Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5208-56351

Bc. Michal Žilinčík

PRIESKUMNÉ VYHĽADÁVANIE NA SOCIÁLNYCH SIEŤACH SO ZAMERANÍM NA DYNAMICKÉ KRITÉRIA A VZŤAHY METADÁT K OBSAHU

Diplomová práca

Študijný program: Informačné systémy

Študijný odbor: 9.2.6 Informačné systémy

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava

Vedúci práce: prof. Ing. Pavol Návrat, PhD.

máj 2013

ANOTÁCIA

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: INFORMAČNÉ SYSTÉMY

Autor: Bc. Michal Žilinčík
Diplomová práca: Prieskumné vyhľadávanie na sociálnych sieťach so zameraním na dynamické kritéria a vzťahy metadát k obsahu
Vedúci diplomovej práce: prof. Ing. Pavol Návrat, PhD.
máj, 2013

Rôznorodosť obsahu či bezprostrednosť vyjadrení na sociálnej sieti Twitter poskytuje zaujímavý zdroj informácií. Avšak práve diverzita obsahu a krátky rozsah predstavujú problémy pre úlohy vyhľadávania informácií. Pre jednoduchšie nájdenie toho, čo používateľa zaujíma sa existujúce metódy sústreďujú na ohodnocovanie relevancie a kvality príspevkov podľa rôznych charakteristík mikroblogov a ich autorov. Často sa experimentálne snažia nájsť najlepšiu kombináciu charakteristík a ich váh vo funkcii, ktorá ohodnocuje jednotlivé príspevky. V diplomovej práci je na základe analýzy týchto prístupov navrhnutá nová metóda prieskumného vyhľadávania na sieti Twitter. Namiesto objavenia a využitia funkcie ohodnocujúcej príspevky na základe fixných parametrov sa sústredí na snahu vytvoriť spôsob, ako modelovať preferencie používateľa v reálnom čase. To znamená, že používateľ počas prezerania obsahu poskytuje implicitnú a explicitnú spätnú väzbu k príspevkom a na tomto základe sa vytvára model jeho preferencií pre dané sedenie.

ANNOTATION

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: INFORMATION SYSTEMS

Author: Bc. Michal Žilinčík
Master's Thesis: Social networks exploratory search focused on dynamical criteria and relationships of metadata to content
Supervisor: prof. Ing. Pavol Návrat, PhD.
2013, May

Diversity of content created by users of Twitter social network provides an interesting source of information. But the diversity itself and the small space for textual content create issues for information retrieval tasks. Existing methods focus on ranking the relevance and quality of microblogs in order to make it easier for user to find interesting content. They often attempt to find the best combination of characteristics and their weights in ranking functions experimentally. This master's thesis proposes new method for exploratory search on Twitter. Instead of discovering a new ranking function that uses fixed parameters, it focuses on inventing a method for modelling user's preferences in real time. This means that during browsing content, user provides implicit and explicit feedback for tweets and based on this feedback, user's preference model for current session is trained and used for classification of unread tweets into interesting or not interesting.

Prehlásenie

Týmto čestne prehlasujem, že diplomovú prácu „Prieskumné vyhľadávanie na sociálnych sieťach so zameraním na dynamické kritéria a vzťahy metadát k obsahu“ vrátane zdrojových kódov som vypracoval sám, na základe mojich vedomostí a použitej literatúry uvedenej v závere práce.

V Bratislave, 8.5.2013 Michal Žilinčík

Poďakovanie

Týmto by som chcel úprimne poďakovať vedúcemu diplomovej práce prof. Ing. Pavlovi Návratovi, PhD. za jeho pomoc a podporu pri vypracovaní a riešení práce.

Obsah

1 Úvod

S neustálym zvyšovaním množstva obsahu zverejneného na internete je nevyhnuté venovať veľa úsilia metódam vyhľadávania informácií. Nezastupiteľné miesto v tejto oblasti informatiky má prieskumné vyhľadávanie zamerané na používateľa, ktorý nepozná cieľovú doménu, nie je si istý, ako nájsť to, čo hľadá alebo si dokonca nie je istý tým, čo presne chce nájsť.

Informácie zverejňované na internete sa líšia v mnohých aspektoch. Či už ide o ich povahu, tému, formu, obsah alebo zdroj. Špeciálne miesto zastupujú sociálne siete, ktoré v súčasnej dobe zažívajú rozmach. Počet používateľov týchto služieb sa zvyšuje, a tým narastá aj množstvo generovaného verejného obsahu. Od ostatných zverejňovaných informácií sa líšia predovšetkým krátkym rozsahom a silným sociálnym či emocionálnym podfarbením. Takýto obsah pochádzajúci zo sociálnych sietí je často zameraný na autora alebo jeho blízke okolie. To je v kontraste s odbornými článkami v digitálnych knižniciach, odborných webových stránkach alebo informačných portáloch venujúcich sa špecifickým témam. Istý presah je možné sledovať na tzv. blogoch, ktoré poskytujú príležitosť poskytnúť ako informácie k určitej tematike, tak aj autorov osobitý pohľad na vec. Napriek tomuto konceptuálnemu rozdielu neexistuje jasná hranica medzi sociálnym webom a webom zaoberajúcim sa faktickým obsahom. Sociálne siete prinášajú tiež možnosti akéhosi individuálneho pohľadu na konkrétne fakty, udalosti, média a prvky webu. Používatelia môžu obsah odporúčať, hodnotiť, upozorňovať naň a celkovo ho tak obohatiť svojimi vyjadreniami a názormi.

Sprístupnenie možnosti rýchleho a stručného vyjadrenia akejkoľvek myšlienky každému používateľovi internetu zároveň prináša obrovské množstvo neštruktúrovaného až chaotického obsahu. Zatiaľ čo niektoré príspevky používateľov sú relevantné časti jeho známych, iné môžu svoje cieľové publikom hľadať v anonymnom dave podľa spoločných záujmov alebo napríklad aj podobným geografickým umiestnením.

Z toho vyplýva nepopierateľná nutnosť vyvíjať a zdokonaľovať algoritmy, metódy a prístupy k prieskumnému vyhľadávaniu na sociálnych sieťach. Ich obsah je potrebné analyzovať a klasifikovať, aby sa používateľ uskutočňujúci vyhľadávanie dokázal zorientovať v spleti nejasného, stručného obsahu a mohol nájsť práve ten zlomok informácií, ktorý ho zaujme a vie ho zužitkovať.

2 Analýza problému

V tejto časti predstavujem sociálnu sieť Twitter ako vybratú sociálnu sieť pre diplomovú prácu. Je opísaná s dôrazom na charakterizovanie jej dôležitých čŕt a na jej aplikačné rozhranie, pretože pre implementáciu systému pre vyhľadávanie je nevyhnutné poznať základnú štruktúru a obmedzenie poskytovaného API. Ďalej stručne opisujem niekoľko vybratých metód pre optimalizáciu vyhľadávania na webe ako ukážku dosiahnutej a prebiehajúcej práce v oblasti vyhľadávania vo veľkom množstve často rozsiahlych dokumentov. Táto časť slúži predovšetkým pre inšpiráciu k vyhľadávaniu na sociálnych sieťach. Ťažisko analýzy problematiky tvorí prehľad a rozbor práce na postupoch optimalizácie výsledkov vyhľadávania na Twitteri.

V súvislosti so zadaním diplomovej práce, ktorého cieľom je metóda prieskumného vyhľadávania na sociálnej sieti sa zameriavam na dosiahnuté výsledky výskumu v oblasti vyhľadávania a jeho optimalizácie. Súvisí s tým identifikácia charakteristík, ktorých sledovanie umožní zvýšiť kvalitu ponúknutého obsahu. Stručný, neformálny obsah príspevkov a nízka úroveň kvality jazyka tweetov predstavujú komplikácie pre metódy vyhľadávania, preto rozoberám aj postupy prekonávania týchto problémov.

Záver analýzy problému tvorí spresnenie cieľa diplomovej práce a spôsobu jeho dosiahnutia.

2.1 Sociálna sieť Twitter

Twitter je webová služba umožňujúca svojim používateľom uverejňovať príspevky, ktoré nazýva slovom tweet. Dĺžka príspevkov je obmedzená na 140 znakov. Vo všeobecnosti sú všetky tweety verejné, avšak používateľ má možnosť zverejniť ich obmedzenému publiku. V diplomovej práci sa budem zaoberať len verejnými tweetmi. Twitter je možné použiť na vyjadrenie postoja, myšlienok, šírenie odkazov na webové stránky alebo konverzácie s inými používateľmi.

Text tweetu je možné obohatiť dvoma značkami. Tzv. hashtag je spôsob zvýraznenia témy príspevku, napríklad pre zimné olympijské hry (Winter Olympic Games) by anglický hashtag mohol byť #WOG. Hash tag je ale definovaný používateľom, čiže #WOG, #wog, #winterolympics alebo #WOG2010 označujú tú istú udalosť. Použitie hashtagu je dobrovoľné, kvôli čomu nie je možné jednoducho určiť príspevky vzťahujúce sa k rovnakej téme len na základe jedného alebo viacerých použitých hashtagov. Pre hashtagy, ktoré sa začínajú objavovať v čoraz väčšom množstve tweetov Twitter používa označenie trend [1].

Používateľ môže svoj príspevok nasmerovať na jedného alebo viacerých používateľov použitím tzv. mention, čiže akéhosi spomenutia používateľa. Realizuje ho zahrnutím mena používateľského účtu, pred ktoré napíše znak @. Čiže ak používateľ chce v tweete spomenúť používateľov user1 a user2, v texte príspevku použije reťazec „@user1 @user2“. Ak sa tweet

začína touto štruktúrou, je namiesto mention nazývaný reply čiže odpoveď. Ostatným používateľom to môže poskytnúť informáciu o tom, že tweet je viac určený spomenutým používateľov než ostatným napriek tomu, že je verejný [1].

K tweetu je možné vložiť aj geografické umiestnenie, používateľské rozhranie na twitter.com tiež umožňuje jednoduché nahratie obrázku, ktorý sa v tweete zobrazí ako URL tohto obrázku. Tweety je možné označiť za svoje obľúbené (z angl. favorite) [1].

Používateľ tiež môže obsah niekoho iného znovu uverejniť, čo sa nazýva retweet. Používateľské rozhranie Twitteru obvykle zobrazuje pôvodný obsah rovnako, ale pod textom tweetu uvádza, že sa jedná o retweet spolu s menom účtu pôvodného autora. Niektorí používatelia môžu vykonať retweet aj skopírovaním niekoho príspevku, kedy pred pôvodný obsah umiestnia text „RT“, za ktorým sa nachádza pôvodný obsah tweetu. Niekedy pred „RT“ používatelia uvedú aj komentár alebo svoju myšlienku [1].

Pre používateľov poskytuje Twitter ďalšie dve funkcie. Jednou sú zoznamy. Ide o používateľmi vytváraný zoznam používateľov Twitteru. Ak niekto nasleduje zoznam, vidí všetky príspevky od používateľov zaradených do zoznamu. Môžu byť verejné alebo súkromné. Do zoznamu môže používateľ zaradiť kohokoľvek, aj keď ho nenasleduje. Každý zoznam musí mať definovaný názov. Druhou časťou stránky je tzv. Time Line, čo ďalej nazývam nástenkou. Ide totiž o prúd tweetov, ktoré Twitter odporúča každému používateľovi osobitne. Vyberané sú predovšetkým tweety od autorov, ktorých používateľ nasleduje alebo retweety vykonané nasledovanými používateľmi [1].

Pre vyhľadávanie a prístup k jeho výsledkom z programových prostriedkov slúži Twitter Search API. Ide o index reálneho času, ktorý obsahuje najnovšie tweety. Obmedzené je na posledných 6 až 9 dní, staršie tweety nie je možné vyhľadať pomocou tohto API. Veľmi komplexné dopyty sú obmedzované a nevracajú výsledky. Twitter Search API je zamerané na relevanciu a nie na úplnosť, čiže napríklad tweety od používateľov bez mena a opisu v profile sa nemusia nachádzať v indexe. Ani používatelia, ktorí uverejňujú tweety veľmi sporadicky, nerobia retweety a nikdy v príspevkoch nespomínajú iných používateľov nemusia mať svoje tweety zaradené do indexu [2].

Twitter Search API kladie isté obmedzenia pri používaní. Presné časové obmedzenie nie je stanovené, ale závisí od frekvencie a komplexnosti dopytov. Používatelia tohto API by nemali použiť v dopyte viac ako 10 výrazov vrátane slov a tzv. operátorov vyhľadávania. Pre každý tweet vo výsledkoch vyhľadávania je možné získať obsiahnuté hashtagy, URL, čas uverejnenia a spomenutých používateľov. Ak používateľ využil možnosť zahrnúť svoju geografickú polohu, je možné získať aj súradnice. Ak je tweet odpoveďou, obsahujú metadáta aj identifikačné údaje príjemcu. Okrem toho sa tu nachádza tiež URL fotografie používateľa, jeho meno, login a identifikačné číslo [2].

2.2 Metódy vyhľadávania na webe zamerané na personalizáciu

V tejto časti sa zameriavam na metódy a algoritmy, ktorých cieľom je optimalizovať výsledky vyhľadávania analyzovaním práce používateľa s vyhľadávačom. Obvykle ide o analýzu vyhľadávacích záznamov, čím je možné zistiť, ktoré výsledky používateľa zaujali, a tak získať ďalšie informácie o kvalite či zaujímavosti výsledkov. Tieto metódy sa používajú predovšetkým pri prehľadávaní webu, ktorého charakter sa môže líšiť od sociálnych sietí. Niektoré aspekty optimalizácie je ale možné vztiahnuť aj na vyhľadávanie obsahu pochádzajúceho zo sociálnych sietí alebo obohatiť vďaka nim krátke príspevky na sociálnych sieťach. Práca a poznatky o vyhľadávaní sú v tejto oblasti ďalej ako v prípade sociálnych sietí, čím niektoré aspekty týchto prístupov tvoria silnú inšpiráciu práve pre vyhľadávanie na sociálnych sieťach.

2.2.1 Spätná väzba používateľa podľa návštev výsledkov vyhľadávania

Pre vyhľadávanie na webe existuje viacero vyhľadávačov, ktoré obsahujú algoritmy a mechanizmy určené na optimalizáciu výsledkov vyhľadávania. Tie sú obvykle založené na výskyte slov dopytu v indexovaných dokumentoch, inými slovami sú zamerané na obsah. QueryFind je navrhnutá metóda, ktorá výsledky respektíve poradie výsledkov z takého vyhľadávača berie ako základ pre hodnotenie založené na obsahu. Okrem toho sa berie do úvahy aj hodnotenie založené na spätnej väzbe používateľa, ktorá je získaná z počtu kliknutí používateľov na základe záznamov vyhľadávačov. Výsledné hodnotenie stránky je teda priamo úmerné obom hodnoteniam [3].

Metóda rozlišuje kliknutia na tú istú stránku podľa aktuálneho dopytu. Čiže ak sa zmení dopyt, môže sa zmeniť aj hodnotenie založené na počte kliknutí, pretože sa berú len kliknutia, ktoré boli vykonané pri takom istom dopyte ako je aktuálny. Súčasne je známe, že spätná väzba v kolaboratívnom vyhľadávaní je založená na väčšinovom prístupe v hlasovaní. Čiže to, čo sa páči väčšine používateľov dostáva najvyššie ohodnotenie. QueryFind sa na výsledné hodnotenia jednotlivých výsledkov vyhľadávania pozerá ako na lineárnu kombináciu normalizovaných hodnotení založených na spätnej väzbe a na obsahu. Základ tvorí pohľad z prístupov frekvencie výrazov (z angl. Term Frequency) a inverznej frekvencie v dokumentoch (z angl. Inverse Document Frequency). Kde prvú zložku reprezentuje hodnotenie založené na spätnej väzbe, na mieste inverznej frekvencie v dokumentoch je potom hodnotenie vypočítané z poradia výsledkov získaných z vyhľadávača. Použitý je aj predpoklad, že hodnotenie používateľmi je dôležitejšie ako poradie z vyhľadávača. Experimenty ukázali, že QueryFind dosahuje dobré výsledky, predovšetkým pri populárnych dopytoch [3].

Podobný pohľad na optimalizáciu výsledkov vyhľadávania majú aj autori metódy CRRA [4] čiže Kolaboratívneho prístupu k znovu ohodnocovaniu výsledkov vyhľadávania. Metóda je založená na predpoklade existencie skupín používateľov s podobnými záujmami, požiadavkami, očakávaniami a motiváciou k hľadaniu podobného obsahu. Autori zdôrazňujú tri kľúčové entity vo webovom vyhľadávaní: používateľ, dopyt a výrazy v dokumentoch. Všetky tri entity vplývajú na

kontext dopytu, pretože napríklad používatelia zadávajúci rovnaký dopyt môžu mať úplne iné potreby voči hľadaným informáciám. Aplikáciou pravdepodobnostného modelu vzťahov medzi týmito entitami je možné vybudovať používateľské modely a vypočítať tzv. komunity používateľov [4].

CRRA vo fáze predspracovania analyzuje záznamy z vyhľadávania a zostavuje maticu, v ktorej riadky tvoria dopyty, stĺpce predstavujú jednotlivé výrazy v dokumentoch z výsledkov vyhľadávania a prvkami sú pravdepodobnosti kliknutia na dokument s daným výrazom v stĺpci pri dopyte v riadku. Vo fáze učenia sa používateľských profilov je vytvorená podobná matica, avšak riadok obsahuje identifikáciu používateľa a identifikáciu dopytu. Čiže v tejto fáze sú vypočítané pravdepodobnosti kliknutia konkrétneho používateľa pri konkrétnom dopyte na dokument s daným výrazom. Oproti prístupu spomínanému vyššie je tu počas ohodnocovania výsledkov vyhľadávania obohateného o spätnú väzbu používateľov prítomný aj pojem komunity. Komunita je množina používateľov, ktorí pri najviac podobných dopytoch klikali na dokumenty s podobnými obsiahnutými výrazmi. Hodnotenie výsledkov vyhľadávania je potom kombináciou používateľovej aktivity (klikania na dokumenty) a aktivitou používateľov v jeho komunite. Zaujímavý výsledok je, že CRRA v porovnaní s inou metódou založenou na analýze aktivity používateľov síce dáva len o niečo lepšie výsledky, dosahuje však omnoho vyššiu stabilitu, čiže jej výkon veľmi nekolíše pri rôznych dopytoch [4].

2.2.2 Optimalizácia zoradenia výsledkov vyhľadávania

Personalizáciu vyhľadávania je možné vykonať aj iným prístupom než analýzou kliknutí používateľov vo vyhľadávacích záznamoch, ako vo svojej práci o automatizovanom parametrizovaní prispôsobiteľnej funkcie ohodnocujúcej vyhľadávanie [5] ukázali Pohorec, Čeh, Zorman a Kokol.

Ako vstup použili dva spôsoby spätnej väzby od používateľov. Prvým bola možnosť preusporiadať výsledky vyhľadávania podľa relevancie k ich dopytu, druhou alternatívou voľba vyhľadávača, ktorý už výsledky zoradil v požadovanom poradí. Prispôsobiteľnosť hodnotiacej funkcie spočívala vo výbere metrík, parametrizovateľnosť je vnímaná ako výber váh jednotlivých metrík. Pre účel svojej metódy použili genetický algoritmus, ktorého cieľom bolo zistiť, ktoré z dostupných metrík by mali byť použité vo výslednej funkcii a aké váhy by im v tejto funkcii mali prislúchať. Ako príklad sú ale uvádzané len dve metriky: či sa dopyt nachádza v názve výsledku vyhľadávania a či bol vo výsledkoch vyhľadávania dopyt nájdený vo svojej pôvodnej podobe. Počiatočná populácia pre genetický algoritmus sa zostavuje dvoma spôsobmi. Časť tvoria náhodne vygenerovaní kandidáti a druhá časť kandidátov je náhodne vybratá spomedzi výsledných kandidátov algoritmu, ktorý bol vykonaný pre ostatných používateľov. Z populácie sa vyberá 5% najlepších kandidátov, pričom je ako funkcia spôsobilosti (z angl. fitness function) použitá vzdialenosť zoradenia výsledkov podľa kandidáta od zoradenia používateľa. Podarilo sa im ukázať,

že genetický algoritmus dokáže vygenerovať požadovanú hodnotiacu funkciu pri splnení predpokladov, že je definovaný dostatočný počet a typ metrík, že požadovaný zoradený zoznam výsledkov je dostatočne dlhý a že je možné nájsť výsledky pre zadaný dopyt [5].

2.2.3 Relevancia podľa kontextu dopytu

Jedným z problémov vyhľadávania je vzťah dopytu k zamýšľanému cieľu vyhľadávania. K zadanému dopytu používateľa môže existovať veľké množstvo dokumentov, z ktorých niektoré vyhľadávací mechanizmus považuje za vysoko relevantné, ale napriek tomu nevyhovujú cieľu používateľa. Relevantné dokumenty môžu byť nové, avšak ich hodnotenie je z toho istého dôvodu nízke. Metóda Q-Rank [6] berie do úvahy aj fakt, že používatelia často svoje dopyty modifikujú, kým poskytnuté výsledky zodpovedajú ich predstave alebo obsahujú také dokumenty, ktoré uspokoja ich zámer [6].

Zmenu dopytu používateľ vykonáva napríklad vtedy, keď chce dopyt zovšeobecniť, spresniť alebo pridať nové informácie. Vzájomný vzťah dopytu s dopytom, ktorý po ňom nasleduje je preto možné považovať za dôležitú stopu k pochopeniu informačnej potreby používateľa. Q-Rank je založený na princípe, že najčastejšie použité rozšírenia dopytov (výrazy, ktoré sú pridané k dopytom) a susedné dopyty (dopyty, ktoré sú použité tesne pred alebo po danom dopyte) jasne poukazujú na zámer používateľov. Funkcia hodnotiaca výsledky vyhľadávania sčítava skóre pre rozšírenia dopytov a susedné dopyty. Pre oba je počítané TF-IDF skóre (pre výskyt rozšírení dopytu alebo susedných dopytov vo všetkých dokumentoch z výsledkov vyhľadávania), ktoré má váhu pozostávajúcu z logaritmu normalizovaného výskytu dopytu, čím sa adresuje rôzna popularita dopytov. Oba výrazy sú vydelené pôvodným hodnotením (podľa poradia z vyhľadávača), ktoré je dôležité vďaka tomu, že už zahŕňa hodnotenie podľa rôznych charakteristík, ktoré Q-Rank neanalyzuje. Práca ukazuje, že tento prístup zvyšuje kvalitu vyhľadávania práve pre tzv. informačné dopyty, kedy sa používateľ snaží po zadaní dopytu dopátrať k súvisiacim informáciám [6].

Pre vyhľadávanie na sociálnych sieťach je priame použitie tejto metódy obmedzené. Predovšetkým za predpokladu, že pre nízky počet znakov v príspevkoch, neobsahuje jeden príspevok dostatočný počet slov na to, aby mohol byť relevantný pre toľko dopytov ako webová stránka. Na druhej strane však Twitter API vracia menší počet výsledkov ako webové vyhľadávače. Spomínaná metóda je preto zaujímavým kandidátom pre rozšírenie súboru výsledkov o ďalšie relevantné výsledky, ktoré je možné získať použitím rozšírených a susedných dopytov. Otázkou ale stále ostáva zdroj pre získavanie týchto dopytov, pretože pokiaľ viem, neexistuje analýza zaoberajúca sa aplikovaním metódy na sociálne siete a efektivitou použitia záznamov z webového vyhľadávania. Na určenie kontextu či zámeru používateľa by ale mohlo byť užitočné použiť aj analýzu jeho často navštevovaných stránok, stránok v jeho záložkách vo webovom prehliadači alebo využiť princíp zo spomínanej práce, avšak aplikovať ho v kontexte jedného sedenia. To

znamená, že systém pre prieskumné vyhľadávanie na sociálnej sieti by po modifikovaní aktuálneho dopytu nemusel zahodiť aktuálne výsledky vyhľadávania, ale naopak ich obohatiť výsledkami nového dopytu so súčasným použitím zmeny ohodnotenia všetkých zobrazovaných výsledkov, čiže relevancia zobrazených príspevkov by závisela od relevancie príspevku k obom (alebo viacerým) dopytom.

