01]. Dostupné z: http://www.pewinternet.org/2015/01/09/social-media-update-2014/

[5] MOORE, H.; ROBERTS, D.: AP Twitter hack causes panic on Wall Street and sends Dow plunging. [online]. 2013. [cit. 2015-04-01]. Dostupné z: http://www.theguardian.com/business/2013/apr/23/ap-tweet-hack-wall-street-freefall

[6] MICHL, P.: Infografika: Sociální sítě v Česku. [online]. 2013. [cit. 2015-02-20]. Dostupné z: http://www.m-journal.cz/cs/internet/socialni-site/infografika--socialni-site-v-cesku__s416x9788.html

[7] Klábosení. [online]. [cit. 2015-04-11]. Dostupné z: http://www.klaboseni.cz/

[8] BERKA, P.: *Dobývání znalostí z databází.* Academia, 2003, ISBN 80-200-1062-9.

[9] HAN, J.; KAMBER, M.; PEI, J.: *Data Mining: Concepts and Techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011, ISBN 978-0123814791.

[10] Twitter API. [online]. 2015. [cit. 2015-04-11]. Dostupné z: https://dev.twitter.com/rest/public

[11] Twitter Advanced Search. [online]. 2015. [cit. 2015-04-11]. Dostupné z: https://twitter.com/search-advanced

[12] Terms: Agreement and Policy. [online]. 2015. [cit. 2015-04-11]. Dostupné z: https://dev.twitter.com/overview/terms/agreement-and-policy

[13] Introducing new metadata for Tweets. [online]. 2013. [cit. 2015-04-11]. Dostupné z: https://blog.twitter.com/2013/introducing-new-metadata-for-tweets

[14] SHUYO, N.: Language Detection Library for Java. [online]. 2010. [cit. 2015-04-11]. Dostupné z: https://code.google.com/p/language-detection/

[15] MySQL. [online]. 2015. [cit. 2015-04-11]. Dostupné z: http://www.mysql.com/

[16] Quarterly report pursuant to section 13 or 15(d) of the securities exchange act of 1934. [online]. 2014. [cit. 2015-03-09]. Dostupné z: http://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm?_ga=1.172053859.1713638677.1406023469

[17] CHU, Z.; GIANVECCHIO, S.; WANG, H.; aj.: Detecting Automation of Twitter Accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, ročník 9, č. 6, 2012: s. 811–824. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6280553

[18] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H.: Introduction to information retrieval. Cambridge: Cambridge university press, 2008, ISBN 978-0-521-86571-5.

[19] MCCORD, M.; CHUAH, M.: Spam detection on twitter using traditional classifiers. *Autonomic and trusted computing*, 2011: s. 175–186.

[20] HALL, M.; FRANK, E.; HOLMES, G.; aj.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, ročník 11, 2009.

[21] CHU, Z.; WIDJAJA, I.; WANG, H.: Detecting Social Spam Campaigns on Twitter. In *Applied Cryptography and Network Security*, Springer, 2012, s. 455–472. Dostupné z: `http://link.springer.com/10.1007/978-3-642-31284-7_27`

[22] MILLER, Z.; DICKINSON, B.; DEITRICK, W.; aj.: Twitter spammer detection using data stream clustering. *Information Sciences*, ročník 260, 2014: s. 64–73. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S0020025513008037`

[23] ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; aj.: A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, ročník 96, 1996, s. 226–231.

[24] ARTHUR, D.; VASSILVITSKII, S.: k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, 2007, s. 1027–1035.

[25] PUSTEJOVSKY, J.: Natural language annotation for machine learning. O'Reilly, 2013, ISBN 978-1-449-30666-3.

[26] GO, A.; BHAYANI, R.; HUANG, L.: Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, 2009: s. 1–12.

[27] KOSARA, R.: Parallel Coordinates. [online]. 2010. [cit. 2015-05-16]. Dostupné z: `https://eagereyes.org/techniques/parallel-coordinates`

[28] BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R. T.; aj.: LOF. *ACM SIGMOD Record*, ročník 29, č. 2, 2000: s. 93–104, ISSN 01635808. Dostupné z: `http://portal.acm.org/citation.cfm?doid=335191.335388`

[29] TORGO, L.: Package DMwR. [online]. 2015. [cit. 2015-02-20]. Dostupné z: `http://cran.r-project.org/web/packages/DMwR/DMwR.pdf`

[30] PFAHRINGER, B.; DRIESSENS, K.; REUTEMANN, P.: Semi-Supervised Learning and Collective Classification. [online]. [cit. 2015-04-24]. Dostupné z: http://www.cms.waikato.ac.nz/~fracpete/projects/collective-classification/

[31] PFAHRINGER, B.; DRIESSENS, K.; REUTEMANN, P.: Collective Classifiers. [online]. [cit. 2015-04-24]. Dostupné z: http://www.cms.waikato.ac.nz/~fracpete/projects/collective-classification/documentation/classifiers.pdf

[32] BREIMAN, L.: Random forests. *Machine learning*, ročník 45, č. 1, 2001: s. 5–32.

[33] CHANG, C.-C.; LIN, C.-J.: LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, ročník 2, č. 3, 2011: s. 1–27, ISSN 21576904.

[34] POKRAJAC, D.; LAZAREVIC, A.; LATECKI, L. J.: Incremental local outlier detection for data streams. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, IEEE, 2007, s. 504–515.