2.3 Metódy vyhľadávania a hodnotenia príspevkov na sociálnej sieti Twitter

V tejto časti sa venujem doterajším poznatkom a postupom pre ohodnocovanie príspevkov na základe ich obsahu, autora alebo iných špecifických faktorov súvisiacich so sociálnou sieťou Twitter. Ďalej opisujem metódy, ktoré sú zamerané viac špecificky na konkrétny aspekt príspevkov a postupy, ktoré využívajú ďalšie zdroje informácií umožňujúce vylepšenie hodnotenia príspevkov alebo poskytujú iný druh optimalizácie výsledkov vyhľadávania na sociálnej sieti.

2.3.1 Metódy zamerané na príspevky a autora

Jednou z prvých akademických prác, ktoré sa snažia navrhnúť stratégiu pre ohodnocovanie príspevkov zo sociálnych sietí je Ranking Approaches for Microblog Search [7] z roku 2010. Motiváciou k navrhnutiu metódy ohodnocujúcej príspevky pochádzajúce z vyhľadávača pre sociálnu sieť je práve absencia zoradenia príspevkov podľa relevancie. Twitter API dnes síce ponúka možnosť vyhľadať najviac relevantné tweety namiesto najnovších týkajúcich sa daného dopytu, výsledky sú však takmer výhradne zoradené reverzne chronologicky. Pre používateľa je nespornou výhodou čítať najviac relevantné príspevky namiesto náhodne usporiadaných viac či menej relevantných. Navrhovaná metóda skúma sociálnu sieť Twitter, ale navrhuje stratégie, ktorých použitie nie je silno viazané len na túto sociálnu sieť. Zavádza dve kritériá týkajúce sa autorov tweetov, ktoré spolu s niektorými charakteristikami samotných príspevkov ako je ich dĺžka alebo fakt, či obsahujú URL tvoria základ pre ohodnocovanie tweetov [7].

Pri ohodnocovaní podľa autorov príspevkov je do úvahy braná myšlienka, že aktívni používatelia skrývajú väčší potenciál pre cenné príspevky. Na tomto základe je postavené jednoduché hodnotenie v počte uverejnených príspevkov. Druhým predpokladom je, že užitočnosť príspevkov autora sa odvíja od jeho pozície v sieti. Čím viac ľudí nasleduje autora, tým je jeho postavenie lepšie. Použitý je Vzorec 2.1, v ktorom *i(a)* označuje počet používateľov, ktorí nasledujú autora *a*, *o(a)* je počet používateľov, ktorých autor *a* nasleduje. Účelom zahrnutia hodnoty *o(a)* je zmierniť hodnotu kritéria FollowerRank pre autorov, ktorí majú veľa nasledovateľov nie pre kvalitu svojich príspevkov, ale pre svoju vysokú sociálnu aktivitu (čiže nadväzovanie a udržovanie vzťahov s inými používateľmi vo veľkej miere). Zavedené kritérium ale napríklad neuvažuje, že vplyv autora sa môže meniť v závislosti od oblastí tém [7].

$$\mathrm{FR}(a) = \frac{\mathrm{i}(a)}{\mathrm{i}(a) + \mathrm{o}(a)}$$

Vzorec 2.1 FollowerRank [7].

Analýza príspevkov uvažuje ich dĺžku, ktorá do hodnotenia vstupuje ako pomer počtu znakov príspevku k počtu znakov v najdlhšom príspevku z výsledkov vyhľadávania. Okrem toho je definovaný aj binárny ukazovateľ, ktorý nadobúda hodnotu 1, ak príspevok obsahuje URL, inak nadobúda hodnotu 0. Nevýhodou tohto ukazovateľa je, že na jeho základe nie je možné príspevky zoraďovať, keďže nadobúda len dve hodnoty. V práci sú definované aj charakteristiky kombinujúce uvedené kritériá. Ich úspešnosť je meraná podľa zhody ohodnotenia dvojíc tweetov na základe charakteristiky a na základe manuálneho ohodnotenia človekom. Úspešnosť 62% dosahuje charakteristika sčítavajúca FollowerRank, hodnotenie podľa dĺžky príspevku a hodnotenie podľa výskytu URL. Tieto tri kritéria je možné získať veľmi rýchlo pri použití Twitter API pre vyhľadávanie, čo robí túto stratégiu hodnotenia vhodnú pre vyhľadávanie v reálnom čase [7].

Podobný prístup k ohodnocovaniu tweetov má aj práca An Empirical Study on Learning to Rank of Tweets [8], ktorej cieľom je podať dôkaz o pozitívnom vplyve ohodnocovania tweetov podľa vplyvu autora, dĺžky tweetov a výskytu URL, pričom sa snaží za vplyv autora zaviesť lepšiu metriku ako počet nasledovateľov [8].

Learning to Rank nazývajú autori práce prístup založený na údajoch, ktorý v modeli efektívne zahŕňa množinu charakteristík. Z toho vyplýva nutnosť tieto charakteristiky zvoliť, pokiaľ možno čo najvýhodnejšie. Rozdeľujú ich do troch skupín [8]:

1. charakteristiky relevancie obsahu označujú také metriky, ktoré opisujú relevanciu tweetov k dopytu,
2. charakteristiky špecifické pre Twitter sú také, ktoré sa pozerajú na údaje, ktoré je možné nájsť len v tweetoch, napríklad počet opakovaného uverejnenia tweetu (čiže retweet),
3. charakteristiky vplyvu autora pojednávajú o ukazovateľoch hodnotiacich akúsi autoritu a vplyv používateľov uverejňujúcich tweety.

K charakteristikám relevancie obsahu patrí napríklad ukazovateľ Okapi BM25, ktorý analyzuje vzťah príspevkov k dopytu. Na to používa vzorec, ktorý pre slová z dopytu kombinuje ich TF a IDF skóre. Druhou charakteristikou je podobnosť obsahu, ktorá sleduje počet tweetov s podobným obsahom. Využíva kosínusovú mierku podobností TFIDF vektorov dvojíc tweetov. Napokon ako miera informatívnosti tweetu slúži počet jeho slov [8].

Medzi charakteristikami špecifickými pre Twitter je aj parameter, ktorý hovorí o prítomnosti URL v tweete. Ďalej sa sleduje aj počet tweetov s rovnakým URL. Počet retweetov je sledovaný na základe zhody textu v tweete s textami iných tweetov za dvojicou znakov „RT“. To znamená, že iba presne odcitované retweety zvyšujú počet retweetov v tejto metóde. Nasleduje hashtag skóre, ktoré sa odvíja od frekvencie výskytu hashtagu v analyzovanom korpuse. Takzvaná odpoveď je binárny ukazovateľ, ktorý nesie hodnotu podľa toho, či je tweet odpoveďou alebo nie. Názov ukazovateľa OOV nesie význam „mimo slovníka“ (z angl. out of dictionary). V práci je

použitý na hrubý odhad kvality jazyka tweetov. Slová mimo slovníka v tweetoch obsahujú okrem iného preklepy a pomenované entity, kde podľa malého výskumu sú za 90% slov mimo slovníka zodpovedné práve preklepy nerátajúc slová napísané veľkými písmenami, hashtagy, spomenutia používateľov a URL. Kvalita jazyka je teda vypočítaná ako podiel slov mimo slovníka voči počtu slov v tweete [8].

Autori práce sledovali vplyv používateľov Twitteru na základe dvoch predpokladov. Prvým je, že používatelia často spomínaní v tweetoch, nachádzajúci sa vo viacerých zoznamoch, majúci veľa znovu uverejnených tweetov dôležitými používateľmi a majúci veľký počet nasledovateľov sú viac vplyvní. Druhý predpoklad tvrdí, že tweet má vyššiu pravdepodobnosť byť informatívny, ak je uverejnený alebo znovu uverejnený vplyvným používateľom. Aby bolo možné rozoznať vplyv uvedených faktorov, boli vytvorené štyri nezávislé ukazovatele. Follower Score je počet nasledovateľov používateľa, Mention Score predstavuje počet spomenutí používateľa v tweetoch, List Score hovorí o počte zoznamov, v ktorých bol používateľ uverejnený a Popularity Score je vypočítaný PageRank algoritmom operujúcim nad grafom, v ktorom vrcholmi sú používatelia a ak používateľ V_1 urobí retweet príspevku používateľa V_2, orientovaná hrana smeruje od V_1 k V_2 [8].

Viac ako 81% vyhľadávaní tvoria témy týkajúce sa noviniek o známych osobách alebo týkajúce sa konkrétnych miest, témy týkajúce sa produktov, čiže ich opisu alebo komentárov k nim a témy týkajúce sa zábavy predovšetkým o filmoch vrátane ich recenzií a úvodom k zápletke filmu. Na tomto základe bolo postavené overenie opisovanej metódy, preto boli vybraté dopyty týkajúce sa týchto oblastí tém. Retweety neberú autori práce do úvahy, pretože neprinášajú žiadne nové informácie než pôvodný tweet. Spomedzi troch spomínaných skupín charakteristík sa ukázalo použitie charakteristík relevancie obsahu ako najmenej prínosné. Autori to odôvodnili tým, že len ohodnocujú výsledky vyhľadávania zabudovaného do Twitteru, ktorý relevanciu určuje podľa obsahu dopytu v tweetoch, čiže napríklad BM25 nespôsobí takmer žiadnu zmenu v ohodnotení relevancie príspevkov. Naopak ako najužitočnejšia sa ukazuje skupina charakteristík podľa vplyvu používateľa. Najlepšie výsledky ale dosahuje model využívajúci SVM natrénovaný na všetkých charakteristikách. Pri pohľade na prínos jednotlivých metrík sa prítomnosť URL javí ako nekompromisne najdôležitejší ukazovateľ. Po ňom je ešte veľmi dôležitý aj počet uverejnení autora tweetu v zozname. Užitočné je použiť aj počet spomenutí používateľa a sledovať výšku hodnotenia najlepšie hodnoteným používateľom, ktorý spravil retweet daného tweetu (použité v Popularity score na základe algoritmu PageRank). Nízka užitočnosť použitia hashtag skóre alebo počtu retweetov môže byť ale podľa autorov práce spôsobená aj malým trénovacím súborom [8].

Zaujímavý pohľad na triedenie príspevkov poskytuje práca [9] zaoberajúca sa fenoménom retweetov. Autori svoju prácu delia na ohodnocovanie tweetov podľa pravdepodobnosti, že ich niekto znova uverejní a na určenie používateľov, ktorým by sa skúmaný tweet mal zobraziť, pretože existuje vysoký predpoklad, že vykonajú jeho retweet [9].

Autori zvolili podobné rozdelenie charakteristík ako vo vyššie opisovanom článku. Na autora tweetu sa pozerajú cez jeho popularitu, či už na tzv. lokálnu popularitu pri 10 000 až 50 000 nasledovateľoch alebo globálnu pri ich vyššom počte. Skúmajú aj vek používateľa v kontexte jeho existencie na sociálnej sieti, počet uverejnených príspevkov, počet obľúbených príspevkov, frekvenciu prispievania, počet uverejnení v zozname, či ide o overeného používateľa, či jeho profil obsahuje aj opis používateľa, URL a či je jeho jazykom angličtina [9].

Pri tweetoch sa metóda zameriava na to, či obsahujú hashtagy, URL, či spomínajú používateľov, či ide o retweet alebo či tweet samotný je retweet. Sleduje sa aj TFIDF skóre tweetu a jeho dĺžka. Charakteristiky sú oproti predchádzajúcemu postupu obohatené aj o sledovanie výskytu otáznika, výkričníka, úvodzoviek, zámena prvej osoby jednotného čísla, emotikon a znakov opakovaných viac ako trikrát za sebou (napríklad slovo cooool) [9].

V súvislosti s obsahom sa sleduje novota príspevku voči ostatným príspevkom, ktoré používateľ vidí na svojej Twitter nástenke meraná kosínusovou mierkou podobnosti. Tiež sa sleduje na koľko je neočakávané, že daný tweet jeho autor uverejnil (podľa podobnosti s ostatnými tweetmi autora). Použitý je aj ukazovateľ založený na pravdepodobnostnom jazykovom modeli autora tweetu. Ak analyzujeme používateľa za účelom určenia, ktoré tweety sa mu majú na nástenke zobraziť berie sa do úvahy, či s autorom ohodnocovaného tweetu komunikuje, či ho označil do zoznamu, či vykonal retweet jeho tweetu, či použil hashtag alebo doménu z URL v ohodnocovanom tweete, prípadne či autor spomenul používateľa alebo naopak [9].

Z overenia metódy určovania pravdepodobnosti retweetov vyplynulo, že spomedzi charakteristík založených na príspevkoch sa ako najefektívnejšie ukázalo sledovať prítomnosť URL, či je tweet retweetom a TFIDF skóre. Z kategórie založenej na používateľovi to bola analýza toho, či sa používateľ s autorom zapája do rozhovorov a či používateľ použil vo svojom tweete URL s rovnakou doménou ako URL v analyzovanom tweete [9].

Téma vplyvu používateľov Twitteru patrí k skúmaným faktorom, ktoré umožňujú zvýšiť kvalitu ohodnocovania tweetov. TwitterRank [10] je metóda, ktorá sľubuje lepšie výsledky pri odhaľovaní vplyvných používateľov podľa tém než algoritmus používaný Twitterom.

Približne 72% používateľov Twitteru nasleduje viac ako 80% svojich nasledovateľov. Tieto vzájomné vzťahy môžu mať dve vysvetlenia. Nasledovanie môže byť používateľmi vnímané ako neformálny vzťah, kedy jeden používateľ začne nasledovať druhého a ten druhý bude nasledovať prvého zo zdvorilosti. Alebo to môže byť práve naopak, čiže nasledovanie je silným ukazovateľom podobnosti používateľov. Napríklad používateľ nasleduje kamaráta, pretože ho zaujímajú témy, o ktorých píše a kamarát si všimne, že obaja majú záujem o podobné témy, preto bude tohto používateľa tiež nasledovať. V tomto prípade ide o fenomén s anglickým názvom *homophily*, čo je tendencia jednotlivcov združovať sa s inými podobnými jednotlivcami. Podarilo sa dokázať, že tento fenomén na Twitteri existuje, čiže existujú používateľa nasledujúci iných kvôli spoločnému záujmu o podobné témy. Na tomto základe postavili TwitterRank ako rozšírenie

algoritmu PageRank, ktorý berie do úvahy štruktúru sociálnej siete a podobnosť tém medzi autormi na Twitteri [10].

Pre identifikáciu tém, o ktorých používatelia Twitteru píšu, a teda sa o ne zaujímajú je použitý model strojového učenia bez učiteľa LDA (Latent Dirichlet Allocation). Je založený na reprezentácií tém ako pravdepodobnosti distribúcie slov v dokumentoch. Výstupom je reprezentácia dokumentov ako pravdepodobností ich príslušnosti k témam. Vplyv používateľa Twitteru je podobný ako vplyv web stránky, kedy vplyv používateľa rastie s vplyvom jeho nasledovateľov. Vplyv používateľa na jeho nasledovateľov sa dá vnímať ako relatívne množstvo obsahu, ktoré od neho jeho nasledovatelia dostali. To je dôvod, prečo je PageRank motiváciou pre návrh podobného algoritmu pre Twitter. Existujú tu však aj rozdiely, pretože používatelia majú rôzne poznatky v rôznych témach a nie každý príspevok môže byť zaujímavý a niektoré tweety si používatelia vôbec nemusia prečítať od autora, ktorého nasledujú. TwitterRank je preto zameraný aj na témy [10].

Vrcholy grafu, s ktorým algoritmus pracuje tvoria používatelia a hrany sú orientované od nasledovateľa k nasledovanému používateľovi. Ohodnocovanie grafu je vykonávané náhodným prechádzaním grafu po hranách. Rozdiel oproti PageRank algoritmu spočíva v tom, že pravdepodobnosť prechodu cez konkrétnu hranu sa zvyšuje so zvyšujúcimi sa spoločnými záujmami používateľov, ktorí sú hranou spojení. Praktické experimenty preukázali, že spomedzi podobných algoritmov založených na PageRank a algoritme Twitteru založeného na počte nasledovateľov dosahuje TwitterRank najlepšie výsledky [10].

Zaujímavým spôsobom zvyšovania kvality výsledkov vyhľadávania je aj sémantická analýza s následným klastrovaním príspevkov. Opisovaný postup [11] prináša lepšie výsledky vďaka použitiu sémantickej podobnosti namiesto triviálneho počítania výskytu rovnakých slov.

K prvému kroku algoritmu patrí získanie maximálneho počtu príspevkov z Twitter API (konkrétne 1500) pre zadaný dopyt. Nasleduje zostavenie matice pre výrazy a dokumenty, kde dokumenty predstavujú tweety. Tweety predstavujú stĺpce, riadky tvoria výrazy bez stop slov. Pre vytvorenie klastrov sa používa algoritmus faktorizácie matice NMF (non-negative matrix factorization). Ďalším krokom je výpočet súdržnosti klastrov na základe vzájomnej podobnosti všetkých tweetov v klastri. Vo výsledkoch vyhľadávania sa zobrazia len príspevky z tých klastrov, ktorých súdržnosť je vyššia ako zvolená hraničná hodnota. Klastre sú zobrazené podľa veľkosti zostupne. Pre tento posledný krok zaviedli autori aj alternatívu, pri ktorej nie sú zobrazené všetky príspevky z klastru, ale len jeden reprezentatívny tweet, aby používateľ dostal prehľad o hlavných témach súvisiacich s dopytom [11].

Opis metód založených na analýze používateľov a obsahu tweetu uzavriem prácou o prekonávaní riedkosti informácií v mikroblogoch a nízkej kvality dokumentov [12].

V bežnom scenári vyhľadávania informácií sú súkromné a osobné správy medzi používateľmi nie veľmi zaujímavé, pretože nespĺňajú konkrétnu potrebu pre nájdenie

požadovaných informácií. Z toho vyplýva, že pri vyhľadávaní tweetov by hlavným kritériom mala byť zaujímavosť tweetu. Klasické modely vyhľadávania ale neuvažujú tieto aspekty, pretože sú zamerané na dlhé texty a predpokladajú, že dokument je primárne určený na prenos informácií. Podobne je to aj pri súčasných snahách využiť na Twitteri prístupy, ktoré sa používajú primárne na stanovanie vplyvu web stránok podľa citácie alebo hyperlinkov (napríklad adaptácia PageRank pre sociálnu sieť). Neberú totiž do úvahy, že sémantika sociálnej siete nie je ekvivalentná sieti, kde je odkazovanie zamerané vyslovene na obsah [12].

Prvým adresovaným problémom v tejto práci je opodstatnenosť normalizácie dĺžky dokumentov vzhľadom na riedky výskyt výrazov v tweetoch. Výskum v článku ukazuje, že mikroblogy vo všeobecnosti obsahujú len niekoľko výrazov a iba veľmi zriedka sa v nich jeden výraz vyskytuje viackrát. Pri moderných modeloch vyhľadávania informácií je ich základnou zložkou normalizácia dĺžky dokumentu. To je motivované snahou eliminovať potenciálnu výhodu dlhších dokumentov. Táto výhoda sa vysvetľuje na základe dvoch hypotéz. Prvá hovorí o tom, že dlhý dokument spracováva tú istú tému vo väčšom rozsahu a opakovane. Čiže tie isté výrazy sa v ňom nachádzajú v hojnejšom počte bez toho, aby dokumentu pridávali ďalšie informácie. Druhou je hypotéza rozsahu a tvrdí, že dlhý dokument adresuje viac tém. To znamená, že dlhý dokument obsahuje rozličnejšie výrazy, kvôli ktorým sa môže dokument zdať relevantným k širokému rozsahu tém. Vo všeobecnosti je ale možné predpokladať, že používateľ preferuje krátky dokument zameraný priamo na tému pred dlhým dokumentom pojednávajúcim o viacerých témach [12].

Autori práce vykonali experiment a zistili, že pre tweety tieto hypotézy s vysokou pravdepodobnosťou neplatia. Preto vykonávať normalizáciu dĺžky tweetov nie je potrebné. Dokonca to môže byť kontraproduktívne, pretože tento prístup potom uprednostňuje krátke mikroblogy pred dlhšími, čo ale nemá opodstatnenie [12].

Druhou témou článku je definícia zaujímavosti ako statického ukazovateľa kvality. Robia tak cez pravdepodobnosť, že daný tweet bude znova uverejnený iným používateľom. Do úvahy ale pripadajú len charakteristiky založené na obsahu tweetu. Prístup nepozerá na autora príspevku ani na čas jeho uverejnenia. K binárnym ukazovateľom patria URL, mená používateľov a hashtagy v tweete. Ďalšími dvoma je aj použitie výkričníka a otáznika na konci príspevku. Sledujú sa aj emotikony a pre výskyt negatívnych a pozitívnych emócií sa tiež zavádza dvojica binárnych ukazovateľov, podobne je to aj pri výskyte pozitívnych a negatívnych výrazov podľa krátkeho preddefinovaného zoznamu. Autori potom definujú aj charakteristiky založené na citovom podfarbení tweetu, ktoré je určené podľa jednoduchého prístupu s použitím slovníka Affective Norms of English Words [13]. Charakteristika podľa príslušnosti textu k témam používa už vyššie spomenutý algoritmus LDA. Na záver sa analyzujú aj výrazy v tweete podľa pravdepodobnostnej Bayes teorémy. Na základe rozboru príspevkov v korpuse a výskytu výrazov v tweetoch a retweetoch sa určuje pravdepodobnosť, že analyzovaný tweet bude tiež retweet. Tieto

charakteristiky tvoria trénovaciu množinu pre logistický regresný model, ktorý umožní pre tweet získať pravdepodobnosť retweetu. Tá je interpretovaná ako kvalita mikroblogu [12].

Na základe experimentálneho overenia sa podarilo dokázať, že normalizácia dĺžky je kontraproduktívna. Pri krátkych a málo špecifických dopytoch dosahuje jednoduché ohodnocovanie podľa relevancie slabé výsledky. Začlenenie zaujímavosti ako statického ukazovateľa kvality vedie k veľkému zlepšeniu. To motivuje k jej použitiu, pretože štatistiky ukazujú, že dopyty na Twitteri sú priemerne dlhé 1,64 slova. Využitie zaujímavosti ako ukazovateľa dosahuje najlepšie výsledky pri jednoslovných dopytoch. Vyhodnotenie experimentu ale ukazuje, že účinok tohto statického ukazovateľa kvality pri dlhších dopytoch je zoslabený a môže medzi najlepšie výsledky vyhľadávania zaradiť síce veľmi zaujímavé, ale zároveň príspevky irelevantné k dopytu. Preto aj autori odporúčajú použiť pravdepodobnosť retweetu pri nešpecifických a krátkych dopytoch [12].

Uvedené výsledky zreteľne hovoria, že pri analýze relevancie tweetov k dopytu je zlou taktikou penalizácia dlhých tweetov. Podobne ako je tomu pri ostatných prácach, aj tu sa ukazuje pravdepodobnosť vykonania retweetu ako dobrá taktika. Avšak článok Searching Microblogs: Coping with Sparsity and Document Quality [12] jasne poukazuje na jej obmedzenie pri špecifických dopytoch. Práca [8] využívajúca adaptáciu PageRank algoritmu na určovanie pravdepodobnosti retweetov je založená na veľmi podobných charakteristikách a ukazuje, že tento prístup dosahuje vo všeobecnosti dobré výsledky. V prvom prípade autori nepoužili charakteristiky súvisiace s autorom tweetu, v druhom prípade boli použité aj takéto ukazovatele. Tento nesúlad ale pravdepodobne s týmto faktom nesúvisí, pretože najužitočnejšími charakteristikami sa ukázali práve tie, ktoré sú založené na metadátach tweetu a jeho obsahu. Autori ale v druhom prípade neskúmali vplyv dĺžky dopytu na výkon metódy, preto nie je možné určiť, či práve dĺžka dopytu nespôsobuje odlišné výsledky.

2.3.2 Využitie rozmanitosti informácií

V tejto časti uvádzam osobitne jeden výskum [14], ktorý síce používa podobné charakteristiky ako v predchádzajúcej časti, avšak jeho cieľom nie je jednoznačne zoradiť výsledky vyhľadávania na Twitteri podľa ich relevancie. Dôraz je kladený na rozmanitosť informácií podľa preferencií používateľa. Metóda overenia tiež ide hlbšie než explicitná spätná väzba formou porovnávania dvojíc príspevkov.

Môžeme intuitívne predpokladať, že pri vyhľadávaní a preskúmavaní obsahu sociálnych médií sa môže používateľ zaujímať o homogénny obsah filtrovaný podľa jedného atribútu, ale v inej situácii môže byť pre neho užitočný heterogénny obsah charakterizovaný mixom rôznych vlastností. Príkladom môžu byť tweety relevantné pre „Windows Phone“, ktoré používateľ vyhľadáva po komerčnom spustení tejto platformy. Pravdepodobne ho bude zaujímať len obsah homogénny vzhľadom na autora, napríklad tweety expertov na technológie. Naopak pri hľadaní

informácií o ropnej škvrne v Mexickom zálive v lete 2010 sa ako dobré výsledky javia tweety od rôznych autorov, z rôznych geografických umiestnení a týkajúce sa viacerých tém ako politika alebo ekonomika. Motiváciou je teda využitie bohatej zásoby atribútov tweetov pri identifikácii relevantných informácií [14].

Predpokladom pre návrh metódy, ktorá berie do úvahy rozmanitosť informácií je absencia najlepšieho tweetu pre dopyt. Naopak je tomu pri vyhľadávaní na webe. Preto je pri návrhu a overení metódy dôraz kladený na to, že ako zaujímavý a informatívny bude ohodnotený obsah, ktorý pri čítaní autora viac zapojí do premýšľania a neskôr si ho bude lepšie pamätať. Definované je spektrum rozmanitosti ako štruktúra založená na entropii, kde jeho jeden koniec tvoria homogénne informácie a druhý informácie heterogénne. Každému bodu tohto spektra zodpovedá hodnota parametru diverzity. Výber entropie ako metriky pre charakterizovanie rozmanitosti je založený na predchádzajúcich prácach v oblasti sociálneho obsahu [14].

Okrem toho sú definované aj tzv. atribúty obsahu sociálneho média, ktoré tvorí množina ukazovateľov podobná tým, ktoré už boli spomínané. Patria k nim nasledujúce: či je tweet retweetom, či ide o odpoveď, výskyt URL v tweete, čas uverejnenia, geografická poloha podľa časovej zóny v profile autora, počet nasledovateľov, počet nasledovaných používateľov, počet príspevkov a asociovaná téma tweetu podľa systému, ktorý analyzuje text v tweete a aj na prípadne odkazovanej web stránke. Autori zdôrazňujú, že používajú konečný počet atribútov, ale využitie ďalších ako analýza citového zafarbenia textu, jazykového štýlu, vzťah autora a používateľa, sofistikované metriky sociálnej siete a pod. môže zvýšiť užitočnosť tejto metódy [14].

Navrhovaný je algoritmus, kedy sa pre každý z množiny tweetov vytvorí vektorová reprezentácia, ktorá nesie hodnoty pre vyššie spomínané atribúty. Potom nasleduje krok vytvorenia množiny tweetov podľa zvolenej hodnoty parametra diverzity. Na záver sa tweety z tejto množiny zoradia podľa povahy obsahu použitím entropie [14].

Pri experimentálnom vyhodnotení mali používatelia k dispozícii znenie niekoľkých tweetov. Po ich prečítaní odpovedali na 4 otázky. Mali odhadnúť čas v minútach a sekundách, ktorý podľa nich potrebovali na prečítanie tweetov. Na zvyšné tri otázky odpovedali na 7 stupňovej škále a mali sa vyjadriť k tomu, ako boli dané tweety zaujímavé, rozmanité a informatívne. Druhá časť získavajúca spätnú väzbu zisťovala, či používatelia budú vedieť z ďalšieho zoznamu tweetov určiť, ktoré tweety už čítali. Porovnávané boli 4 metódy, ktoré pozostávali zo všetkých kombinácií využitia minimalizácie entropie a priradenia váh atribútom. Navrhovaná metóda využívala oba prístupy, zvyšné tri kombinácie slúžili na porovnanie. Najmenšie chyby pri hodnotení relevancie tweetov dosahovala navrhovaná metóda, ale vyhodnotenie tiež ukazuje, že z uvedených dvoch prístupov viac pomáha minimalizácia entropie. Spätná väzba používateľov ukazuje, že dôraz na úroveň rozmanitosti spôsobil väčší záujem a informatívnosť výsledkov vyhľadávania, ak bol ich obsah veľmi heterogénny alebo naopak veľmi homogénny [14].

Štúdia teda dokazuje, že rozmanitosť informácií hrá dôležitú úlohu pri vyberaní relevantných výsledkov. Používateľov vo všeobecnosti zaujíma obsah od podobných autorov, s podobnými metadátami v príspevkoch a rovnako je pre nich užitočný aj rozmanitý obsah, ktorý podáva pohľad na rôznorodé typy mikroblogov.

2.3.3 Jazyková analýza príspevkov a jej využitie

V tejto časti analyzujem metódy, ktoré sa zaoberajú jazykovými aspektmi tweetov. Súvisiacimi problémami sú krátky rozsah mikroblogov, čo znemožňuje použitie metód sémantickej analýzy pre vyhľadávanie a kvalita jazyka tweetov, pre ktoré je charakteristická neformálna štylizácia. Postupy prekonávajúce tieto prekážky sú predpokladom pre dobré výsledky pri vyhľadávaní na sociálnej sieti Twitter.

Nielen vo vyššie uvedených prístupoch k zvyšovaniu kvality výsledkov vyhľadávania sa často spomína príslušnosť príspevkov k témam. Práca Expert-Driven Topical Classification of Short Message Streams [15] adresuje problém malého rozsahu príspevkov na sociálnych sieťach v súvislosti s ich klasifikáciou k témam. Metóda je založená na viacerých postupoch, spomeniem ale predovšetkým tie, ktoré sú relevantné k tejto diplomovej práci.

V krátkych správach ako sú statusy na Facebooku, SMS komunikácia či tweety sa fixne vyskytujú témy ako šport alebo správy, avšak obsah k týmto témam sa s časom mení. Z toho dôvodu nie je so súčasnými prístupmi možná efektívna klasifikácia príspevkov podľa ich tematiky. Pre prekonanie spomínaných prekážok je navrhnutý spôsob klasifikácie s ohľadom na časom meniaci sa charakter príspevkov príslušných k jednotlivým témam. Významnými črtami navrhovaného prístupu sú klasifikátor založený na expertnom posúdení, adaptívne trénovanie s časovým oknom a súprava metód obohacujúcich obsah príspevkov [15].

Zatiaľ čo na určenie vhodnej kategórie pre príspevok môže poslúžiť viacero algoritmov ako Naive Bayse alebo SVM (Support Vector Machines), autori sa sústredia na klasifikáciu postavenú na maximálnej entropii, ktorá preukázala svoju efektivitu v modelovaní domén s riedkym obsahom informácií. Na prekonanie uvedenej nízkej hustoty informácií v krátkych príspevkoch je možné použiť viacero spôsobov obohatenia krátkych správ [15].

K použitým technikám rozšírenia lexikálnych čŕt patria tieto [15]:

- znakové n-gramy: táto technika je založená na n po sebe idúcich znakoch,
- slovné n-gramy: zo správ sa vyberá a používa n po sebe idúcich slov,
- orthogonal sparse word bigrams: ide o techniku, ktorá sleduje páry slov, ktoré sú oddelené maximálne troma slovami.

Rozšíriť informácie v krátkych textoch je možné aj použitím externých zdrojov [15]:

- rozšírenie založené na odkazoch: príspevky na sociálnych sieťach často obsahujú URL, v ktorých sa nachádzajú informácie napovedajúce niečo o obsahu cieľovej web stránky, napríklad z pohľadu na URL:

http://www.nytimes.com/2010/07/27/sports/football/27concussion.html?ref=sports je jasné, že cieľová web stránka sa týka futbalu a hovorí o otrasoch mozgu (z angl. concussion). Pre prípady, kedy sú použité služby skracovania odkazov sa samozrejme analyzuje pôvodná, dlhá forma URL,

- rozšírenie založené na slovných druhoch: v krátkych správach môže pri pochopení príslušnosti príspevku k téme pomôcť identifikácia podstatných mien.

Obohatiť správy je možné aj uvažovaním slov, ktoré majú silné spojenie s inými slovami, čiže analyzovať správy a slová v nich na základe zoznamu často používaných slovných spojení. Pri analýze sú ale využívané predovšetkým slovné spojenia typu „kobe bryant“ (známy americký basketbalista) alebo „boston celtics“ (basketbalový tím). To znamená, že príspevok môže hovoriť o takomto koncepte, ale pomenovať ho len jeho časťou. Ak teda tweet napríklad pre obmedzenie dostupnej dĺžky obsahoval slovo „bryant“ bez uvedenia krstného mena basketbalistu, vieme na základe analýzy slovných spojení doplniť pravdepodobnú frázu, a tým aj určiť tému príspevku [15].

Z výsledkov experimentov vyplýva, že z lexikálnych prístupov k obohateniu krátkych správ je najvhodnejšie použiť unigramy. Prístup obohatenia využívajúci detekciu slovných druhov pomáha menej ako použitie unigramov. Podobne je to aj pri analýze slov v URL, ktorá nepreukazuje žiadny prínos. Tieto zistenia ale autori považujú za pozitívne práve preto, že analýza slovných druhov a extrakcia slov z URL je časovo náročná a pri analýze v reálnom čase je práve z technického hľadiska najefektívnejšie využiť unigramy [15].

Aby bolo možné využiť veľký objem dát, ku ktorým Twitter poskytuje prístup v reálnom čase, je nutné tieto veľmi zašumené dáta skvalitniť. V práci zaoberajúcej sa lexikálnou normalizáciou krátkych textových správ [16] sa autori sústreďujú na identifikáciu a normalizáciu nespisovných výrazov bez nutnosti anotácie príspevkov.

Normalizácia správ je podobná kontrole pravopisu, ale líši sa v tom, že nespisovný výraz je zámerom a nie omylom. Opisovaný spôsob sa snaží ísť ďalej ako kontrola pravopisu zámenou skratiek za plnú formu výrazov vo vhodných prípadoch ako i nahradzovaním často používaných skrátenín ako „b4“ za „before“, čo sa považuje mimo kompetencie funkcií kontroly pravopisu. Autori práce vykonali štúdiu, ktorá analyzovala slová mimo slovníka v SMS správach a tweetoch. Zistili, že pre Twitter sú charakteristické dlhé rady slov mimo slovníka, čo napovedá, že klasické učenie s učiteľom by v tomto probléme nepomohlo kvôli nízkej hustote informácií. Navyše mnoho nespisovných slov je nejednoznačných a na ich jednoznačnú identifikáciu je nevyhnutný kontext (ide napríklad o anglické slová, v prípade „Gooood“ môže ísť o slovo Good alebo God) [16].

Postup začína identifikovaním všetkých alfanumerických výrazov, z ktorých sú vybraté všetky výrazy nenachádzajúce sa v slovníku. Istá miera nedokonalosti je spôsobená prístupom, ktorý ako slová mimo slovníka označuje aj neologizmy alebo niektoré vlastné podstatné mená, ktoré sa do slovníka ešte nedostali. Riešenie spočíva aj v rozdelení slov mimo slovníka na zámerne

zmenené slová, preklepy, zadefinovaní špecifických skrátenín slov (ako „lv“ nesúce význam slova „love“) a na fonetické náhrady („2morrow“ namiesto „tommorrow“). Autori demonštrujú použitie GNU aspell slovníka na detekciu slov mimo slovníka a vyhľadávanie v slovníku internetového slangu s 5021 záznamami [16].

Tab. 2.1 zachytáva výsledky analýzy kategórií slov mimo slovníka. „Písmená“ označujú výrazy, v ktorých chýba písmeno a pôvodnú formu slova je možné jednoducho určiť (napr. „shuld“ namiesto „should“). Náhrada číslami hovorí o spomínanom nahradzovaní častí slov pre fonetickú podobnosť. Pri kategórií „písmená a slová“ ide o obe tieto situácie, čiže chýba písmeno a časť slova je nahradená číslom (už spomínané „b4“ namiesto „before“). Slangom sa označuje internetový slang ako skratka „lol“ vo význame „laugh out loud“. Kategória „iné“ obsahuje v najväčšej miere výrazy vzniknuté nevložením medzery („sucha“ namiesto „such a“). Tieto výsledky ovplyvnili návrh stratégie normalizácie, ktorá berie do úvahy predovšetkým chýbajúce písmená v slovách a náhradu časti slov kvôli fonetickej podobe s číslicami [16].

Tab. 2.1 Rozdelenie nespisovných slov podľa kategórií [16].

Kategórie	Podiel
Písmená a slová	2,36%
Písmená	72,44%
Náhrada číslami	2,76%
Slang	12,20%
Iné	10,24%

Výber kandidátov na opravu slov mimo slovníka prebieha najpoužívanejším spôsobom kontroly pravopisu, čiže ku každému výrazu sa vypočíta, koľko znakov je potrebné meniť, aby sa z nespisovného výrazu stalo slovníkové slovo. Ďalej sa v tomto kroku použije algoritmus pre dekódovanie výslovnosti slov mimo slovníka a zo slovníka sa k nim potom vyberú slová podľa postupu v predchádzajúcom kroku [16].

Podobnosť slov je analyzovaná na základe editačnej vzdialenosti slov (z angl. edit distance), na reťazcoch ako predpona a prípona a na hľadaní najdlhšieho spoločného podreťazca. Kontext sa snažia autori odvodiť použitím jazykového modelu, ale pri tweetoch tento prístup dosahuje nízku úspešnosť, preto ho autori obohacujú o vzťahy závislosti medzi slovami z databanky vzťahov. V treťom kroku dochádza k výberu kandidátov pre slová mimo slovníka [16].

Výsledky ukazujú, že vyhľadávanie v slovníku slangu dosahuje pre tweety najvyšší precision ukazovateľ spomedzi analyzovaných metód. Avšak ukazovateľ recall už nenadobúda najlepšie hodnoty. Najlepšie výsledky autori dosiahli svojou metódou, kedy sa nespisovné slová

najprv vyhľadajú v slovníku slangu a až ak sa nenájde žiaden význam sa použije kombinácia podobnosti slov a podpory kontextu [16].

Vyhľadanie slov mimo slovníka v internetovom slovníku slangu nepredstavuje problémy pre implementáciu pri optimalizovaní výsledkov vyhľadávania, ale môže spomaliť proces rekonštrukcie informačnej hodnoty tweetov. Analýza v práci [16] ale jasne ukazuje, že bez použitia slovníka slangu je možné dosiahnuť len veľmi slabé výsledky pri lexikálnej normalizácii. Problém s rýchlosťou vyhľadávania v slovníku je ale možné odstrániť použitím lokálneho slovníka slangu. V kombinácií s podobnosťou slov na základe analýzy podreťazcov je možné dosiahnuť dobré výsledky. Podpora kontextu pomáha pri výbere správnych kandidátov v najmenej možnej miere, čo z nej robí metódu neefektívne využívajúcu prostriedky pri vyhľadávaní v reálnom čase.

Napriek tomu, že vyhľadávanie na Twitteri sprevádza problém určovania relevancie krátkych správ s nekvalitnou jazykovou stránkou a môže tak vznikať nedostatok relevantných výsledkov, je dôležité nezabúdať na duplicitu a redundanciu informácií v tweetoch. Podľa článku Linguistic Redundancy in Twitter [17] nebola tomuto faktu aspoň do roku 2011 kladená takmer žiadna pozornosť.

Redundancia informácií v tweetoch čiže uverejňovanie veľmi podobných alebo rovnakých informácií používateľmi Twitteru často vzniká pri komentovaní alebo samostatnom zviditeľňovaní nejakého novinového článku alebo udalosti. Bolo dokonca objavené, že veľká časť tweetov týkajúcich sa udalostí je redundantná. Detekcia rovnakého alebo podobného obsahu v mikroblogoch je dôležitá, pretože väčšina aplikácií využívajúcich Twitter má za cieľ poskytnúť informatívny a tiež rozmanitý obsah [17].

Tab. 2.2 Výsledky štúdie redundancie tweetov [17].

redundantné	**367**	**(29,5%)**
entailment	195	(15,7%)
parafrázovanie	172	(13,5%)
neredundantné	**875**	**(70,5%)**
príbuzné	541	(43,6%)
nepríbuzné	334	(26,9%)

V literatúre je detekcia redundancie študovaná v rámci viacdokumentovej sumarizácie, kde je hlavným cieľom vybrať najinformatívnejšie vety alebo úryvky. Keďže ale tweety sú krátke, je veľmi ťažké na ne použiť tieto metódy. Vhodnejšími sú skôr detekcia parafráz a rozpoznávanie tzv. textual entailment (skratka RTE z angl. textual entailment recognition). V RTE úlohách je cieľom zistiť, či z textu (zvyčajne ho tvoria jedna až dve vety) vyplýva iný text. Väčšina navrhovaných prístupov k RTE úlohám používa strojové učenie. V opisovanej práci sú dva tweety považované za

redundantné, ak nesú rovnakú informáciu (ide o parafrázu) alebo z informácie v jednom tweete vyplýva informácia v druhom respektíve jeden tweet zahŕňa druhý. Dôkaz nutnosti riešiť redundanciu tweetov ukazujú výsledky štúdie v Tab. 2.2. Vidno totiž, že takmer 30% tweetov je redundantných. Vo väčšine takýchto tweetov sa nachádza rovnaká informácia, ktorá je vyjadrená inými slovami. Niekedy komentovaná nepodstatným osobným komentárom alebo citovým podfarbením [17].

Na detekciu redundancie sa navrhovaný prístup pozerá ako na problém klasifikácie párov tweetov. Rozhodnutie o redundancii sa teda vykonáva pre dva príspevky. Jeden z najjednoduchších prístupov je tzv. bag-of-word model (skratka BOW) a je využívaný ako základ pre hodnotenie výkonu ostatných metód. Zakladá sa na intuitívnom predpoklade, že ak väčšina použitých slov je v oboch textoch rovnaká, je vysoká pravdepodobnosť, že vyjadrujú rovnakú informáciu, a sú tak redundantné. Tento prístup je ale príliš prostý, pretože príspevky s rovnakou informáciou je jednoduché vyjadriť úplne odlišnými slovami, zatiaľ čo jednoduché pridanie záporu do identických textov z nich spraví neredundantné texty. BOW je možné rozšíriť tak, že sa podobnosť sleduje na sémantickej úrovni namiesto lexikálnej. Na tento účel je možné využiť slovnú databázu anglického jazyka WordNet. Tento prístup je teda BOW založený na WordNet databáze (skrátene WBOW). Je tak možné rozoznať, že príspevky sú podobné, aj keď jeden hovorí o Oskaroch a druhý o „Academy Awards“. Práca tiež sleduje výkon modelu lexikálneho obsahu (LEX), ktorý využíva SVM. V princípe ale nemodeluje vzťahy medzi slovami v tweete [17].

Pre prekonanie obmedzení metódy LEX je možné využiť model syntaktického obsahu (SYNT), ktorý reprezentuje dvojicu tweetov ako dvojicu fragmentov syntaktického stromu. Podobné sú potom také tweety, ktorých syntaktické stromy sú rovnaké. Naopak problémom v tomto prístupe je, že sleduje syntax a zámena slova za neekvivalentné slovo môže zmiasť použitý algoritmus. Eliminovať tento problém je možné obohatením syntaktického stromu o premenné, ktoré reprezentujú konkrétne slová, ako to robí model tzv. syntaktického pravidla prvého rádu (FOR). Pri zámene slov ale nezmenenej syntaxi už syntaktické stromy nie sú rovnaké [17].

Tab. 2.3 Výsledky experimentálneho porovnania metód odhaľovania redundancie [17].

Model	AROC
BOW	0,592
WBOW	0,578
LEX + BOW	0,725
LEX + WBOW	0,728
SYNT + BOW	0,736
SYNT + WBOW	0,737
FOR + BOW	0,739
FOR + WBOW	0,747

Výsledky experimentálneho porovnania metód sa nachádzajú v Tab. 2.3. Pre porovnanie sa používa hodnota AROC, čo označuje obsah pod ROC krivkou pre skóre klasifikácie použitím SVM. Na prvý pohľad je jasné, že využitie WordNet databázy nie je veľmi užitočné. V článku je ako dôvod uvádzaná nízka kvalita jazyka v príspevkoch. Čiže pri použití slangu, nespisovných výrazov alebo preklepoch sa tieto slová v databáze nenájdu, čiže potenciál tejto metódy ostáva nevyužitý. Pokročilé metódy dosahujú štatisticky významné zlepšenie, kde najlepšie funguje kombinácia metód FOR a WBOW [17].

Vyššie je spomenuté, že využitie WordNet slovníka nie je pre Twitter vhodné, pretože tweety obsahujú veľa výrazov mimo slovníka. Pri použití už analyzovaného prístupu pre lexikálnu normalizáciu by ale následne bolo možné zužitkovať potenciál databázy WordNet. Preto je určite vhodné preskúmať úspešnosť WBOW po nahradení slov mimo slovníka za ich slovníkové ekvivalenty. Takáto kombinácia môže predstavovať zvýšenie kvality rozpoznávania redundancie pri menšej námahe než je tomu pri aplikovaní metód strojového učenia.

Niekoľkokrát som už spomenul, že kategorizácia obsahu nepreukazuje dobré výsledky pri veľmi krátkych textoch, ktoré navyše obsahujú veľa šumu. Jeden zo spôsobov ako prekonať tieto prekážky je predstavený v článku Tweets Mining Using WIKIPEDIA and Impurity Cluster Measurement [18].

Na prekonanie krátkej dĺžky tweetov je použité ich rozšírenie o výsledky vyhľadávania na Wikipedii a pre klastrovanie je predstavený revidovaný K-NN klasifikátor používajúci meranie nečistoty klastru. Kvôli veľkej a zašumenej trénovacej množine autori najprv vykonávajú rozdelenie príspevkov do klastrov a až potom ich klasifikáciu. Takto sa im podarilo dosiahnuť efektívnejšiu a presnejšiu klasifikáciu. Tweety sú najprv delené do k klastrov použitím bežného postupu, kedy pozorovanie (čiže v tomto prípade tweet) patrí do klastru s najbližším stredom. Použitým je deliaci K-means algoritmus. Keďže príspevky v jednom klastri majú zodpovedať práve jednej téme a počet tém je na začiatku neznámy, algoritmus nekončí pri vytvorení k klastrov. Počas behu algoritmu sa sleduje rozsah klastra a po dosiahnutí zvolenej hraničnej hodnoty sa proces zastaví [18].

Kvôli rozsahu tweetov a ich nízkej jazykovej kvalite nie je možné jasne určiť, k akej téme patria. Trénovacia množina tiež neobsahuje dostatok informácií na opísanie tém, pretože obsah tweetov vždy čiastočne zodpovedá aj príbuzným témam. Preto sa pre každý tweet najprv vykoná odstránenie stop slov, slangu, neznámych slov a podobne a až potom sú extrahované entity, ktoré pravdepodobne súvisia s témou. V tomto kroku autori hovoria o téme ako trende. Názov trendu spolu s extrahovanou entitou tvorí dopyt vyhľadávania na Wikipedii. Z výsledkov sa vyberú tie najrelevantnejšie, čiže fixný počet výsledkov na začiatku zoznamu. Každý výsledok vyhľadávania obsahuje úryvok článku s asi 30 slovami. Tento úryvok je pridaný do trénovacej množiny, čím sa jej rozsah značne zväčší. Kvôli veľkému množstvu článkov na Wikipedii a ich formálnemu jazyku

je takto rozšírená trénovacia množina považovaná za vhodnú, aby ju bolo možné analyzovať podľa frekvencie výrazov [18].

Pri tvorbe klastrov sa používa taká funkcia, ktorá penalizuje nečistotu klastrov. Prístup tiež penalizuje nečisté susedné klastre pri zaraďovaní nového príspevku. Pri zaraďovaní príspevkov sa teda dáva prednosť čistejšiemu klastru [18].

V Tab. 2.4 sa nachádza porovnanie výsledkov klasifikácie. Ako základ pre hodnotenie navrhovaného algoritmu nazvaného NNEW sa použil Naive Bayes algoritmus, jeho obmena kladúca veľký dôraz na silne súvisiace kľúčové slová (v tabuľke nesie názov Dôležitosť tém) a Tag klasifikátor. Pri pohľade na F-Measure (harmonický priemer ukazovateľov Precision a Recall) je vidieť, že použitie upraveného deliaceho K-means algoritmu nedosahuje veľmi dobré výsledky, keď trénovaciu množinu tvorí len obsah tweetov. Ak sa pre trénovanie použijú len úryvky z vyhľadávania na Wikipedii, je dosiahnuté výrazné zlepšenie oproti bežným metódam. Najlepšie výsledky sú však dosiahnuté, ak trénovaciu množinu tvoria tweety spolu s úryvkami z vyhľadávania na Wikipedii [18].

Tab. 2.4 Porovnanie výsledkov klasifikácie [18].