[35] AHMED, F.; ABULAISH, M.: A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, ročník 36, č. 10-11, 2013: s. 1120–1129. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0140366413001047

[36] Hierarchical Clustering. [online]. [cit. 2015-04-11]. Dostupné z: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html

[37] K-Means Clustering. [online]. [cit. 2015-04-11]. Dostupné z: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

[38] PocasiPraha (Twitter). [online]. 2010. [cit. 2015-03-31]. Dostupné z: https://twitter.com/PocasiPraha/status/1514265563369473

[39] TV_program_CT24 (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/TV_program_CT24/status/591255252403388416

[40] JulezPooh (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/JulezPooh/status/591377830178476032

[41] FotoHanus (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/FotoHanus/status/587312906003423232

[42] webfaircz (Twitter). [online]. 2011. [cit. 2015-04-24]. Dostupné z: https://twitter.com/webfaircz/status/152749051807416320

[43] EngineersDay (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/EngineersDay/status/580353927373303808

[44] big_ben_clock (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/big_ben_clock/status/591376008374329344

[45] vorloj (Twitter). [online]. 2015. [cit. 2015-04-24]. Dostupné z: https://twitter.com/vorloj/status/25819772673

[46] PENNACCHIOTTI, M.; POPESCU, A.-M.: Democrats, republicans and starbucks afficionados. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, New York, New York, USA: ACM Press, 2011, str. 430. Dostupné z: http://dl.acm.org/citation.cfm?doid=2020408.2020477

[47] YU, L.-C.; WU, J.-L.; CHANG, P.-C.; aj.: Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, ročník 41, 2013: s. 89–97. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S095070511300004X

[48] HABERNAL, I.; PTÁČEK, T.; STEINBERGER, J.: Supervised sentiment analysis in Czech social media. *Information Processing & Management*, ročník 50, č. 5, 2014: s. 693–707. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S0306457314000399`

[49] SAIF, H.: Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, the STS-Gold. *CEUR Workshop Proceedings*, ročník 1096, 2013: s. 9 – 21.

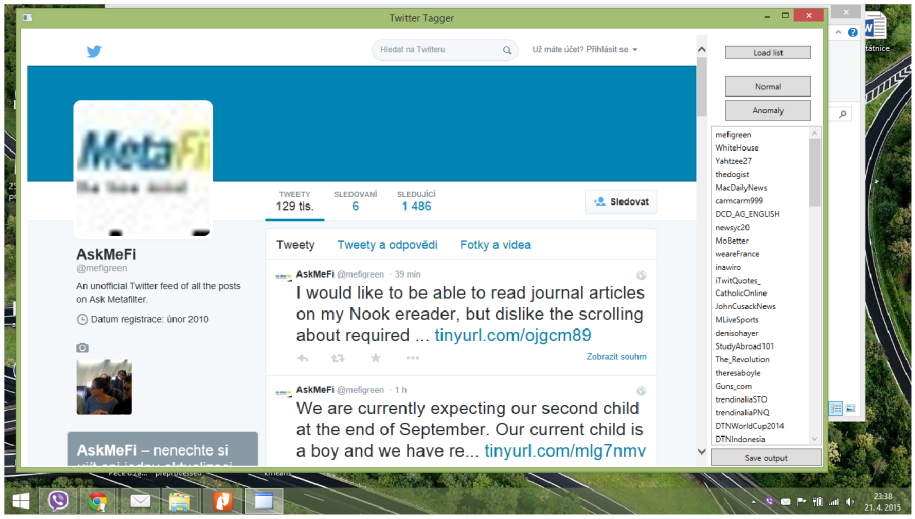
[50] GROOT, R. D.: *Data mining for tweet sentiment classification*. LAP LAMBERT Academic Publishing, 2012, ISBN 978-3-659-29517-1.

A Nástroj pro anotaci uživatelů

Nástroj pro anotaci uživatelů vznikl za účelem zefektivnění práce při anotaci normálních a anomálních uživatelů. Jedná se o jednoduchou aplikaci napsanou v programovacím jazyce C#, která umožňuje procházet a přiřazovat třídu uživatelům ve frontě čekajících na její přiřazení.

Po spuštění programu je nutné načíst frontu uživatelů z textového souboru. Formát textového souboru vyžaduje, aby každý uživatel reprezentovaný jeho uživatelským jménem na sociální síti Twitter byl na samostatném řádku, přičemž uživatelské jméno uvádíme bez uvozujícího symbolu „@“. Anotátor poté pomocí rozhraní programu postupně přiřazuje třídu jednotlivým uživatelům. Jakmile přiřadí třídu poslednímu, uloží seznam uživatelů i s přiřazenými třídami ve formátu CSV, kde oddělovačem je středník. CSV soubor lze pak podrobit dalšímu zpracování.

Aplikace je součástí přílohy práce včetně ukázkového vstupu a výstupu. Obrázek A.1 zobrazuje náhled aplikace při reálném použití.



Obrázek A.1: Nástroj pro anotaci uživatelů

B Obsah přiložených souborů

Součástí práce je příloha, která obsahuje doplňující informace k této práci. Součásti přílohy jsou:

- obrázky použité v textu práce v plném rozlišení
- výstupy všech probíraných experimentů z programu Weka při prováděných experimentech
- grafické výstupy metod
- nástroj pro anotaci uživatelů včetně ukázkového konfiguračního souboru a náhledu

Příloha je dostupná ke stažení v elektronickém archivu závěrečných prací Masarykovy univerzity.