	Naive Bayes	**Dôležitosť tém**	**Tag klasifikátor**	**NNEW Twitter**	**NNEW Wikipedia**	**NNEW kombinovaný**
Precision	0.23	0.41	0.53	0.37	0.67	0.69
Recall	0.8	0.75	0.93	0.86	0.78	0.84
F-Measure	0.36	0.53	0.68	0.52	0.72	0.76

Pre optimalizáciu vyhľadávania na Twitteri nemusí klasifikácia tweetov do tém priamo hrať dôležitú úlohu. Ak používateľ pri vyhľadávaní použije špecifický dopyt, dá sa očakávať, že tweety vo výsledkoch vyhľadávania budú patriť k jednej téme. Keďže cieľom diplomovej práce je navrhnúť metódu prieskumného vyhľadávania, je dôležité brať do úvahy aj veľmi všeobecné dopyty. V takom prípade sa môžu tweety týkať celého množstvá tém a pre optimalizáciu ponúknutých výsledkov je nevyhnutné pochopiť, čo používateľa zaujíma. Ak o používateľovi nie sú k dispozícii žiadne ďalšie informácie, témy jeho záujmu je v prípade potreby možné objaviť len z analýzy tweetov, ktoré označí ako zaujímavé. Podľa výsledkov metódy využívajúcej Wikipediu [18] sa tento prístup javí ako užitočný.

2.4 Spôsob riešenia

V tejto časti je opísaný spôsob riešenia zadania diplomovej práce a zdôvodnené vybraté riešenie. Časť riešenia bude založená na existujúcich metódach, preto sa ďalej sústredím na spôsob ich využitia a modifikácie.

2.4.1 Zhodnotenie analyzovaných prístupov

Cieľom analyzovaných prístupov v časti 2.3.1 je ohodnotiť kvalitu alebo relevanciu výsledkov vyhľadávania z dôvodu, že Twitter Search ako zabudovaný vyhľadávací mechanizmus tejto služby vracia výsledky predovšetkým podľa výskytu slov dopytu v texte tweetov. To znamená, že používateľ nemá žiadnu záruku vidieť najzaujímavejšie, najinformatívnejšie a najrelevantnejšie tweety medzi prvými výsledkami vyhľadávania. Tieto metódy sa k spomínanému problému stavajú tak, že hľadajú súbor charakteristík, ktoré používajú v hodnotiacej funkcii. Cieľom prác je opísanie funkcie, ktorá na základe najužitočnejších charakteristík dokáže čo najlepšie ohodnotiť tweety. To znamená, že ich vie ohodnotiť a zoradiť tak, ako to urobili používatelia pri experimentoch.

V tomto spôsobe optimalizácie výsledkov vyhľadávania ale vidím zásadný problém. Dalo by sa povedať, že takéto postupy sa snažia hľadať jednu odpoveď na všetky otázky. Výsledkom je jedna funkcia, ktorá výsledky vyhľadávania hodnotí na základe rovnakých charakteristík, kde každá má experimentálne stanovenú fixnú váhu. Dalo by sa teda povedať, že predpokladajú, že každý používateľ má o relevantných príspevkoch rovnakú predstavu. Dokonca pri akomkoľvek dopyte majú používatelia rovnaký cieľ, ktorý je možné dosiahnuť použitím rovnakej hodnotiacej funkcie.

Práca analyzovaná v časti 2.3.2 je postavená na úvahe, že používatelia niekedy preferujú homogénny obsah, zatiaľ čo inokedy im viac vyhovuje, ak je čo najviac rozmanitý. Rozmanitosť obsahu je založená na atribútoch tweetov a ich autorov. Práca zavádza metódu, ktorá umožňuje dynamicky meniť požadovanú hodnotu rozmanitosti výsledkov. Táto metóda tvorí základ a inšpiráciu pri voľbe prístupu k riešeniu diplomovej práce.

2.4.2 Definícia zvoleného prístupu k riešeniu

Domnievam sa, že pre dosiahnutie čo najlepších výsledkov vyhľadávania, čiže najzaujímavejších a najužitočnejších pre používateľa, je nutné brať do úvahy rôzne ciele a odlišnú motiváciu používateľov pri vyhľadávaní. Myslím si, že je preto nevyhnutné zbaviť sa predpokladu, že je možné na základe jedinej funkcie úspešne klasifikovať tweety ako zaujímavé alebo nezaujímavé bez ohľadu na konkrétneho používateľa, ktorý vyhľadáva. Ako bolo uvedené, ešte i jeden používateľ môže od vyhľadávania očakávať odlišné výsledky. Preto som sa rozhodol navrhnúť metódu, ktorá bude ohodnocovať výsledky vyhľadávania z Twitter Search API na základe parametrizovateľnej a prispôsobiteľnej funkcie (v podobnom duchu ako v článku [5] analyzovanom v časti 2.2.1). Táto adaptácia funkcie by mala prebiehať v kontexte jedného sedenia, ktoré môže pozostávať z viacerých súvisiacich dopytov.

2.4.3 Opis celkového spôsobu riešenia

Cieľom mojej diplomovej práce je navrhnúť a implementovať metódu prieskumného vyhľadávania na sociálnej sieti Twitter. Túto metódu chcem zamerať na skupinu používateľov, od ktorých sa

nevyžaduje aktívne používanie zvolenej sociálnej siete. To znamená, že nie sú k dispozícií napríklad ich tweety alebo zoznam používateľov, ktorých na Twitteri nasledujú.

To prináša obmedzenia v súvislosti s analyzovanými postupmi ohodnocovania užitočnosti, zaujímavosti alebo vhodnosti podľa záujmov. Metódy personalizácie vo všeobecnosti využívajú istú znalosť používateľa, jeho návykov alebo záujmy. Z toho vyplýva nutnosť zamerať sa skôr na spätnú väzbu od používateľa, ktorá pomáha vytvoriť model jeho záujmov a predovšetkým preferencií vo vzťahu k tweetom.

Navrhovaná metóda a systém pre prieskumné vyhľadávanie na Twitteri bude pre zadaný dopyt čerpať výsledky vyhľadávania z Twitter Search API. Použitím najčastejšie využívaných prístupov k optimalizácii vyhľadávania na sociálnej sieti sa dané tweety a prípadne ich autori budú analyzovať, čím vznikne základné zoradenie výsledkov vyhľadávania. Preddefinovaný počet týchto zoradených výsledkov bude predostretý používateľovi.

Metóda bude založená hlavne na explicitnej spätnej väzbe používateľa, napríklad označovaním zobrazených tweetov za zaujímavé alebo nezaujímavé. Po každom jednom ohodnotení sa pre používateľa bude dotvárať model jeho preferencií použitím jednej z metód strojového učenia. Na jej základe sa z ešte nezobrazených tweetov vyberú práve tie, ktoré model preferencií používateľa považuje za preferované. Samozrejme bude potrebné zahrnúť aj náhodné tweety, pretože v prípade neúplného alebo zle naučeného modelu preferencií by mohol byť používateľ ukrátený o tie tweety, ktoré by ho v skutočnosti zaujali. Model bude ako požadovaný výsledok vnímať spätnú väzbu používateľa a učiť sa bude na základe množstva charakteristík, ktoré boli prezentované v tejto časti.

Pri návrhu riešenia bude nutné definovať konkrétny súbor atribútov, ktoré budú slúžiť ako trénovacia množina pre model preferencií. Keďže esenciálnou časťou metódy je spätná väzba používateľa, je nevyhnutné navrhnúť čo najmenej rušivý spôsob explicitnej spätnej väzby. Používateľovi nemôže prekážať pri práci s aplikáciou a zároveň ho musí motivovať k poskytnutiu spätnej väzby. Návrh bude vykonaný s ohľadom na fakt, že ide o zostavovanie modelu preferencií v reálnom čase. Preto výber charakteristík, výber spôsobu ich vyhodnocovania ako i výber metódy strojového učenia pre model preferencií by mal spĺňať požiadavky na efektivitu.

3 Konceptuálny návrh riešenia

Cieľom tejto časti je navrhnúť riešenie predostretého problému. Na to je samozrejme nutné definovať požiadavky, ktoré navrhované riešenie musí spĺňať. V nadväznosti na tieto požiadavky opisujem možné riešenia opierajúc sa o analýzu problému a pokúšam sa navrhnúť, ako modifikovať a kombinovať existujúce prístupy tak, aby priniesli úžitok v kontexte zadania diplomového projektu. K týmto kandidátskym riešeniam sa snažím identifikovať a navrhnúť zatiaľ nepoužité spôsoby riešenia. Na základe predpokladu, že možných alternatív riešenia je mnoho sa snažím jednotlivé možnosti otestovať tak, aby sa dala vyhodnotiť ich vhodnosť pre riešenie zadaného problému.

Vzhľadom k povahe problematiky je možné riešenie rozdeliť na viacero častí. Cieľom systému je poskytovať používateľovi čo najzaujímavejšie príspevky vzhľadom na spätnú väzbu poskytovanú v reálnom čase. To znamená, že počas čítania tweetov a poskytovania spätnej väzby od používateľa systém dynamicky mení svoj pohľad na predostreté príspevky a snaží sa zlepšovať model naučených preferencií používateľa v kontexte jedného sedenia. Prezentácia výsledkov a získavanie implicitnej i explicitnej spätnej väzby je zodpovednosť používateľského rozhrania. Algoritmus, ktorý riadi hlavnú funkcionalitu ohodnocovania príspevkov operuje s touto spätnou väzbou a so samotnými príspevkami tak, že z tweetov extrahuje informácie a pracuje s ich metadátami. Pomocou zvolenej metódy strojového učenia vytvára model preferencií používateľa podľa poskytovanej spätnej väzby. Táto úloha je komplexná, preto sa zameriam osobitne na to, aké charakteristiky tweetov budú sledované a akú metódu zvolím pre vytváranie modelu preferencií.

V tejto časti navrhujem riešenie pre tieto časti systému osobitne, pretože predpokladám, že najlepší výsledok je možné dosiahnuť tak, že použijem najlepší spôsob získavania spätnej väzby, použijem charakteristiky príspevkov s najväčšou výpovednou hodnotou a najlepší algoritmus pre vytváranie modelu preferencií. Konkrétne kombinácie by nemali mať vplyv na výkon jednotlivých častí, pretože vhodnosť hodnotiaceho algoritmu sa nemení v závislosti od kvality zistenej spätnej väzby. Kvalita spätnej väzby, ak ju vnímame ako ukazovateľ presnosti vystihnutia názorov používateľa na príspevky, ovplyvňuje len kvalitu výsledkov, čiže to, ako zodpovedajú používateľovým preferenciám. Neovplyvňuje samotnú schopnosť zvoleného algoritmus naučiť sa na základe vstupov požadované výstupy.

Aby bolo možné vytvoriť návrhy zmysluplné a zodpovedajúce rozsahu, v nasledujúcej časti upresňujem požiadavky na systém a definujem nadväzujúce požiadavky na jednotlivé časti systému. Chcem ale zdôrazniť, že návrh jednotlivých častí bude len čiastočne oddelený, pretože napríklad požiadavka na rýchlosť ohodnocujúceho algoritmu závisí aj od množstva tweetov, ktoré bude naraz používateľské rozhranie prezentovať.

3.1 Špecifikácia požiadaviek

Všeobecný pohľad na zaujaté stanovisko k riešeniu problému je načrtnutý v časti 2.4. Na tomto mieste je stručne zhrnuté vo forme požiadaviek na riešenie.

Navrhovaný systém musí:

- ohodnocovať tweety na základe parametrizovateľnej a prispôsobiteľnej funkcie,
- ohodnocovať tweety podľa modelu preferencií aktuálneho používateľa,
- vytvárať model preferencií v kontexte jedného sedenia (ktoré môže zahŕňať jeden alebo viac súvisiacich dopytov),
- vytvárať model preferencií dynamicky a neprestať ho zlepšovať, pokiaľ používateľ pracuje so systémom,
- pracovať bez ohľadu na to, či je používateľ aktívnym používateľom Twitteru,
- brať do úvahy explicitnú ako i implicitnú spätnú väzbu od používateľa.

V prípade používateľského rozhrania sú požiadavky relatívne triviálne, čím ale dávajú priestor na rozmanité návrhy a štúdium ich vhodnosti. Sú rozdelené do povinných a nepovinných.

Používateľské rozhranie systému:

- musí umožniť používateľovi čítať obsah tweetov a pristupovať k odkazovanému obsahu (externé prepojenia),
- musí poskytovať možnosti ako sledovať aktivitu používateľa respektíve ako extrahovať spätnú väzbu používateľa,
- nemôže používateľa obmedzovať nadmerným množstvom vyžadovanej spätnej väzby,
- malo by poskytovať možnosti na vyjadrenie explicitnej spätnej väzby,
- malo by byť stavané tak, aby spôsob jeho používania podnecoval používateľa k aktivitám, ktoré sa dajú chápať ako implicitná spätná väzba,
- malo by byť zaujímavé pre používateľa alebo poskytovať mu zábavu pri jeho používaní,
- môže používateľovi prihlásenému na sieti Twitter umožniť typickú prácu s tweetmi (napr. vykonať retweet, označiť tweet za obľúbený a pod.).

Na algoritmus riadiaci hlavnú funkcionalitu ohodnocovania príspevkov a vytvárania modelu preferencií používateľa sú kladené požiadavky podľa jednotlivých jeho súčastí.

Model preferencií používateľa:

- musí byť vytváraný v kontexte jedného sedenia,
- musí byť vhodný pre použitie pri malom množstve trénovacích vzoriek ako i pri ich väčšom množstve (rádovo desiatky),

- musí si zachovávať variabilitu, nemôže predpokladať, že presnosť naučenia preferencií používateľa je možné neustále zvyšovať (optimalizácia proti preučeniu),
- musí byť schopný pracovať s množstvom rozmanitých vstupov (rôzny spôsob merania charakteristík, ich rôzna povaha a pod.),
- musí poskytovať pridruženú funkcionalitu rozširovania kolekcie dokumentov pre zvýšenie kvality poskytovaných výsledkov používateľovi (na základe aktuálneho stavu modelu by malo byť možné formulovať dopyt pre získanie príspevkov, ktoré rozšíria aktuálne načítanú, obmedzenú kolekciu príspevkov tak, aby bolo možné používateľovi poskytnúť ďalší pre neho pravdepodobne zaujímavý obsah).

Charakteristiky čiže vstupy pre model preferencií používateľa:

- musia tvoriť množinu ukazovateľov, ktoré pojednávajú o čo najrozmanitejších vlastnostiach príspevkov,
- musia hovoriť o obsahu príspevkov ako aj o jeho metadátach,
- môžu pojednávať aj o autoroch príspevkov ako aj o externom obsahu, ktorý je odkazovaný v príspevku,
- musia byť vhodné na získavanie v reálnom čase.

Metóda strojového učenia, ktorá realizuje budovanie modelu preferencií používateľa na základe spätnej väzby a charakteristík:

- musí byť svojou povahou vhodná pre vstupy odlišnej povahy a rozsahu,
- musí dosahovať primerane dobré výsledky aj pri malej trénovacej množine,
- musí obsahovať mechanizmus zabezpečujúci ochranu pred pre- a podučením,
- musí pracovať dostatočne rýchlo (množstvo času potrebného na vykonanie iterácií trénovania nemôže v konečnom dôsledku spôsobiť, že používateľovi nie sú prezentované žiadne ďalšie príspevky).

Vyššie uvedené podmienky sa vzťahujú k zvolenému návrhu používateľského rozhrania, charakteristík tweetov a spôsobu vytvárania modelu preferencií používateľa. To napríklad znamená, že navrhovaná metóda strojového učenia musí byť vhodná pre takú množinu vstupov, ktorá bola zadefinovaná návrhom sledovaných charakteristík. Predpokladám totiž, že prílišné zameranie na univerzálnosť metódy by zbytočne skomplikovalo návrh v kontexte tejto diplomovej práce.

3.2 Zvolený prístup k návrhu

Keďže systém sa skladá z viacerých častí opisovaných vyššie, ovplyvňuje kvalita týchto prvkov priamo kvalitu celého systému. Pre zvýšenie kvality modulov systému som sa rozhodol identifikovať a analyzovať možné prístupy pre každý z nich. Po sformulovaní jednotlivých alternatív a voliteľných častí (ktorých súčasné použitie sa navzájom nevylučuje) tieto návrhy

overím aspoň do takej miery, aby bolo možné rozhodnúť, ktorá alternatíva a ktoré menšie časti budú tvoriť finálny systém. Pre tieto zvolené spôsoby riešenia potom sformulujem definitívny návrh riešenia.

3.3 Prostredie a softvérové prostriedky

Práca s tweetmi prostredníctvom Twitter API je možná ako z klientskych, tak i webových aplikácií. Pre výpočtový výkon, ktorý však môže byť vyžadovaný pre spracovanie príspevkov a vytváranie modelu používateľa môže byť výhodnejšie umiestniť aplikáciu na webový server. Rovnako je výhodnejšie načítavať príspevky z Twitteru do centrálnej databázy, pretože sa tak nemusia načítavať rovnaké príspevky, pokiaľ sú relevantné pre viacerých používateľov. Kontrola duplikátov sa tým môže tiež značne zefektívniť. Preto riešenie bude implementované vo forme webovej aplikácie.

Používateľské rozhranie teda tvorí webová stránka v prehliadači. Použité jazyky a framework:

- HTML,
- CSS,
- JavaScript,
- jQuery.

Zvyšná časť riešenia je umiestnená na webovom a databázovom serveri, kde budú použité tieto jazyky resp. systémy:

- PHP,
- MySQL.

3.4 Používateľské rozhranie

Ak chceme používateľovi odporúčať pre neho zaujímavý obsah, musíme získať informáciu o tom, čo sa mu páči. To však nie je zďaleka také jednoduché, preto sa zameriavam na zisťovanie toho, ktoré z ponúknutých tweetov sa používateľovi páčia alebo ho zaujali. Z toho vyplýva dôvod vysokej dôležitosti používateľského rozhrania. Ide práve o získanie toho, čo je unikátne a najdôležitejšie pri prieskumom vyhľadávaní zameranom na dynamicky meniace sa kritéria, o získanie spätnej väzby od používateľa.

Ale aj získanie spätnej väzby viažucej sa k prezentovaným tweetom predstavuje problém. Spôsobov síce existuje niekoľko, no je potrebné vybrať taký, ktorý bude vhodný pre použitie na špecifické médium, akým Twitter nepochybne je.

Spätnú väzbu môžeme rozdeliť na implicitnú a explicitnú. Správne zachytiť používateľove preferencie je možné explicitným vyžiadaním si jeho spätnej väzby k relevantnosti výsledkov, avšak iba málo používateľov je ochotných ju poskytnúť, keďže si to od nich vyžaduje veľké množstvo času [19]. Hodnotenie a klasifikácia sú typickým príkladom takejto explicitnej spätnej

väzby, avšak pokiaľ používateľ nehodnotí úprimne, odporúčacie systémy vracajú nesprávne odporúčania [20]. K implicitnej spätnej väzbe naopak patrí správanie používateľa, ktoré je možné analyzovať bez kladenia záťaže na používateľa, no je potenciálne nejednoznačné a jeho analýza je preto komplikovaná [19]. K obťažnosti jeho analýzy prispieva aj veľké množstvo údajov spojených s implicitnou spätnou väzbou, preto aj získanie informácií o správaní používateľa zaberá omnoho viac času [20]. Implicitné charakteristiky je možné získať jednoducho a sú k dispozícii pre každý objekt, s ktorým používateľ interaguje, ide napríklad o čas pozornosti, kliknutia a pohyby myšou [19]. Pri internetových obchodoch k týmto charakteristikám okrem kliknutí spadá aj otvorenie detailu obrázkov, vloženie produktov do košíka a nákup [20].

V nasledujúcich dvoch častiach uvádzam prehľad a analýzu používaných metód pre získanie spätnej väzby. Je ale dôležité si uvedomiť, že takéto štúdie boli doteraz vykonávané predovšetkým na entitách ako filmy, produkty alebo výsledky vyhľadávania v dokumentoch. Pokiaľ viem, nebol vykonaný výskum zameraný na porovnanie metód získavania spätnej väzby pre príspevky zo sociálnych sietí. Z tohto dôvodu sa opieram o tieto štúdie existujúcich metód, ale zameriavam sa na ich pretransformovanie do kontextu prieskumného vyhľadávania v tweetoch.

3.4.1 Explicitná spätná väzba

Hodnotiace rozhrania sú široko používaným spôsobom ako zistiť názory ľudí na internete. Existuje niekoľko rôznych metód, je preto dôležité skúmať ich efektívnosť. Prístupy v takýchto hodnotiacich metódach je možné rozdeliť do dvoch skupín. V angličtine existujú výrazy ranking a rating, ktoré sa ale do slovenčiny niekedy prekladajú rovnako, napr. hodnotiť alebo klasifikovať. Ranking však označuje relatívne hodnotenie zoraďovaním položiek, zatiaľ čo rating znamená priradzovanie absolútnych hodnôt týmto položkám. Okrem mierky hodnotenia (čiže rating verzus ranking) tvorí spôsob návrhu takýchto rozhraní aj tzv. recall support čiže podpora pre používateľa s cieľom pripomenúť mu, ako hodnotil predchádzajúce položky [21].

Súčasné systémy využívajúce hodnotenie používajú n-bodovú škálu, kde n je najčastejšie väčšie ako dva a zároveň nepárne číslo, škála je zároveň umiestnená tak, aby stred reprezentoval neutrálny postoj a vyššie či nižšie pozície reprezentovali pozitívny alebo negatívny názor. Napríklad Amazon používa 5 bodový rating, cestovná príručka Michelin Guide používa trojstupňovú škálu vyjadrujúcu len intenzitu pozitívneho názoru a Facebook len jednostupňové hodnotenie prostredníctvom „Like“ tlačidla. Častou praktikou je využitie dvojstupňových rozbiehajúcich sa škál ako na Youtube prostredníctvom tlačidiel palec nahor a nadol. Majú teda spoločnú schému absolútneho ohodnotenia, kde je každá položka hodnotená samostatne čiže bez ohľadu na ostatné. To používateľovi umožňuje rýchlo a jednoducho vyjadriť názor a je to rovnako výhodné pre systém pri agregovanom zobrazovaní výsledkov [21].

Niektorí výskumníci však zastávajú názor, že ranking sa viac zhoduje s konceptom názorov a postojov, pretože sú vo svojej podstate komparatívne a súťaživé. Môžeme ale povedať, že ak je

cieľom výber z možností, ranking môže byť preferovanou cestou, avšak pri kategorizovaní prvkov množiny môže byť vhodnejší rating. Ranking je však kognitívne náročnejší a vyžaduje si koncentráciu, čo predstavuje problém pri veľkom množstve položiek. Odpoveďou na otázku, prečo je však rating výrazne rozšírenejší ako ranking môže byť aj fakt, že k ratingu sa pristúpilo hlavne pre skrátenie času potrebného na telefonické prieskumy. Lenže zjednodušiť úlohu môže mať za následok aj zníženie presnosti. Pri niektorých používateľoch nastáva tzv. ends-aversion bias čiže tendencia vyhýbať sa extrémom, čo je vyhnutie sa výberu najnižšieho alebo najvyššieho hodnotenia. Uvádzaným dôvodom môže byť neohodnotiť film najvyššou známkou, pretože potom môže prísť lepší a používateľ chce mať možnosť odlíšiť ich rôznym ohodnotením. Ranking môže byť v tomto prípade elegantným riešením tendencie vyhýbať sa extrémom [21].

Z rôznych výskumov motivácie k hodnoteniu vyplynulo, že produkty a služby ľudia na internete hodnotia vtedy, keď majú vyhranené názory, pri hodnotení filmov je motivujúce zlepšovať odporúčania, mať pocit zábavy z hodnotenia a udržiavať si zoznam videných filmov. Štúdia vykonaná za pomoci siedmich účastníkov tiež ukazuje zaujímavé výsledky. Všetci zúčastnení sa zhodli, že hodnotia, pretože pociťujú zodpovednosť za to, aby informovali ostatných o ich vlastných skúsenostiach, obzvlášť ak ide o výnimočne dobré alebo zlé skúsenosti. Päť zo siedmich sa zhodlo na tom, že chcú svojím hodnotením zlepšiť odporúčanie systému. Jeden zúčastnený ale vyjadril starosti o súkromie, kvôli ktorým je zdržanlivý z vyjadrovania svojich názorov v rôznych systémoch. Naopak štyria zo siedmich ale hodnotia práve preto, že chcú dať najavo svoj názor. Zaujímavé je aj to, že všetci účastníci výskumu chceli svojím hodnotením ovplyvniť celkové (agregované) hodnotenie jednotlivých položiek, ale zároveň pociťovali len malú možnosť ich ovplyvniť. Nakoniec len traja zo siedmich preto považujú ovplyvnenie súhrnného hodnotenia za motivujúce pri hodnotení. Ako prostriedok pre uchovávanie informácií o svojich zážitkoch pre vlastnú potrebu v budúcnosti vnímalo až päť účastníkov zo siedmich. Čo sa týka zábavy z hodnotenia, jedna štúdia hovorí o tom, že aspoň pre niektorých používateľov nie je hodnotenie prostriedkom ale cieľom. Naopak iná analýza hovorí, že pociťovaná zábava pramení z potešenia, ktoré je výsledkom dosiahnutia iných cieľov, akými sú napríklad vyjadrenie názoru alebo organizovanie vlastných kolekcií položiek [21].

V tomto článku je teda okrem spomínaných spôsobov ratingu a rankingu sledovaný aj vplyv recall support-u, čo je, ako už bolo spomenuté, použitie ďalších informácií ako napríklad predchádzajúce ohodnotené položky pri hodnotení novej položky. Obr. 3.1 zachytáva prototyp používateľského rozhrania, ktoré tvorilo súčasť výskumu v článku. Časť *a* ukazuje ranking, filmy sa hodnotia tak, že ich používateľ zoradí od najhoršieho po najlepší. Časť *b* zobrazuje binárny ranking, kedy používateľ kliknutím vyberie film, ktorý je podľa neho zo zobrazenej dvojice lepší. Časť *c* zobrazuje rating päťstupňovou škálou, používateľ hodnotí film, ktorého plagát je zobrazený vľavo a pri umiestnení kurzora myši nad príslušnú hviezdičku je vpravo zobrazený plagát filmu, ktorý predtým používateľ ohodnotil príslušným počtom hviezdičiek. Ide práve o spomínaný recall

support, ktorý by mal rating teoreticky zjednodušiť, keďže používateľ si nemusí rozpomínať na to, aké filmy hodnotil akým počtom hviezdičiek. Pri medium-fidelity návrhu autori zobrazovali namiesto jedného plagátu väčšie množstvo plagátov pod škálou [21].

Obr. 3.1 Low-fidelity prototyp pre porovnávanie spôsobov ohodnocovanie filmov [21].

Autori článku sledovali viacero ukazovateľov, ku ktorým sa zároveň vyjadrili aj účastníci výskumu. Kvantitatívna analýza názorov zúčastnených ukázala, že hviezdičkové rozhranie a hviezdičkové rozhranie s históriou boli používateľmi vnímané ako výrazne rýchlejšie než binárne a rozhranie so zoznamom. To potvrdili aj reálne výsledky získané meraním času počas používania rozhrania. Podľa používateľov bolo binárne hodnotenie ku koncu veľmi nudné. Názory na rýchlosť rozhodovania sa pri hodnotení hviezdičkami s a bez histórie sa ale líšili. Niektorým vyhovovalo to, že pri ukázaní na niektorú z hviezdičiek videli rovnako hodnotené filmy a pokiaľ mali pocit, že rovnako dobrý je aj aktuálne hodnotený film, stačilo kliknúť a nemuseli zbytočne premýšľať. Opačný názor bol zdôvodnený aj tak, že pri hodnotení hviezdičkami nejde o porovnávanie, ale jednoducho treba vyjadriť svoj názor na daný film bez súvislostí s ostatnými [21].

Hviezdičkové rozhranie s históriou zúčastnení vnímali tiež ako výrazne presnejšie než ostatné a taktiež najvhodnejšie pre organizáciu položiek pre vlastnú potrebu. Zobrazenie histórie pre hviezdičky dokonca u niektorých používateľov vyvolalo zvýšený pocit nutnosti hodnotiť presnejšie. Stále však boli takí, ktorí chcú jednoducho priradiť k filmu nejaký počet hviezdičiek podľa toho, čo im napadne ako prvé. Rozhranie s utrieďovaním zoznamu filmov dopadlo v súvislosti s presnosťou veľmi zle, čo autori pripisujú viacerým faktorom. Namiesto zoradenia niektorí používatelia jednoducho kategorizovali filmy medzi tie, ktoré sa im nepáčia, tie, ktoré sa im páčia a tie, ku ktorým majú neutrálny postoj. Rozhranie taktiež nepodporovalo automatické posúvanie sa pri ťahaní položky smerom ku koncom zoznamu, čiže bolo náročnejšie nájsť položke správne miesto. Presnosť ovplyvnilo aj požadované množstvo úsilia pre nájdenie správneho miesta pre položku. Neúspech binárneho hodnotenia je pripisovaný dvom faktorom. Používatelia sa pri rýchlejšom tempe hodnotenia občas pomýlili a vybrali nesprávnu položku a presnosť hodnotenia znižovali aj zložité rozhodnutia medzi dvoma ťažko porovnateľnými položkami [21].

Čo sa týka porovnania mentálnej záťaže, vstupujú do hry viaceré činitele. Napríklad k obom rozhraniam používajúcim ranking sa jeden z používateľov vyjadril, že majú veľkú výhodu nad ratingom, pretože nemusí svoj názor kvantifikovať a je jednoduchšie položky zoradiť. Tí, ktorí sa snažili hodnotiť konzistentne boli samozrejme spokojní so zobrazovaním histórie. Z porovnania názorov na binárne a zoznamové rozhranie je vidieť, že binárne je všeobecnejšie mentálne nenáročnejšie, pretože používateľ nemusí porovnávať jednu položku s celým zoznamom. Na druhej strane však predstavuje až príliš vysokú záťaž v situáciách, kedy sú prezentované položky z rôznych dôvodov neporovnateľné [21].

Pre organizáciu položiek sa používateľom ako vhodnejšie javili rozhrania pre rating čiže hviezdičkové hodnotenia. Kategórie sú pre nich dôležité, pretože ak majú niekomu odporučiť filmy, môžu vybrať tie, ktoré ohodnotili maximálnym počtom hviezdičiek. Na druhej strane nie je pre niektorých dôležitá presná klasifikácia, stačí vedieť, ktoré filmy sú tie lepšie. Binárne hodnotenie nie je akceptovateľné, pretože neumožňuje určiť hranicu, pokiaľ sú položky dobré a odkiaľ zlé. Väčšina ale vyhodnotila obe rozhrania pre rating ako vhodnejšie než usporadúvanie zoznamu a binárne rozhodovanie. Účastníkom štúdie vo všeobecnosti neprinášalo používanie žiadneho z rozhraní obzvlášť viac zábavy. Najlepšie sa v tejto kategórií ukázalo binárne rozhodovanie, pretože na nich pôsobilo ako „súťaž medzi filmami“ a bolo to pre nich „takmer ako hrať hru“ [21].

Iba jeden zo zúčastnených hodnotil hviezdičkové hodnotenie bez histórie ako najlepšie. Dôvodom je, že chcel hodnotiť film sám o sebe, napríklad z pohľadu toho, ako rád by si ho znova pozrel. Najvyššie známky pre hviezdičkové hodnotenie s históriou má na svedomí jeho nízka kognitívna záťaž a zvyk na tento typ hodnotenia. Pre binárne hodnotenie ako najlepšie sa rozhodli tí, ktorí preferovali jednoduchosť a zábavu. Používatelia hľadajúci jemnú granularitu a preferujúci ranking pred ratingom boli najviac spokojní s utrieďovaním zoznamu [21].

Výskum opisovaný v danom článku dokázal, že obohatenie štandardného hodnotiaceho rozhrania položkami, ktoré boli ohodnotené skôr dokáže podporiť vyjadrenie názoru bez obetovania rýchlosti. Hodnotenie hviezdičkami s históriou bolo rovnako alebo viac presné než ostatné a rovnako rýchle ako bez histórie, čím bolo stále rýchlejšie ako zvyšné rozhrania. Pri rozhraniach pre ranking existuje zásadný problém so situáciami, kedy sú podľa používateľov dve položky neporovnateľné. Analyzované rozhrania rankingu (binárne a utrieďovanie zoznamu) generovali rôzne výsledky pre tie isté položky a používatelia neboli so žiadnym z dvoch výsledných zoradení spokojní. Autori štúdie odhadujú, že to vyplýva z fundamentálneho rozdielu medzi matematickým modelom rankingu (plne zoradená výsledná množina) a ľudským mentálnym modelom (ktorý môže byť len čiastočne zotriedený). To poukazuje na ďalší možný výskum pre overenie tejto hypotézy a bádanie po tom, aké mentálne modely majú ľudia aj pre iné typy údajov než filmy. Ďalším objavom je, že zahrnutím recall support-u v čo najväčšej miere nemusí viesť k najvyššej úspešnosti hodnotiacich rozhraní. Utrieďovanie zoznamu nebolo preferovanejšie než

hodnotenie hviezdičkami s históriou alebo binárne hodnotenie, aj keď zoznam obsahoval najviac predtým hodnotených položiek. Pre vyvolávanie pocitu zábavy z hodnotenia sa binárne rozhranie ukázalo ako potenciálny kandidát pre zbieranie informácií bez nutnosti použiť ho ako primárne hodnotiace rozhranie. Či už kvalitatívne výsledky alebo slabá zhoda medzi účastníkmi pokusov zvýrazňujú individuálne preferencie v postojoch používateľov k týmto rozhraniam a poukazujú na príležitosť ďalšieho výskumu. Aj recall support sa ukázal ako neistá voľba, pretože niektorí používatelia ho využívajú, zatiaľ čo ostatní ho ignorujú. Priestor pre ďalšie bádanie je otvorený aj pre hodnotenie abstraktných konceptov a entít, ktoré sa ťažšie rozpoznávajú napríklad aj pre nemožnosť vizuálnej reprezentácie obrázkom. Rovnako je potrebné skúmať aj rozdiely medzi škálami klasifikačnými a zoraďovacími, rozchádzajúcimi sa a sekvenčnými (s nulou v strede a začínajúcich v nule) a spojitými či diskrétnymi [21].

Z tohto výskumu vzišli zaujímavé poznatky. Najdôležitejší je podľa mňa fakt, že vhodnosť konkrétneho rozhrania pre hodnotenie položiek závisí hlavne od preferencií používateľa. Tam sa ako dôležité ukazuje, či používateľ uprednostňuje rating alebo ranking a či preferuje hodnotiť položky osobitne alebo v kontexte celej množiny. Recall support sa ukazuje ako dobrá voľba, pretože pri jeho použití v správnej miere hodnotenie nespomaľuje a súčasne zvyšuje jeho presnosť. Používateľom, ktorí preferujú hodnotenie položiek osobitne, recall support neprekáža, pretože ho vedia ignorovať. Relatívne hodnotenie čiže ranking však zaostáva v presnosti respektíve poskytuje výsledky (výsledné zoradenie položiek), s ktorými sa používatelia stotožňujú najmenej. Naopak je ale menej mentálne náročné pri správnom počte porovnávaných položiek, pretože zoradiť dlhý zoznam si vyžaduje veľa sústredenia. Je tu ale riziko vyvolávania frustrácie, ak je používateľ nútený zoradiť podľa neho neporovnateľné položky. Je potrebné tiež myslieť na to, že výsledky naznačujú možné vlastnosti mentálnych modelov pre rôzne typy entít, zdá sa totiž, že pre človeka je prirodzené mať podľa preferencie položky zoradené len čiastočne a neplatí jasné porovnanie medzi všetkými inštanciami jednotlivých entít.

Vo svojej práci nemôžem všetky uvedené výsledky brať ako všeobecne platné, pretože už intuitívne je možné tušiť, že existujú rozdiely v tom, akým procesom človek hodnotí svoje preferencie voči filmom či produktom a voči tweetom. Filmy a produkty sa človeku pripomínajú pomocou názvu a obrázka a má k nim isté postoje na základe skúsenosti a zážitku. Tweety naopak predstavujú názory ľudí, kúsky informácií alebo nesú odkaz na rôzne multimédiá ako obrázky, videá, hudbu alebo články. Všetky tieto informácie sa pritom môžu vzťahovať nielen k filmom, produktom ale aj na služby, udalosti, dokonca na súkromné dianie alebo ponuky práce. Výsledky výskumu by v tejto oblasti mohli byť diametrálne odlišné od predchádzajúcej štúdie.

Spätná väzba viažuca sa k nejakej položke však samozrejme neznamená nutne len vyjadrenie toho, ako veľmi sa položka používateľovi páči. Entity, ktoré sú predmetom hodnotenia môže byť užitočné klasifikovať bez toho, aby výsledkom bolo zoradenie podľa zaujímavosti alebo

preferencií. Napríklad aj spomínané filmy môže používateľ zaradiť do kategórií podľa toho, akú náladu v ňom vyvolávajú, čím si vie lepšie spätne vybrať film alebo rýchlejšie nájsť film, ktorý chce odporučiť známemu. Pri tweetoch to platí podobne a dokonca takáto situácia je môj základný predpoklad pre túto prácu. Pre používateľa môžu byť najzaujímavejšie tie tweety, ktoré sú vtipné, ak s jeho názorom korešponduje satira k nejakej udalosti, voči ktorej uskutočňuje prieskumné vyhľadávanie. Avšak ak naopak hľadá skôr viac (serióznych a vecných) informácií k nejakej udalosti, už mu vtipné tweety vôbec vyhovovať nemusia. Prvým problémom súvisiacim s používateľským rozhraním je teda spôsob zisťovania, do akej kategórie spadajú prezentované príspevky. Kategórií, do ktorých by sa tweety dali zaradiť existuje veľmi veľa a pre používateľa by určite nebolo pohodlné hľadať v takomto zozname jednu či viac vhodných kategórií pre každý zobrazený príspevok. Omnoho elegantnejšie riešenie môže byť abstrakcia kategórií až na úroveň emócií, ktoré tweet v používateľovi vyvoláva. Na tento účel som našiel zaujímavý nástroj. Ide o tzv. AffectButton [22], ktorý opíšem v nasledujúcich odstavcoch. Pre úplnosť len dodám, že získaná informácia pomôže tweety klasifikovať, ale sama o sebe nehovorí o tom, či príspevok je alebo nie je pre používateľa zaujímavý.

AffectButton je nový digitálny nástroj pre meranie explicitnej emocionálnej spätnej väzby. Meraný emocionálny postoj označuje to, čo človek vo všeobecnosti pociťuje voči niekomu alebo niečomu. Jedným z cieľov pri vývoji AffectButton-u bolo vytvoriť metódu merania pre zadávanie zmiešanej a odstupňovanej emocionálnej spätnej väzby bez ohľadu na to, či sa vzťahuje k postoju, pocitu alebo nálade. Autori využili teóriou využívajúcu tri faktory respektíve dimenzie, ktoré tvorí potešenie, vzrušenie (v zmysle vyburcovanosti alebo nabudenosti) a dominancia (hovorí o tom, či emócia zvádza človeka k dominantnému alebo submisívnemu správaniu). V angličtine sa daný model nazýva PAD emotional model podľa slov Pleasure, Arousal a Dominance. Je však stále nejasné, či dominancia tvorí jadro emocionálneho postoja k veciam a ľuďom. PAD model je teda založený na troch faktoroch a teória hovorí, že každý objekt, emóciu, náladu a pod. je možné mapovať na trojicu hodnôt tohto modelu. Naopak to však neplatí, pretože na PAD trojice môžu byť mapované viaceré veci, nálady a emócie [22].

AffectButton bol vyvinutý ako jednoduché tlačidlo v tvare štvorca, ktorého veľkosť je možné meniť. Môže byť považovaný za statický element rozhrania, pretože sa nezväčšuje, ani neotvára ďalšie okno s možnosťami. Tlačidlo samo o sebe vykresľuje tvár, ktorá sa mení priamo podľa pozície kurzora myši v tlačidle. Emócie sú reprezentované aktuálne vykreslenou tvárou a používateľ kliknutím zadá aktuálne vyobrazenú emóciu. Výstupom tlačidla sú P, A a D hodnoty a samotný názov typických emócií pre tieto hodnoty. Definovaných je deväť prototypových výrazov tváre, čiže pre každý extrém v PAD priestore jeden výraz a jeden pre stred označujúci neutrálne pocity. Výrazy tváre pozostávajú z konfigurácie obočí, očí a úst na tvári. Na základe PAD súradníc je zobrazovaná tvár interpoláciou týchto deviatich prototypových výrazov tváre.

Tým môže používateľ zadať zmiešané emócie a tiež prototypové emócie v rôznej intenzite. Tvár je elementárna a neutrálna vzhľadom k pohlaviu [22].

Autori štúdie ukázali, že AffectButton-om je možné správne a spoľahlivo odmerať emocionálny postoj človeka k emócie vyvolávajúcim slovám. Pri meraní emócií je dôležitou vecou aj kognitívna záťaž. Výsledky práce ukázali, že čas potrebný na vybratie výrazu tváre pri prvom použití AffectButton-u nie je výrazne vyšší než pri ďalšom používaní a ani používatelia nevnímajú vynaložené úsilie ako problematické. Dôležité je tiež, že podľa výsledkov je možné tento spôsob použiť na zaznamenávanie emocionálnej povahy stimulov a generovanie užitočnej spätnej väzby používateľmi. Jeden z potenciálnych nedostatkov je použitie teórie založenej na faktoroch, pretože viaceré emócie môžu byť mapované na rovnakú PAD trojicu. To je však dôsledok zvoleného psychologického modelu, no to neprekáža, ak je tento spôsob použitý na meranie emocionálnej vlastnosti respektíve povahy objektu, citového stavu, nálady a podobne. AffectButton je tiež možné použiť na zaznamenanie dynamiky emócie, napríklad ak je používateľ požiadaný ohodnotiť nejakú položku dvakrát s odstupom času. Štúdia považuje ako jeden z ďalších dôležitých krokov vyhodnotenie tlačidla v bežných situáciách napríklad na poskytnutie emocionálnej spätnej väzby k skutočným výrobkom či obsahu alebo v kontexte sociálneho softvéru [22].

Na možnosti používateľského rozhrania pre prieskumné vyhľadávanie na Twitteri a získavanie explicitnej spätnej väzby sa pozriem ešte z jedného uhla. Najspoľahlivejšia spätná väzba je logicky tá, ktorá sa dá ľahko interpretovať čiže nie je nejednoznačná. Presnosť kategorizácie alebo hodnotenia je možné zvýšiť aj granularitou mierky hodnotenia. Lenže čím viac možností používateľ má pri akomsi hodnotení obsahu, tým je tento proces pre neho náročnejší. Ak neberiem do úvahy dokumenty odkazované v tweetoch, tak pri tak krátkych textových útvaroch naberá tento fakt prudko na svojej dôležitosti. Samotný akt zadávania hodnotenia nemôže trvať dlhšie ako prečítanie obsahu príspevku. Existuje však i možnosť, že zvyšovaním možností, z ktorých používateľ vyberá pri hodnotení nemusím dosiahnuť lepšie výsledky. Lepšie povedané, predpokladám, že existuje hranica, od ktorej sa zvyšovaním komplexnosti systému hodnotenia už hodnota spätnej väzby zvyšovať nebude. Dokonca môže klesať.

Tweety sú predsa len ťažšie analyzovateľné klasickými metódami vyhľadávania informácií než rozsiahle dokumenty s formálnou štylizáciou. A existuje i možnosť, že nie je potrebné odlišovať veľkým množstvom stupňov používateľov záujem o obsah tweetu alebo rozdeľovať tweety do veľkého počtu kategórií. Rád by som preto navrhol aj odlišný spôsob získavania spätnej väzby. Taký, ktorý podľa môjho názoru má potenciál svojou jednoduchosťou a konceptom motivovať používateľa k jeho intenzívnemu využívaniu a poskytovaniu kvalitnej spätnej väzby.

Pinterest je novšia internetová služba so silným sociálnym charakterom, ktorá je založená na vytváraní tzv. boards čiže násteniek. Používatelia môžu na nástenky pripínať len obrázky, ktoré nájdu na webe alebo nahrajú zo svojho počítača. Nástenky sú tematické a radené do kategórií ako

dizajn, umenie, cestovanie, vzdelávanie, výrobky či humor. Používatelia okrem pridávania obrázkov na existujúce nástenky môžu tiež vytvárať vlastné nástenky alebo existujúce objekty znova pripnúť na inú nástenku. Nechýbajú časté sociálne aspekty ako možnosť komentovať obrázky alebo aktivitu používateľov a objekty zdieľať alebo označiť tlačidlom Like sociálnej siete Facebook.

Uvedením výsledkov nedávnej štúdie [23] sa pokúsim odpovedať na otázku, prečo je vytváranie kolekcií obsahu také zaujímavé z pohľadu používateľov. Nie všetky kolekcie sú však zaujímavé, jedna z teórií hovorí, že kolekcie by mali byť expresívne, k čomu prispievajú tri vlastnosti. Vyberavý či eklektický cieľ pre zber a opis položiek, unikátny autorský hlas a záväzok k emocionálnemu zážitku. Týmto sa zbierka odlíši od akéhosi zoznamu obľúbených vecí a stane sa niečím hodnotným [23].

Výskum ďalej zahŕňal experiment s vytváraním kolekcií z videí, ktoré boli zamerané skôr na prezentovanie informácií. Všetci zúčastnení uviedli ako motiváciu vytvárania kolekcií osobný manažment informácií. Žiaden z nich naopak nemyslel na to, že by kolekcie chcel neskôr ukázať iným. V podstate väčšina používateľov nemala existujúcu koncepciu expresívnych možností osobných digitálnych kolekcií. Humanistická teória v analytickej časti štúdie naznačovala, že význam kolekcie je tvorený pomocou výberu, opisu a naaranžovania zdrojov. Nikto zo zúčastnených nespomenul, že pripisoval význam opisu alebo zoradeniu. Len štvrtina výsledných kolekcií zahŕňala hlbšie opisy a len jeden zúčastnený si starostlivo pripravil poradie videí v kolekcii. Naopak všetci využili celú radu sofistikovaných kritérií pri vytváraní zbierky. Mnohí súčasne zvažovali rôzne aspekty ako obsah, osobné záujmy voči predmetu videa, príťažlivosť pre prípadné publikum a vhodnosť vzhľadom k zamýšľanému cieľu kolekcie. Tab. 3.1 obsahuje frekvenciu využitia výberových kritérií pre kolekcie vytvorené v rámci experimentu [23].

Tab. 3.1 Výberové kritériá a množstvo kolekcií, v ktorých ich použili účastníci experimentu [23].

Výberové kritériá	Počet kolekcií (z 24)
Vhodnosť k zvolenej stratégii účastníka experimentu	19
Osobný záujem voči obsahu	12
Príťažlivosť pre publikum (odhadované vlastnosti pre ostatných, napr. zaujímavosť, informačný charakter, zábavnosť)	9
Dôveryhodnosť tvorcu videa, sponzora	4
Dĺžka videa	3
Uvedená popularita videa	1

Tab. 3.1 v podstate zobrazuje, čo je možné očakávať v súvislosti so spätnou väzbou od používateľa. Ak je možné očakávať podobné správanie používateľov aj pri koncepte vytvárania

kolekcií z tweetov, tak väčšina z nich by vybrala také príspevky, ktoré zapadajú do konceptu toho, čo chcú vytvoriť, čo by mohlo korešpondovať s tým, aký obsah chcú nájsť. V polovici kolekcií tiež boli prejavované osobné záujmy v súvislosti s obsahom položiek. Oba aspekty sa zdajú byť vhodné ako spätná väzba, ktorá má byť použitá na odvodenie preferencií používateľa.

3.4.2 Implicitná spätná väzba

Ako som spomenul už v časti 3.4, k implicitnej spätnej väzbe patria napríklad kliknutia na odkaz vo vyhľadávaní alebo otvorenie stránky s detailmi produktu v internetovom obchode a ďalšie akcie ako kliknutie na obrázok s cieľom otvoriť jeho väčšiu verziu, vloženie produktu do košíka alebo samotný nákup. Iné teórie hovoria o implicitnej spätnej väzby všeobecnejšie, keď pozostáva z ukazovateľov, ktoré zachytávajú správanie používateľa a je ich možné sledovať v súvislosti, s akýmkoľvek objektom, s ktorým je používateľ v interakcii. V tejto časti rozviniem práve základné ukazovatele, ktoré sú použiteľné i v kontexte prieskumného vyhľadávania na Twitteri.

Implicitnú spätnú väzbu treba chápať skôr ako informáciu o tom, čo používateľ v aplikácií robí a ako sa správa v súvislosti s obsahom. Táto informácia nie je samotná spätná väzba, ale na základe správnej analýzy je ju možné využiť ako kontext pre odvodenie spätnej väzby. Implicitnú spätnú väzbu potom chápem ako odvodenú spätnú väzbu k obsahu na základe správania používateľa.

Čas pozornosti (z angl. attention time) sa niekedy nazýva aj doba zobrazenia alebo čas čítania. Tento ukazovateľ bol pomenovaný na základe analýzy vzťahu množstva času stráveného čítaním dokumentu s množstvom záujmu používateľa. Jedna zo štúdií brala ohľad na to, že dokumenty obsahujú dva typy elementov, konkrétne text a obrázky. Čas pozornosti je teda čas, ktorý používateľ strávi čítaním textu spolu s prezeraním obrázkov. Bol tiež predstavený algoritmus založený na podobnosti obsahu dvoch dokumentov. Jeho autori predpokladali, že používateľ strávi približne rovnaký čas čítaním oboch dokumentov, ak je ich obsah veľmi podobný. Potom na základe predpovede času pozornosti nanovo zoradili výsledky vyhľadávania v snahe splniť používateľove požiadavky. Výsledky experimentov ukázali, že ohodnocujúci algoritmus s ohľadom na čas pozornosti konkrétneho používateľa môže výrazne vylepšiť výkon vyhľadávacieho systému [19].

Keď používateľ zadá vyhľadávací dopyt, vyhľadávací systém mu vráti tisícky odkazov na stránky s relevantným obsahom a používateľ klikne na tie odkazy, ktoré ho zaujali. Ukázalo sa, že kliknutie na odkaz nie vždy znamená, že cieľový dokument je relevantný, pretože používateľ má veľkú dôveru k schopnostiam vyhľadávacieho mechanizmu. Naopak význam aktu kliknutia sa prejaví, ak ho chápeme ako odhad relatívnej relevancie. V tomto zmysle boli skúmané štyri stratégie odvodené na základe sledovania kliknutí a ich porovnania s explicitnou spätnou väzbou [19]:

1. „click > skip above“ označuje, že kliknuté odkazy sú viac preferované než odkazy uvedené nad nimi (pri zoradení s klesajúcou odhadovanou relevanciou),
2. „click > earlier click“ označuje, že ak máme dva odkazy, na ktoré používateľ klikol, tak používateľ viac preferuje ten, na ktorý klikol neskôr,
3. „last click > skip above“ stratégia je podobná prvej a hovorí, že ak sú za sebou uvedené dva odkazy a používateľ klikne na druhý, určite má byť viac preferovaný než ten, ktorý je uvedený nad ním,
4. „click > no-click next“ označuje ďalšiu podobnú stratégiu, kedy kliknutý odkaz má vyššiu preferenciu než ten, ktorý je uvedený priamo za ním.

Experimentom bolo ukázané, že najvyššia dosiahnutá presnosť použitím týchto stratégií bola až takmer 90% [19].

Analýza kliknutí však nikdy nemôže dosahovať 100% úspešnosť v odvodzovaní relevancie, pretože pri vyhľadávaní v dokumentoch môže používateľ nájsť svoju odpoveď už v krátkej ukážke dokumentu pri odkaze. V tomto prípade používateľ na daný odkaz vôbec neklikne a nemusí ani kliknúť na žiaden iný. Bolo tiež objavené, že niektorí používatelia hýbu kurzorom myši podľa toho, ako čítajú text, čiže pozícia myši na stránke potom naznačuje, ktorú časť zobrazeného textu čítajú. Ďalšia odhalená súvislosť hovorí, že ak počas prehliadania výsledkov vyhľadávania nenastáva žiaden pohyb myšou, kvalita vyhľadávania sa zníži (používateľ nenachádza to, čo chce nájsť). Ukázalo sa ale aj to, že množstvo používateľov necháva kurzor myši bez pohybu na bielych miestach stránky počas toho, ako váhajú a rozhodujú sa, ktorý odkaz je pre nich ten správny [19].

Ak sledujeme, ako dlho používateľ nechá kurzor myši nad ukážkou obsahu dokumentu alebo na odkaz na tento dokument, vieme predpovedať, na ktorý odkaz nakoniec klikne. Je to obvykle ten, na ktorom strávil najviac času. Tým sa otvára možnosť poskytovať dynamický obsah podľa toho, nad ktorými odkazmi a ukážkami používateľ najviac premýšľal (čiže tie, na ktorých mal najdlhšie kurzor myši). Pokiaľ ale používateľ drží kurzor v častiach stránky bez textu, vieme pri vyhľadávaní v dokumentoch určiť len to, či si niekde na stránke všimol odpoveď na svoju otázku. Pozitívne výsledky priniesla štúdia, ktorá dokázala, že sledovanie pohybov kurzora myši prináša lepšie výsledky pri určovaní jeho preferencií než sledovanie kliknutí na odkazy. Žiaľ techniky analýzy pohybov kurzora myši sú zatiaľ najmenej preskúmané a rozvinuté [19].

Z vyššie uvedených techník vnímam čas pozornosti a analýzu pohybov kurzora myši za vhodných kandidátov pre odvodenie implicitnej spätnej väzby na tweety. Nie všetky tweety totiž obsahujú odkazy na dokumenty. Pri tých, ktoré odkazy obsahujú nemusí samotný text tweetu dobre charakterizovať obsah odkazovaného dokumentu či obrázku. Preto aj pri kliknutí na odkaz považujem čas prezerania dokumentu za smerodajný alebo aspoň smerodajnejší než samotný fakt, či nastal klik na odkaz. Okrem prezerania dokumentu sa pokúsim využiť aj čas pozornosti v rámci

samotných tweetov. Preto považujem za dôležité navrhnúť aj takú alternatívu používateľského rozhrania, ktorá vedie používateľa k poskytovaniu tejto informácie.

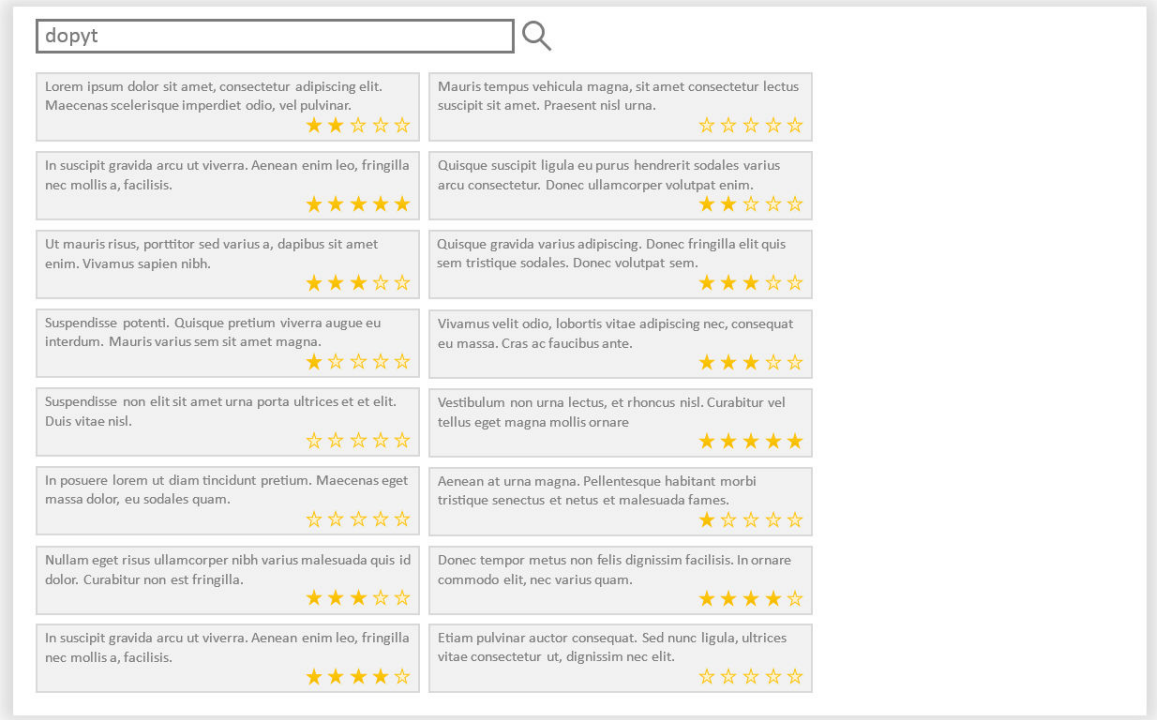
3.4.3 Navrhované alternatívy používateľského rozhrania

V tejto časti uvádzam tzv. medium-fidelity návrhy rôznych alternatív pre používateľské rozhranie. To má podporovať získavanie explicitnej a zároveň implicitnej spätnej väzby, preto tieto prístupy neskôr kombinujem. Primárnou úlohou je ale samozrejme prezentovanie príspevkov, pre ktoré navrhujem niekoľko alternatív. S výberom rozloženia príspevkov súvisí aj prezentovanie odporúčaní alebo akési vyjadrenie systému, v ktorom dáva používateľovi najavo, ako odhaduje jeho postoj k novým príspevkom.

Na Obr. 3.2 sa nachádza klasické rozhranie prezentovania tweetov, ktoré pripomína aj rozhranie Twitteru po zjednodušení. Podľa môjho názoru takéto rozhranie dostatočne nevyužíva priestor, ktorý je k dispozícii. Ďalšie príspevky sa v tomto prípade zobrazia po posunutí posuvníka. Zároveň tu žiadnym spôsobom nie sú prítomné prvky pre hodnotenie príspevkov, v rozhraní Twitteru sú ale zakompované prvky pre retweet alebo označenie tweetu ako obľúbeného.

Obr. 3.3 zobrazuje jednoduchý spôsob rozšírenia tohto konceptu. Zároveň je zahrnutý aj častý spôsob ratingu piatimi hviezdičkami. Počet stĺpcov v príklade slúži len ako ukážka, môže sa napríklad prispôsobovať veľkosti obrazovky. Za zmienku stojí aj orientácia takéhoto zoznamu, avšak rozhodol som sa nenarúšať pôvodný vertikálny koncept webstránky. Horizontálna orientácia a posúvanie zobrazeného obsahu nie je bežný, aj keď tzv. Windows 8 aplikácie prinášajú práve tento spôsob. V budúcnosti preto môže byť vhodné toto rozhodnutie prehodnotiť.

Obr. 3.2 Koncept základného rozhrania pre prezentáciu krátkych príspevkov



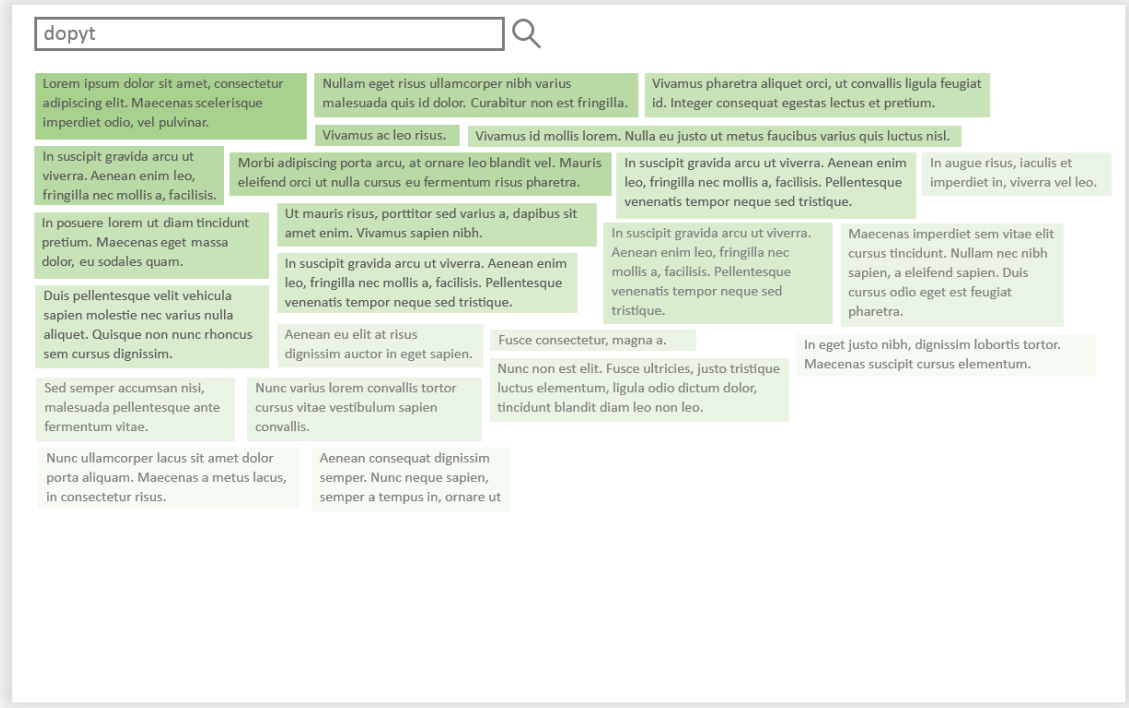
Obr. 3.3 Rozšírenie základného konceptu na viac stĺpcov (v tomto prípade dva kvôli jednoduchosti) a obohatenie o klasický spôsob pre rating hviezdičkami

V prípade konceptov vyššie je dobré mať na pamäti tzv. f vzor čítania (z angl. F-shaped reading pattern), keďže používateľ začína čítať príspevky vľavo hore (za predpokladu, že ho nezaujme niečo nezvyčajné na inom mieste rozhrania) a pokračuje po riadkoch smerom nadol, pričom si postupne menej všíma obsah napravo.

Pokiaľ hovoríme o možnostiach, ako správne rozložiť na stránke prvky, ktoré chceme dať používateľovi do pozornosti (ak vie systém určiť, ktoré prvky budú používateľa pravdepodobne najviac zaujímať), navrhujem tiež iný prístup než zobrazenie zoradeného viacstĺpcového zoznamu. Rešpektujúc spomínaný vzor čítania (alebo tiež prezerania v zmysle rýchleho prebehnutia obsahu) sa domnievam, že s klesajúcou intenzitou odporúčania je lepšie umiestniť príspevky po pomyselných štvrtinách kružníc so spoločným stredom v ľavom hornom rohu. Intenzita odporúčania príspevkov je v návrhu na Obr. 3.4 zvýraznená aj sýtosťou farebného podfarbenia príspevkov. Na tomto koncepte by som tiež rád ukázal, že nepravidelnosť rozloženia a veľkostí blokov s príspevkami šetrí miestom na obrazovke. Zostáva tak viac využiteľného miesta pre tweety než na Obr. 3.3 pri rovnakom počte položiek.

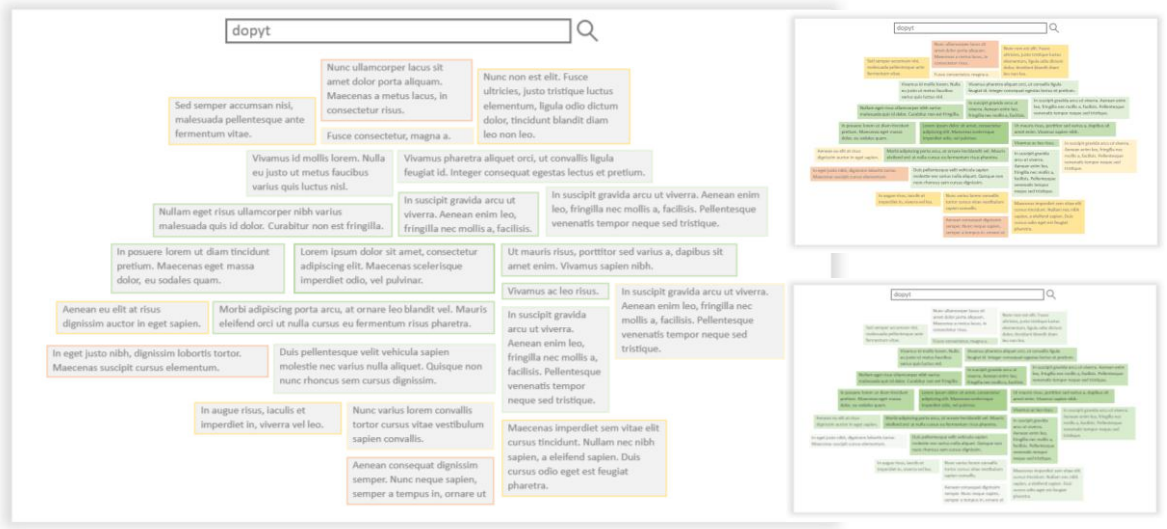
Uvedomujem si tiež, že v tomto prípade sa zobrazovanie ďalších príspevkov komplikuje, pretože posúvanie či už horizontálne alebo vertikálne by muselo mať za následok výrazne preskupenie položiek, ak chcem zachovať koncept pre F vzor. Tu do hry prichádza možnosť odstránenia posúvania zobrazenia a namiesto toho viesť používateľa k odstraňovaniu príspevkov,

aby utvoril miesto pre zobrazenie ďalších. To zároveň môže podporovať aj implicitnú spätnú väzbu.



Obr. 3.4 Rozloženie príspevkov s ohľadom na f vzor čítania spolu s farebným odlíšením prvkov podľa odporúčania systémom

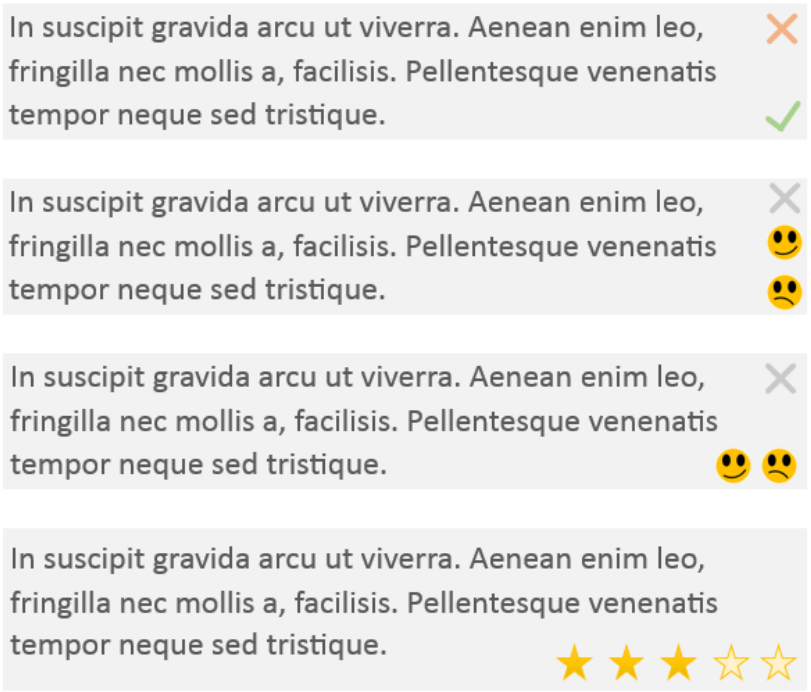
Rád by som tiež preskúmal aj menej koncepčné ponímanie zobrazovania množiny obsahových prvkov. F vzor je prezentovaný v súvislosti s klasickým rozložením textu, môže byť ale možné narušiť ho úplne iným rozložením. Symetrickejšie rozloženie položiek by mohlo byť pre používateľa intuitívnejšie a predovšetkým zaujímavejšie než to, ktoré začína v ľavom hornom rohu. Navrhujem preto aj kruhové rozloženie orientované smerom od stredu k okrajom podľa intenzity odporúčania. Obr. 3.5 tiež zachytáva alternatívy zvýraznenia odporúčania.



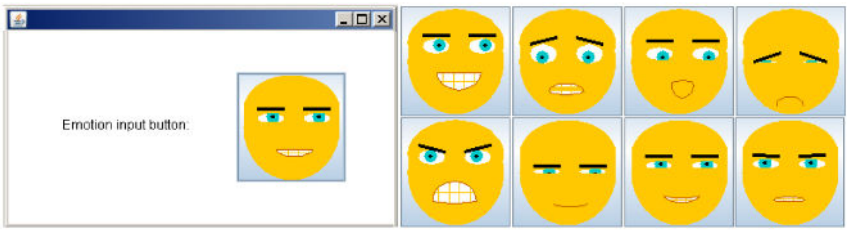
Obr. 3.5 Rozloženie orientované smerom zo stredu k okrajom so súčasným naznačením rôznych možností zvýraznenia odporúčania (napr. škála svetlosti farby verzus škála svetlosti a rôznych farieb, ale tiež možnosť podfarbiť celé bloky s príspevkami alebo len ich okraje)

V súvislosti s hodnotením tweetov navrhujem niekoľko možných spôsobov. Na Obr. 3.6 sa nachádza niekoľko ukážok rozhrania pre hodnotenie ratingom. Prvá alternatíva poskytuje len dve možné vyjadrenia. V druhej už krížik v pravom hornom rohu nie je nástrojom na vyjadrenie negatívneho postoja, ale umožňuje tweet odstrániť zo zobrazenia, pokiaľ je stanovisko používateľa skôr neutrálne. Tretia alternatíva je len malou obmenou. Na úkor mierne zložitejšieho rozloženia sa snaží byť intuitívnejšia pre vyjadrenie rozdielu medzi vyjadrením neutrálneho postoja (krížikom) a vyjadrením vyhraneného názoru (emotikonami). Štvrtá ukážka zachytáva veľmi dobre známe hodnotenie na päťstupňovej škále. Stále však ostáva predmetom skúmania z pohľadu vhodnosti na ohodnocovanie krátkych príspevkov.

Podľa AffectButton-u tiež na Obr. 3.7 uvádzam ukážku tohto tlačidla so zachytením ôsmich emócií, ktoré tvoria rohy PAD priestoru, ktoré slúžia ako ďalšia alternatíva hodnotenia, v tomto prípade však ako kategorizácia namiesto vyjadrenia miery zaujímavosti.

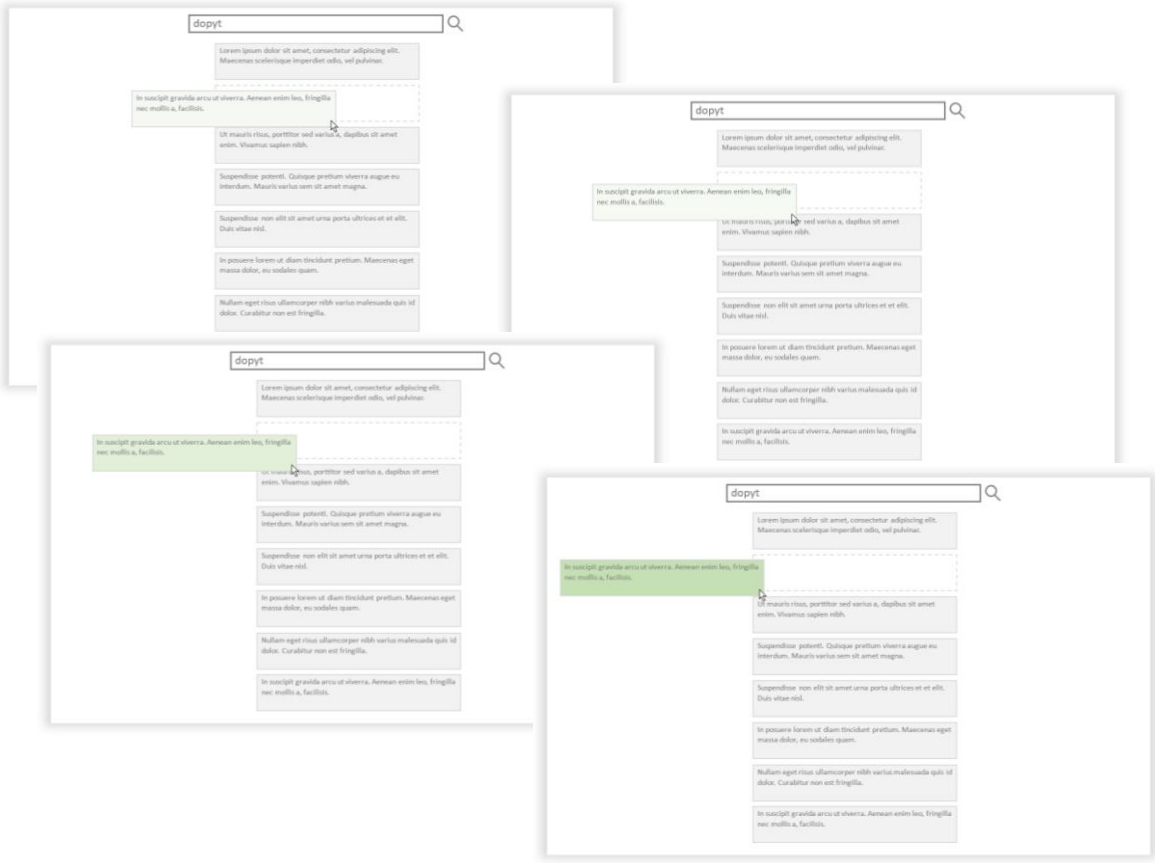


Obr. 3.6 Návrhy rozhrania pre hodnotenie tweetov

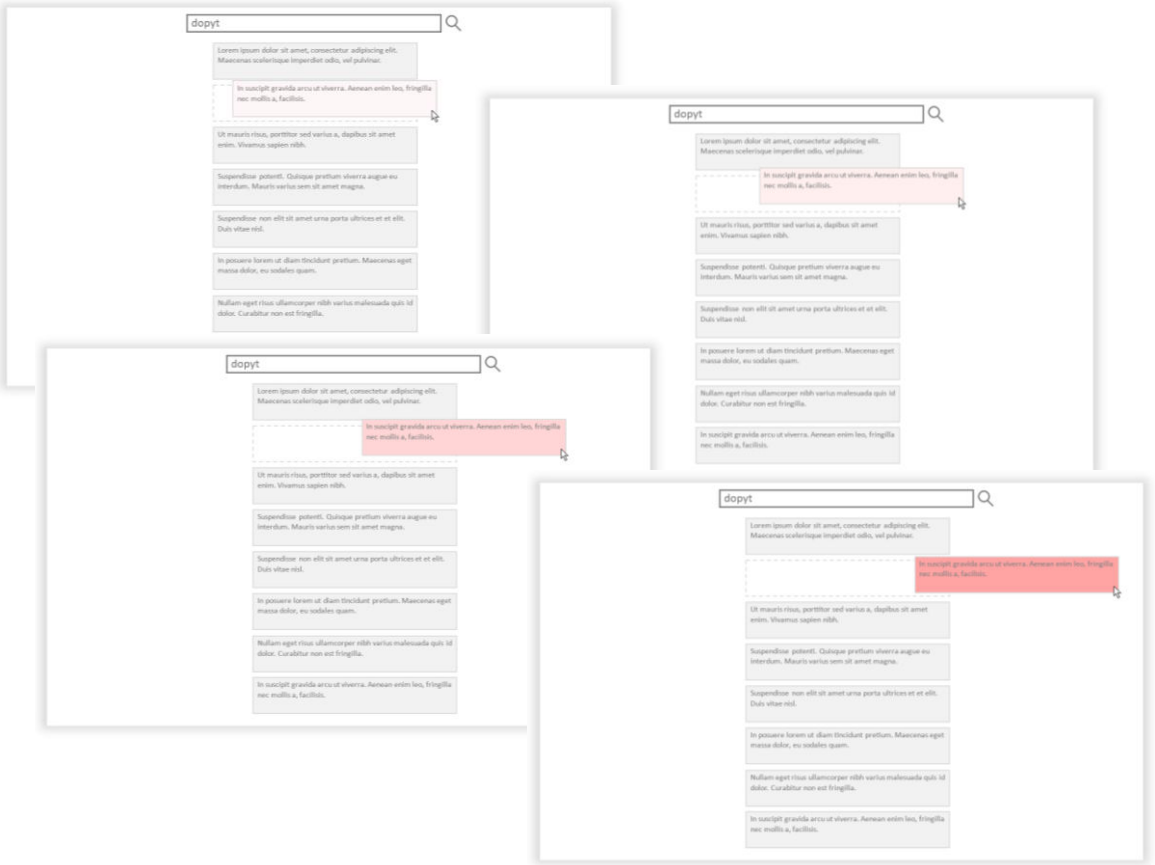


Obr. 3.7 Ukážka AffectButton-u spolu s deviatimi vyhranenými emóciami, ktoré sa nachádzajú v extrémoch PAD priestoru (zľava a po riadkoch: šťastie, strach, prekvapenie, smútok, hnev, uvoľnenosť, spokojnosť a frustrácia) [22].

Používatelia môžu mať však problém s rozhodovaním sa v rámci náročnejšieho hodnotenia (päťstupňová škála či AffectButton), preto navrhujem aj taký spôsob, ktorý nevyžaduje svoj názor kvantifikovať presne. Obr. 3.8 a Obr. 3.9 zobrazujú, ako potiahnutím tweetu smerom k stranám obrazovky mení blok s textom intenzitu farby podfarbenia. Zelená intuitívne vyjadruje pozitívny názor, červená negatívny a intenzita podfarbenia korešponduje s intenzitou pocitu. Návrhy zachytávajú použitie tohto typu hodnotenia pri jednostĺpcovom zozname, ale aplikovať by ho bolo možné aj pri zobrazení orientovanom smerom z ľavého horného rohu do strán alebo od stredu k okrajom obrazovky. Pri viacstĺpcovom zozname by sa mohla veľmi ľahko vytratiť jeho intuitívnosť a priamočiarosť.



Obr. 3.8 Vyjadrenie intenzity pozitívneho názoru na tweet jeho potiahnutím do príslušnej strany

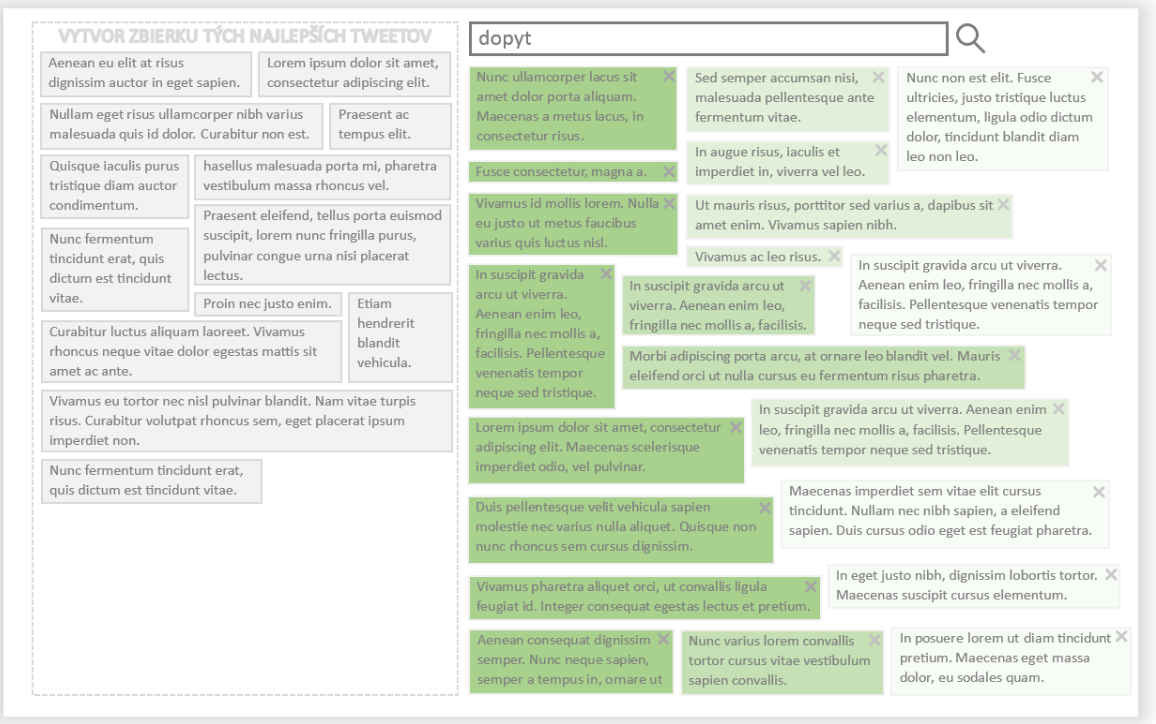


Obr. 3.9 Vyjadrenie intenzity negatívneho pocitu voči tweetu jeho potiahnutím doprava

Posledný koncept pre používateľské rozhranie určené pre získavanie explicitnej spätnej väzby sa nachádza na Obr. 3.10. Základom je vytváranie mozaiky alebo zbierky tweetov používateľom. Základný pokyn pre vytváranie mozaiky by mal byť formulovaný všeobecne, pretože používateľ by mal vyberať tweety podľa vlastných kritérií, aby poskytnutá spätná väzba jasne vyjadrovala, čo je podľa neho zaujímavé, s čím sa stotožňuje, čo ho pobavilo alebo vďaka čomu objavil nové informácie. Určite by nemal vyberať s ohľadom na to, čo by si mysleli ostatní alebo len tweety s odkazom na seriózne informačné zdroje, pokiaľ má práve za cieľ nájsť humorné príspevky.

Ukážka ďalej využíva farebné zvýraznenie výsledkov vyhľadávania podľa odporúčania systémom. Tieto tweety sú orientované zľava doprava podľa intenzity odporúčania, čo je tiež vyobrazené aj intenzitou ich podfarbenia. Výsledné rozloženie však nemusí nutne spĺňať tieto pravidlá, môže ísť aj o niektoré z vyššie uvedených rozložení, pokiaľ sa preukážu ako vhodnejšie. Používateľ myšou presúva tweety do zbierky vľavo, pokiaľ sa mu ich obsah páči. Bloky s textami tweetov obsahujú aj tlačidlo v podobe krížika v pravom hornom rohu, ktoré podobne ako pri predchádzajúcich konceptoch slúži na odstránenie položky zo zobrazeného zoznamu, čím sa uvoľní miesto pre zobrazenie ďalších tweetov.

K možným obmenám tohto konceptu patrí aj zavedenie ďalšej časti na obrazovku napr. k pravej strane, ktorá bude slúžiť ako plocha pre odstraňovanie tweetov. Čiže ak používateľ hodnotí tweet ako zaujímavý, presunie ho doľava. Ak je naopak presným opakom toho, čo ho zaujíma, presunie ho doprava. Vpravo by sa samozrejme už zoznam tweetov nezobrazoval. Takýto koncept by spríjemnil prácu v prípade dotykových displejov. Na vyjadrenie neutrálneho názoru a uvoľnenie priestoru pre zobrazenie ďalších tweetov by mohlo naďalej slúžiť tlačidlo v bloku s obsahom tweetu.



Obr. 3.10 Koncept vytvárania zbierky tweetov ako spôsob vyjadrenia spätnej väzby

V rámci podpory získavania implicitnej spätnej väzby navrhujem otvárať odkazy z tweetov priamo v rámci okna aplikácie. To znamená, že obsah, ktorý je cieľom odkazu sa zobrazí takmer cez celú plochu používateľského rozhrania. To znamená, že pokiaľ bude chcieť používateľ pokračovať v čítaní tweetov, bude musieť daný dokument zavrieť. Tým získam možnosť jednoduchšie a presnejšie merať čas pozornosti v súvislosti s dokumentami, na ktoré smerujú odkazy v tweetoch.

Čas pozornosti pre tweety je však komplikované sledovať, ak sú používateľovi naraz zobrazené viaceré položky. Ako som písal vyššie, používatelia niekedy umiestňujú kurzor myši blízko k textu alebo priamo naň, keď ho čítajú a niekedy len do prázdneho priestoru. Sledovanie takejto udalosti preto podáva skreslené a nejednoznačné výsledky. Zvažujem však dve možnosti, ktoré by pomohli dosiahnuť podobný efekt. Napríklad vyhľadávač Google pri hľadaní obrázkov zobrazuje väčší náhľad obrázku, ak naň používateľ umiestni kurzor myši. Ak by text príspevkov bol čitateľnejší práve v momente, kedy naň používateľ ukáže myšou, získal by som pravdepodobne omnoho presnejšiu informáciu o tom, čo používateľ práve číta. Na základe tohto údaju v spojení s časom pozornosti pre daný príspevok sa potom môžem pokúsiť odvodiť, ako veľmi tweet používateľa zaujal.

Druhým spôsobom je urobiť text prakticky nečitateľným, pokiaľ používateľ nedá najavo, že si ho praje prezrieť. Tým by sa presnosť určenia aktuálne čítaného tweetu ešte zvýšila, keďže nie je možné čítať dva príspevky naraz. Otázne je však, nakoľko komfortná by táto alternatíva bola pre používateľa. To je však stále možné ovplyvniť. Napríklad rozostrenie textu ostatných tweetov,

zmenšenie písma na nečitateľnú veľkosť alebo úplne skrytie textu je extrémnejšou alternatívou napríklad pre mierne rozostrenie textu, pre použitie menej kontrastnej farby písma voči pozadiu alebo len miernemu zmenšeniu textu. Experimentálne preskúmam väčšinu z týchto možností v súvislosti so zaujímavosťou a komfortom pre používateľa.

3.5 Základná charakteristika obsahu pre experimentálny systém

Pre experimentálne skúmanie v nasledujúcej časti som ako základnú charakteristiku vybral charakteristiku pre obsah tweetov. Ide o Hybrid TF-IDF [24], mieru založenú na bežne používanom ukazovateli TF-IDF. Kompenzuje však fakt, že klasická frekvencia výrazu v dokumente nie je vhodnou mierou pre krátke príspevky. Totiž aj relevantné slová sa v nich obvykle nevyskytujú viac ako raz. Tiež nerieši priamo otázku, ako veľmi je konkrétne slovo relevantné pre dokument čiže v tomto kontexte príspevok, ale ako je relevantné pre danú množinu dokumentov. To je dosiahnuté zlúčením príspevkov do jedného dokumentu pri výpočte TF komponentu, ale zachovaním oddelených príspevkov pre získanie IDF komponentu.

Týmto je možné získať zoradený zoznam slov z analyzovaných tweetov podľa ich Hybrid TF-IDF. So znižujúcou sa touto hodnotou stúpa pravdepodobnosť, že dané slovo je stop slovo. Na základe spätnej väzby je stanovené hodnotenie daného príspevku, čo môže v kombinácii s hodnotou Hybrid TF-IDF slov v ohodnotenom príspevku slúžiť ako ukazovateľ relevantnosti slov. Slová s najvyšším hodnotením používateľa a najvyššou hodnotou Hybrid TF-IDF sú potom pravdepodobne kľúčové slová pre príspevky zaujímavé pre používateľa.

$$W(w_i) = \frac{\#OccurencesOfWordInAllPosts}{\#WordsInAllPosts} * \log_2\left(\frac{\#Posts}{\#PostsInWhichWordOccurs}\right)$$

Vzorec 3.1 Váha i-teho slova pri Hybrid TF-IDF [24].

3.6 Experimentálne skúmanie používateľského rozhrania a rozširovania dopytu

V súvislosti so zvoleným prístupom k návrhu z časti 3.2 uvádzam v tejto časti postup pri skúmaní vybratých alternatív návrhu riešenia. Keďže všetky z troch častí systému (používateľské rozhranie, charakteristiky príspevkov a model preferencií používateľa) navzájom súvisia a ovplyvňujú sa, nie je možné ich objektívne overiť nezávisle na sebe. Napríklad kvalitu spätnej väzby používateľa môže v rôznej miere ovplyvňovať kvalita odporúčaní systému. Čiže ak by boli používateľovi ponúkané len náhodne zvolené príspevky a nedochádzalo by k optimalizácii hodnotenia príspevkov systémom, predpokladám, že by následne používateľ prestal poskytovať presnú explicitnú spätnú väzbu, pretože by mu chýbala motivácia toto hodnotenie ďalej vykonávať. Podobnú závislosť predpokladám aj medzi zvolenými charakteristikami či ukazovateľmi tweetov a algoritmom vytvárajúcim model preferencií používateľa. Pri použití slabých charakteristík by ani vhodne zvolený algoritmus nedokázal správne tweety ohodnocovať.

Na tomto základe som zvolil nasledujúci postup. Prvým cieľom implementácie je vytvoriť základný systém, ktorý v návrhovej časti projektu umožní implementovať alternatívy pre jednotlivé časti systému a vyhodnotiť ich. Tento základný systém má byť dostatočne znovupoužiteľný pre implementáciu finálneho návrhu.

3.6.1 Požiadavky na základný systém

V tejto časti sú uvedené požiadavky, ktoré sú spoločné pre systém určený na experimentálne skúmanie a pre finálne riešenie. Základný systém je teda spoločná časť pre tieto dve implementácie. Požiadavky teda tvoria podmnožinu požiadaviek z časti 3.1 alebo sú jej užšou špecifikáciou. Samozrejme k nim patria aj nefunkcionálne požiadavky z časti 3.3.

Základný systém musí:

- poskytovať funkcionalitu pre komunikáciu s Twitter API, čo zahŕňa potrebnú autentifikáciu a získavanie príspevkov podľa základných kritérií ako je slovný dopyt, jazyk a počet,
- uchovávať načítané tweety spolu s metadátami tak, aby ich neskôr bolo možné jednoducho spracovávať a analyzovať,
- poskytovať funkcionalitu na vyhľadávanie v tweetoch uložených v databáze systému,
- poskytovať podporu pre spravovanie sedení (rozlišovanie používateľa aj v prípade rôznych vyhľadávacích sedení),
- zahŕňať rozhranie pre moduly určené na získavanie spätnej väzby, extrakciu charakteristík z tweetov a vytváranie modelu preferencií používateľa.

3.6.2 Požiadavky na rozšírenie základného systému pre experimentálne skúmanie

Okrem predchádzajúcich požiadaviek musí tento systém:

- uchovávať spätnú väzbu používateľa vo forme numerického hodnotenia príspevkov,
- poskytovať funkcionalitu pre vyjadrenie spätnej väzby používateľa dotazníkom ku skúmaným alternatívam riešenia.

3.6.3 Test používateľského rozhrania a získavania explicitnej spätnej väzby

Pre test spôsobov získavania explicitnej spätnej väzby som vybral tri základné rozhrania z časti 3.4.3: hodnotenie hviezdičkami, hodnotenie ťahaním tweetov do strany a hodnotenie pomocou vytvárania zbierky.

Test by mal prebiehať tak, že používateľ zadá dopyt a zobrazia sa mu tweety, ktoré preň vráti Twitter API. Počet zobrazených tweetov je obmedzený (v teste na 15 tweetov) a aby zobrazené tweety používateľ skryl a zobrazili sa nové, musí prečítané príspevky ohodnotiť prvým spôsobom hodnotenia. Po tom, ako používateľ ohodnotí stanovený počet tweetov prvým spôsobom

hodnotenia (v teste zvolená hodnota 60), je nasmerovaný na dotazník. Po vyplnení dotazníka používateľ môže pokračovať v čítaní a hodnotení tweetov k ďalšiemu alebo rovnakému dopytu, ale už ďalším spôsobom hodnotenia.

Ukážka implementácie sa nachádza v Prílohe 1, ktorá obsahuje snímok webovej aplikácie počas sedenia, v ktorom používateľ hodnotí tweety k dopytu „bratislava“ a používa hodnotenie vytváraním zbierky tweetov.

Dotazníkom som sa snažil zistiť subjektívny pohľad používateľov na tieto aspekty rozhrania pre hodnotenie tweetov:

- kognitívna náročnosť,
- časová náročnosť,
- či je počet možností pri hodnotení dostatočný,
- možnosť naozaj vyjadriť svoj názor,
- postoj k nutnosti ohodnotiť tweet (aby ho bolo možné skryť),
- pocit zábavy z hodnotenia (spôsobu hodnotenia),
- či by používateľ odporučil daný spôsob hodnotenia.

K jednotlivým otázkam sa používatelia museli vyjadriť na Likertovej stupnici s piatimi stupňami a bol k dispozícii aj priestor na nepovinné vyjadrenie sa vlastnými slovami na daný spôsob hodnotenia. Ukážka dotazníka sa nachádza v Prílohe 2 vo forme snímky webovej aplikácie pri otvorení dotazníka k hodnoteniu posúvaním tweetov.

Aby mal používateľ motiváciu tweety hodnotiť, musel som implementovať aj jednoduchý spôsob vytvárania modelu jeho preferencií. V čase vykonávania testu som navrhol základnú charakteristiku pre obsah opísanú v časti 3.5, čiže podľa hodnotenia príspevkov sledujem slová s vyšším Hybrid TF-IDF a najvyšším hodnotením používateľa a po presiahnutí hranice sú prehlásené za vhodné rozšírenia dopytu. Z rovnakého dôvodu, z akého používam Hybrid TF-IDF namiesto TF-IDF ani rozšírenie dopytu nechápem v klasickom kontexte. Tieto rozšírenia používam ako samostatný dopyt a zobrazované tweety sú potom načítavané striedavo jeden tweet pre každý dopyt. Zoznam dopytov v tomto prípade teda tvorí pôvodný dopyt spolu so zoznamom rozšírení.

Tweety obsahujúce slová, ktoré boli na základe hodnotenia používateľa a Hybrid TF-IDF zvolené za relevantné zvýrazňujem pozadím zelenej farby, ktorej sila súvisí s hodnotou hodnotenia obsiahnutých slov. Keď používateľ ohodnotí tweet a existuje k nemu jeden alebo viac zobrazených duplicitných tweetov, skryje sa nielen pôvodný, ale aj jeho duplikáty. Tweety, ktoré obsahujú slová s negatívnym hodnotením zvýrazňujem červenou farbou podobným spôsob ako pri slovách relevantných pre používateľa v danom sedení.

Numerickú reprezentáciu hodnotenia tweetu ukladám do databázy spolu s identifikáciou tweetu, sedenia a dopytu, ku ktorému bol daný tweet načítaný. Tým získavam množinu dát, ktoré

použijem pre neskôr navrhnuté algoritmy vytvárania modelu preferencií používateľa. Zároveň môžu po rozdelení slúžiť ako trénovacia, testovacia a kontrolná množina.

Technická dokumentácia k návrhu tohto základného a testovacieho systému sa nachádza v časti 6.1.

Opisovaný systém je v čase písania práce dostupný na `http://dp.zilincik.eu/`.

Počas priebehu testu som zistil, že používateľom vo veľkej miere prekáža nutnosť registrovať sa a prihlasovať sa do systému, preto som nakoniec používal len identifikáciu sedenia.

Hlavným cieľom tohto testu bolo zistiť subjektívny názor na spomínané aspekty týchto troch rozhraní pre hodnotenie. V Tab. 3.2 sa nachádzajú sumarizované výsledky podľa jednotlivých aspektov a podľa rozhraní. Odpovede na otázky boli v päťstupňovej škále od „úplne nesúhlasím“ po „úplne súhlasím“, čomu prislúchali body za odpoveď od -2 po 2. Sumarizované bodové hodnotenie bolo znamienkovo upravené tak, aby pozitívna hodnota odpovede znamenala, že v danom aspekte má rozhranie výhodné vlastnosti a aby negatívne hodnotenie znamenalo, že v danom ohľade má rozhranie nedostatky. Pre hodnotenie hviezdičkami bolo vyplnených 9, pre posúvanie 5 a pre zbierku 4 dotazníky.

Tab. 3.2 Výsledky dotazníka v teste troch rozhraní pre hodnotenie tweetov.

aspekt / rozhranie	hviezdičky	posúvanie	zbierka
kognitívne nenáročné	0,44	1,20	1,50
časovo nenáročné	1,00	1,00	0,75
dostatočný počet možností hodnotenia	0,78	0,40	1,25
umožňuje vyjadriť názor	0,67	-0,60	0,00
neprekáža nutnosť ohodnotiť tweet	-0,11	0,20	1,25
používanie je zábavné	0,44	0,20	0,00
odporúčané rozhranie	0,56	-0,67	0,75
PRIEMER	0,54	0,25	0,79

Vzhľadom na celkovo malý počet vyplnených dotazníkov nemusia byť výsledky smerodajné, napriek tomu si myslím, že je dôležité brať vážne tie názory, ktoré prevládali u väčšiny používateľov. Keďže v priemere vystupujú všetky aspekty s rovnakou váhou, nemusí byť najdôležitejším výsledkom. Zaujímavým faktom podľa mňa je, že napriek tomu, že hodnotenie tweetov vytváraním zbierky nevyvolalo u používateľov prevládajúci názor na to, či umožňuje alebo neumožňuje dostatočne vyjadriť svoj názor, väčšina sa vyjadrila za to, že by dané rozhranie odporúčalo. Celkovo tiež subjektívne názory na vybrané vlastnosti boli najpozitívnejšie.

Ako som ale spomínal vyššie, operujem s predpokladom, že bez dostatočnej motivácie v rámci kvality prezentovaných tweetov bude aj kvalita spätnej väzby nízka (kvalitu spätnej väzby vnímam ako súlad medzi skutočným a zaznamenaným názorom používateľa). Preto v Tab. 3.3

uvádzam aj prehľad hodnotení tweetov pre jednotlivé rozhrania rozdelené podľa toho, či ohodnotený tweet bol výsledkom pre pôvodný dopyt alebo pre rozšírenie dopytu. Zobrazené sú len výsledky pre tie sedenia, počas ktorých bol ohodnotený priemerný alebo nadpriemerný počet tweetov. Tiež sú odstránené sedenia, v ktorých nedošlo k rozšíreniu dopytu. K tomu dochádzalo vtedy, keď bola väčšina tweetov hodnotená negatívnym hodnotením. Je vidieť, že v testovacej implementácii zatiaľ vo všeobecnosti nedochádza k úspešnému prezentovaniu lepších tweetov podľa hodnotenia. Najmenšie zhoršenie v ohodnoteniach tweetov však nastáva pri hodnotení posúvaním tweetov, tu zo štyroch dostatočne dlhých sedení došlo dvakrát k prípadu, kedy boli tweety načítané pre rozšírenie dopytu používateľom lepšie hodnotené.

Tab. 3.3 Hodnotenie tweetov v rôznych rozhraniach rozdelené podľa príslušnosti k pôvodnému alebo rozšírenému dopytu.

ROZHRANIE	HVIEZDIČKY		POSÚVANIE		ZBIERKA	
ROZŠÍRENIE	BEZ	S	BEZ	S	BEZ	S
	-0,24	-1,50	-0,21	-1,72	-0,11	-1,52
	-0,51	-1,38	-1,36	-1,00	-0,52	-1,43
	0,16	-1,29	-1,15	-1,98	-1,00	-1,15
	-0,35	1,00	-1,00	0,05		
	1,16	-0,07				
PRIEMER	0,04	-0,65	-0,93	-1,16	-0,54	-1,37
ROZDIEL	0,69		0,23		0,82	

Je zaujímavé, že subjektívne odporúčanie používateľov pre použitie týchto rozhraní je v rozpore s výsledkami, ktoré dané spôsoby poskytovania spätnej väzby priniesli. Spôsobené to však môže byť mnohými faktormi, predovšetkým malá vzorka používateľov a ešte menšie množstvo reprezentatívnych údajov. Tento test však ukázal, že rozšírenie dopytu má potenciál ako jedna z charakteristík obsahu pri hľadaní modelu preferencií používateľa. Nie je však možné na základe tohto testu vybrať najlepší spôsob získavania explicitnej spätnej väzby. Preto k tomuto kroku pristúpim až po testovaní rôznych charakteristík tweetov a algoritmov pre vytváranie modelu preferencií používateľa.

4 Opis riešenia

V tejto časti sa nachádza opis návrhu a implementácie systému podľa požiadaviek z časti 3.1. Systém je navrhnutý tak, aby spĺňal zadanie, pričom novým spôsobom stavia na niektorých vhodných existujúcich riešeniach a dopĺňa ich o nové potrebné časti pre dosiahnutie stanovených cieľov.

Na základe výsledkov testov pre získavanie explicitnej spätnej väzby z časti 3.6.3 sa pre použitie s tweetmi neukázalo žiadne z uvedených rozhraní ako výrazne lepšie než ostatné. Metóda rozširovania dopytu a načítavanie nového obsahu počas čítania príspevkov sa síce v testoch ukázala v niektorých sedeniach ako prospešná, celkovo však málo podporuje prieskumné vyhľadávanie. Ďalej je uvedený finálny opis systému pre jeho jednotlivé časti.

4.1 Návrh a opis riešenia pre jednotlivé aspekty systému

V tejto časti sú na základe analýzy zhodnotené jednotlivé aspekty pre návrh finálneho systému spolu s opisom riešenia čiastkových problémov.

4.1.1 Prieskumné vyhľadávanie

Prieskumné vyhľadávanie je vo všeobecnosti podľa [25] skúmanie informácií za účelom zistiť viac o nejakej téme. Obvykle zahŕňa iteratívny proces preskúmavania a ako sa používateľ viac dozvedá o téme (alebo súvisiacich podtémach), vyjasňuje sa aj jeho cieľ. To je opak prípadu priameho vyhľadávania faktov, kedy je cieľ dobre definovaný už na začiatku.

Existuje mnoho spôsobov, ako prieskumné vyhľadávanie podporiť z pohľadu celkovej interakcie používateľa so systémom, všeobecne známe je napríklad použitie faziet. Jedným z cieľov práce je dať používateľovi prehľad o tom, čo ľudia píšu na Twitteri v súvislosti so zadanou témou. Na to je vhodné tweety sumarizovať podľa obsahu a prezentovať takto vytvorené skupiny tweetov používateľovi. Spomínané fazety sa často používajú z dvoch dôvodov. V prvom rade používateľa informujú, aké kritériá môže použiť a aké hodnoty môžu tieto kritériá nadobúdať. S tým samozrejme súvisí aj druhá funkcionalita čiže využitie týchto faziet pre zjemnenie výsledkov vyhľadávania.

Používateľ, ktorý sa potrebuje dozvedieť, o čom sa na Twitteri píše a zistiť, ktoré podtémy ho zaujímajú, by pravdepodobne takéto fazety nevyužil dostatočne. Jedna fazeta by mohla obsahovať podtémy, ktoré sa vyskytujú vo výsledkoch vyhľadávania k zadanému kľúčovému slovu (alebo kľúčovým slovám) čiže téme. Ďalšie by mohli obsahovať spomínané metadáta, ktoré boli spomínané pri analýze či návrhu, no to by znamenalo, že si používateľ môže výsledky zoradiť alebo filtrovať podľa toho, koľko ľudí sleduje autorov príspevkov alebo či tweety obsahujú hashtag či spomenutia používateľov. To nepovažujem za dobrý nástroj pre cieľového používateľa, ktorý twitter nemusí vôbec aktívne používať. Ako bolo spomenuté vyššie, pri prieskumnom vyhľadávaní

používateľ na začiatku nemusí mať konkrétny cieľ. Aby zistil, či sa mu viac páčia tweety s referenciami na externý obsah alebo napríklad so spomenutiami používateľov, musel by vyskúšať rôzne kombinácie kritérií a pamätať si, aký obsah našiel. Táto úloha je kognitívne náročná. A nie je vôbec zaručené, že väčšina kritérií môže mať významný vplyv na zaujímavosť príspevkov pre konkrétneho používateľa v konkrétny čas a pre konkrétny dopyt.

Pre podobný prístup k prieskumnému vyhľadávaniu na Twitteri však podľa mojich zistení neexistuje takmer žiadne riešenie, s výnimkou systému TweetMotif [26], ktorý bol istú dobu aj verejne prevádzkovaný. Slúži teda ako jediný príklad metódy prieskumného vyhľadávania na Twitteri. Autori ako základ využívajú tematickú sumarizáciu. Textová sumarizácia obvykle opisuje proces, v ktorom je dokument alebo množina dokumentov transformovaná na výrazne kratšiu verziu za účelom ich stručne charakterizovať. V tejto práci ale používam výraz tematická sumarizácia pre označenie procesu extrakcie často vyskytujúcich sa fráz z množiny dokumentov. Ide teda o sumarizáciu až na frázy s rozdielom, že frázy sú využité na kategorizovanie dokumentov do tém.

Po zadaní dopytu čiže témy pozostávajúceho z jedného alebo viacerých kľúčových slov nastáva v TweetMotif [26] fáza spracovania. Z tweetov, ktoré tvoria výsledok vyhľadávania pre zadaný dopyt (čiže obvykle obsahujú zadaný dopyt) sa extrahujú unigramy, bigramy a trigramy. Nasleduje tzv. fáza syntaktického filtrovania, kedy sa niektoré n-gramy odstránia. Presný postup autori nešpecifikujú, no hovoria o pravidlách, ktoré hľadajú n-gramy obsahujúce niektoré interpunkčné znamienka na špecifikovaných miestach a tiež napríklad n-gramy, ktoré končia slovami ako „the“ alebo „of“. Takto nájdené n-gramy sa odstránia, čím sa dosahuje vyššia súdržnosť extrahovaných n-gramov. Na identifikovanie najvýznačnejších fráz pre množinu tweetov je zvolený jednoduchý prístup k modelovaniu jazyka. Frázy sú ohodnotené pravdepodobnostným pomerom, ktorý uvádza Vzorec 4.1 [26].

$$\frac{\mathbf{Pr}(\,\boldsymbol{fráza} \mid \boldsymbol{množina\ načítaných\ tweetov}\,)}{\mathbf{Pr}(\,\boldsymbol{fráza} \mid \boldsymbol{korpus}\,)}$$

Vzorec 4.1 Ohodnotenie frázy podľa jej význačnosti v množine tweetov [26].

Všeobecný korpus tweetov autori zostavili tak, že vybrali 150 000 tweetov, ktoré tvorili výsledok vyhľadávania pre všeobecné dopyty ako „the“ alebo „of“ v apríli 2009. Prirodzene, je veľké množstvo fráz, ktoré sa v tomto korpuse nenachádza, preto musia byť pravdepodobnosti zjemnené. Najlepšie výsledky boli dosiahnuté použitím Lidstonovho zjemňovania (Vzorec 4.2), kde pre frázu dĺžky *m* je *N* počet výskytov všetkých fráz s dĺžkou *m* v korpuse, *n* je počet všetkých rôznych fráz s dĺžkou *m* v korpuse a δ je zjemňovací parameter s hodnotou 0,5 [26].

$$\Pr(\,fráza \mid korpus\,) = \frac{počet\ výskytov\ frázy\ v\ korpuse + \delta}{N + \delta n}$$

Vzorec 4.2 Lidstonovo zjemňovanie pre výskyt frázy v korpuse [26].

Každá kandidátska fráza definuje tému a množinu tweetov obsahujúcich danú frázu. Veľa fráz sa však vyskytuje v približne rovnakej množine tweetov. Pre zjednotenie n-gramov s rôznou dĺžkou zaviedli autori pravidlo, že na zjednotenie kratších n-gramov obsiahnutých v dlhšom n-grame musí množina tweetov s dlhším n-gramom byť podmnožinou alebo tou istou množinou ako množina tweetov s daným kratším n-gramom. Ak teda všetky tweety pre tému „swine flu“ (prasacia chrípka) sú obsiahnuté aj v téme „flu“ (chrípka), môže sa ponechať len prvá, dlhšia téma. Okrem toho sa tiež zjednotia každé dve témy bez ohľadu na samotné frázy, ak ich množiny tweetov sú vzájomne podobné aspoň na 90% podľa Jaccardovho indexu. Po takomto zjednotení sa ponechá vždy len prienik množín tweetov. Keďže veľké množstvo tweetov je vzájomne veľmi podobných alebo priamo duplicitných, dochádza aj k odhaľovaniu takýchto tweetov. Používateľ potom vidí len jeden spomedzi veľmi podobných tweetov, no môže si zobraziť všetky podobné tweety v prípade záujmu. Napokon TweetMotif zobrazí prvých 40 tém zoradených zostupne podľa skóre uvedeného vyššie a používateľ kliknutím vyberie tému, aby si prečítal tweety s danou frázou [26].

Uvedený prístup tematickej sumarizácie považujem za veľmi vhodný pre riešenie zadania diplomového projektu. Používateľ totiž jasne vidí, aké témy k jeho zvolenému dopytu obsahujú tweety na Twitteri a vie tak rýchlo preskúmať obsah súvisiaci s jeho dopytom. Na základe takéhoto prehľadu si potom môže zvoliť, ktoré podtémy bude čítať a v prípade, že nepozná význam niektorých fráz, prečítaním niekoľkých tweetov k nej zistí, či ho takáto tematika zaujíma. Okrem tohto prehľadu, ktorý používateľ získa hneď na začiatku sa jeho cieľ začne špecifikovať práve tým, že číta tweety, ktoré ho zaujmú. Ako bolo viac krát spomenuté, výskyt fráz čiže tém ale nemusí byť jediným kritériom záujmov používateľa. Niekedy môže viac inklinovať k tweetom, prostredníctvom ktorých sú zdieľané články alebo k tweetom zdieľajúcim obrázky. Naopak inokedy pre neho môže mať vyššiu cenu samotný názor vyjadrený v tweete. Keďže množstvo obsahu je na Twitteri obrovské a jeho povaha je rozmanitá, treba klásť dôraz na to, aby používateľ nestrácal čas čítaním tweetov, ktoré ho nezaujímajú. Dôležité je ho naopak podporiť v tom, aby ďalej skúmal obsah, ktorý mu príde zaujímavý.

Preto sa riešenie diplomovej práce sústredí na návrh a implementáciu prototypu pre podporu prieskumného vyhľadávania prostredníctvom tvorby modelu preferencií používateľa, ktorého účelom je predpovedanie zaujímavosti obsahu pre používateľa výlučne v kontexte aktuálneho sedenia. Používateľovi tak potom môže byť odporúčaný ďalší obsah, ktorý sa mu pravdepodobne bude páčiť a pravdepodobne nezaujímavý obsah ostane v úzadí. Keďže spätná väzba používateľa zaťažuje, návrh sa sústredí na minimalizáciu tejto záťaže. Výsledkom je návrh systému, kedy používateľ číta obsah, hodnotí ho len v nevyhnutnej miere a systém mu indikuje odporúčané tweety. Používateľ tak namiesto čítania všetkých tém môže využiť odporúčanie a čítať predovšetkým tweety z odporúčaných tém.

Implementácia tematickej sumarizácie je postavená na čiastočne zjednodušenej metóde systému TweetMotif [26]. Opísaný prístup zahŕňa niekoľko časovo náročných úloh, preto návrh riešenia berie ohľad na toto obmedzenie a snaží sa opísať systém, ktorý ako prototyp bude použiteľný v reálnom čase. Dopyty používateľov totiž nie sú vopred známe, preto nie je možné extrakciu tém vykonávať vopred. Riešenie sa zameriava na obsah v anglickom jazyku. Po zadaní dopytu používateľom je postup nasledujúci:

1. Cez Twitter API je načítaných (maximálne) 200 tweetov v anglickom jazyku pre zadaný dopyt.
2. Pre každý tweet sa vypočíta jeho podobnosť s ostatnými tweetmi. Použitá je funkcia PHP jazyka `similar_text`. Z tweetov, ktoré sú podľa tejto funkcie podobné aspoň na 80% sa ponechá len prvý.
3. Tweety sú tokenizované na slová, z ktorých sa pre každý tweet zostavia unigramy, bigramy a trigramy. Odstránia sa inštancie n-gramu zhodného s dopytom.
4. N-gramy končiace na slová *and*, *of*, *the* a *a* sú odstránené.
5. Vypočíta sa počet výskytu všetkých n-gramov v načítaných tweetoch a s použitím všeobecného korpusu tweetov sa vypočíta aj ich skóre tak, ako zachytávajú Vzorec 4.1 a Vzorec 4.2.
6. N-gramy so skóre 10 a menej sa odstránia. Ide o veľmi všeobecné frázy ako „the“ „and the“ a podobne.
7. Odstránia sa tie n-gramy, pre ktoré existuje (n+1)-gram, ktorého množina tweetov je podmnožinou alebo tou istou množinou ako množina tweetov pre n-gram.
8. Zjednotia sa množiny tweetov pre tie n-gramy, ktorých množiny tweetov sú podobné na 90% a viac podľa Jaccardovho indexu.
9. N-gramy sa zoradia zostupne podľa skóre vypočítaného v kroku 5.

Všeobecný korpus tweetov pozostáva z 199 318 tweetov, ktoré boli načítané vo februári 2013 pre dopyty „the“ a „of“.

N-gramy sú zobrazené používateľovi ako zoznam fráz zoradený podľa vypočítaného skóre. Jedna strana zoznamu obsahuje vždy 20 tweetov. Používateľ kliknutím vyberie frázu a zobrazí sa mu množina tweetov, ktoré túto frázu obsahujú.

4.1.2 Kompenzácia obmedzeného informačného obsahu tweetov

V časti 2.3.3 je uvedený prístup z [15], ktorý rôznymi spôsobmi zvyšuje informačný zisk z textového obsahu tweetov. Metódy sú však vhodné pre aplikovanie na korpus, kedy vyššia výpočtová náročnosť nepredstavuje problém. Pre praktické využitie je vhodný prístup obohacovania tweetov o slová extrahované z obsiahnutého URL. Prináša to však prekážky; keďže tweety sú obmedzené na 140 znakov, URL sú aktuálne skracované vo všetkých tweetoch. To znamená, že v tweete sa bude nachádzať `http://nyti.ms/c5SzO0` namiesto

`http://www.nytimes.com/2010/07/27/sports/football/27concussion.html?ref=sports`. Preto prístup z [15] nie je možné priamo použiť, pretože by boli extrahované buď bezvýznamné tokeny, alebo vôbec žiadne slová. Keďže bežný spôsob fungovania skracovačov URL funguje na princípe presmerovania prostredníctvom údajov v http hlavičke, je na získanie pôvodnej formy URL potrebné načítať túto hlavičku prostredníctvom http požiadavky. Extrakcia slov z pôvodného URL v príklade môže text tweetu obohatiť o slová *sports*, *footbal* a *concussion*. Zasielanie týchto požiadaviek však už predstavuje ďalšiu réžiu.

Toto URL `http://mahotellaqueens.com/deals/230922493184.html` však obsahuje jediné slovo *deals*. Názov stránky, ktorý je uvedený v tágu `title` jazyka HTML však znie *„FOLK COSTUME BLOUSE Europe Slovakia Ethnic Gypsy Retro Hand Embroidered Kroj $140.00“*. Tento opis obsahuje výrazne viac plnovýznamových slov, a teda aj viac informácií než jediné slovo z pôvodnej URL. Pre predchádzajúci príklad obsahuje `title` tág „*N.F.L. Asserts Greater Risks of Head Injury*“. Ak chceme obohatiť tweety o slová z referencií, je potrebné vykonávať http požiadavky, ktoré pri veľkom množstve už vytvárajú badateľné časové oneskorenie. Aby však táto technika mala vyšší prínos, považujem za vhodné načítať aj samotnú webovú stránku a extrahovať obsah tágu `title`. Autori [15] síce nakoniec neodporúčajú používať extrakciu slov z URL, no ich cieľom bolo zaradiť tweety do širokých tém ako správy alebo šport. Na tento účel bolo rýchlejšie využiť unigramy priamo z textu a podľa ďalšieho korpusu vzťahov medzi slovami a témami určiť toto zaradenie. V tejto práci ale obohatenie slúži pre rozšírenie výstupov extrakcie fráz v tematickej sumarizácií.

4.1.3 Získavanie spätnej väzby

Pre vytvorenie odporúčaní a modelu preferencií používateľa je nutné zistiť, čo ho zaujíma a čo nie. Keďže táto práca sa sústredí na modelovanie preferencií používateľa v kontexte jedného sedenia, nepripadá do úvahy napríklad zisťovanie záujmov používateľa prostredníctvom Twitter účtu, histórie vyhľadávania na webe či záložiek v prehliadači. To ale neprináša znevýhodnenie pre témy, ktoré by používateľa zaujali, ale nikdy ich nevyhľadával, čiže všetky témy v načítaných tweetoch majú relatívne rovnakú pravdepodobnosť byť označené za zaujímavé pre používateľa v súvislosti s aktuálnym dopytom.

Je teda potrebné získať spätnú väzbu používateľa počas jeho čítania obsahu v systéme. V častiach 3.4.1 a 3.4.2 boli spracované možnosti na získavanie spätnej väzby. Keďže explicitná spätná väzba zaťažuje používateľa, je dôležité nevynucovať ju v príliš veľkej miere. Implicitná spätná väzba sa veľmi ťažko interpretuje a podľa mojich doterajších zistení neexistuje štúdia zameraná na krátke útvary, akými sú tweety. Výskum v tejto oblasti sa zaoberá výsledkami vyhľadávania na webe respektíve vyhľadávaním v dokumentoch spravidla rádovo dlhších než sú mikroblogy.

Rozhodol som sa zvoliť mierne odlišný prístup k získavaniu spätnej väzby (SV). Používateľ by mal mať vždy možnosť ohodnotiť zaujímavosť tweetu. Nie je k tomu nútený a využiť možnosť poskytnutia spätnej väzby je na jeho uvážení. Aby bolo nižšie množstvo explicitnej SV kompenzované, používam implicitnú SV ako prostriedok pre odhad explicitnej SV. Aby bolo možné implicitnú SV interpretovať, rozhodol som sa použiť strojové učenie pre vytvorenie modelu používateľa z pohľadu jeho správania sa v súvislosti s čítaním tweetov. To znamená, že počas čítania tweetov zaznamenávam tieto veličiny:

- čas pozornosti pre text tweetu,
- akciu kliknutia na URL (ak sa nachádza v tweete),
- čas pozornosti pre odkazovaný obsah (externá web stránka odkazovaná cez URL).

Aby bolo možné sledovať čas pozornosti pre text tweetu, je používateľské rozhranie navrhnuté tak, aby bol text tweetu dobre čitateľný len vtedy, keď naň používateľ umiestni kurzor myši. Je to dosiahnuté zvolením veľmi málo kontrastných farieb pre text oproti pozadiu pre tweety, nad ktorými sa kurzor myši nenachádza.

Tweety, ktoré používateľ ohodnotí (poskytne explicitnú SV) slúžia ako trénovacia množina. Pre zjednodušenie problému sú tweety rozdeľované do dvoch tried: zaujímavé a nezaujímavé. Potom tweety, ktoré používateľ prečíta ale neohodnotí (čiže k tweetom je zaznamenaná iba implicitná SV) sú prostredníctvom vytvoreného modelu klasifikované do uvedených dvoch tried. Takto je k dispozícii viac informácií o preferenciách používateľa než v prípade, že by sa použili len tie tweety, ktoré používateľ explicitne ohodnotil.

Spomedzi algoritmov strojového učenia vhodných pre klasifikáciu do dvoch tried som pre vytváranie modelu interpretujúceho implicitnú spätnú väzbu zvolil SVM (Support Vector Machines). Trénovaciu množinu teda tvorí množina atribútov, kde prvý označuje zaradenie do jednej z dvoch tried (zaujímavé / nezaujímavé) a ďalšie tri vyjadrujú implicitnú SV. Kliknutie na URL v tweete tvorí binárny atribút. Keďže časy pozornosti sú merané v milisekundách, sú normalizované tzv. min-max metódou, kedy najvyššia hodnota 1 patrí najdlhšiemu času pozornosti v rámci daného atribútu a hodnota 0 najkratšiemu času.

Keďže testy z časti 3.6.3 nepriniesli výrazne rôzne výsledky pre skúmané rozhrania pre hodnotenie tweetov, zvolil som na získavanie explicitnej spätnej väzby päť stupňovú mierku s hviezdičkami, čiže jeden z najčastejšie používaných spôsobov pre absolútne hodnotenie (rating). Aby bolo dosiahnuté jej správne použitie, používateľ je pri otvorení aplikácie informovaný o spôsobe interpretácie hodnotenia hviezdičkami. Jedna alebo dve znamenajú nezaujímavý alebo výrazne nezaujímavý tweet, pri štyroch a piatich hviezdičkách je situácia analogická pre pozitívny vzťah k tweetu. Tri hviezdičky nesú význam neutrálneho postoju.

Keďže je však potrebné klasifikovať len do dvoch tried, jedna a dve hviezdičky predstavujú ekvivalentné ohodnotenie rovnako ako štyri a päť hviezdičiek. Štúdia analyzovaná v časti 3.4.1 hovorí o rôznych názoroch na niektoré rozhrania hodnotenia, no sústredí sa na názory

a merania pre rôzne spôsoby hodnotenia v širšom zmysle, preto obsahuje málo informácií o detailnom výbere počtu stupňov či ich označenia (hviezdičky, palec hore a dole a pod.). Pre niektorých ľudí v danej štúdii bolo veľmi dôležité, či položky ohodnotia napríklad štyrmi alebo piatimi hviezdičkami, pre iných naopak nebolo významné porovnávať, aké hodnotenie dali ostatným položkám, keď sa rozhodovali pre správny stupeň v mierke hodnotenia. Preto považujem za lepšie poskytnúť radšej mierku s piatimi stupňami, aby používatelia, ktorí preferujú jemnejšiu granularitu hodnotenia necítili zbytočné obmedzenie.

Interpretovať implicitnú spätnú väzbu na základe uvedených atribútov nie je jednoduchá úloha. Predovšetkým ide o veľmi malú množinu atribútov a pre tweety bez URL sa mení len čas pozornosti pre text tweetu. Taktiež je samozrejme prítomný šum, kedy sú sledované atribúty ovplyvnené okolnosťami, ktoré nesúvisia so zaujímavosťou obsahu pre používateľa. Napríklad narušenie pozornosti používateľa externými okolnosťami, kedy sa počas čítania tweetu alebo prezerania web stránky začne na chvíľu venovať niečomu inému.

Keďže je nutné rátať aj s možnosťou, že sa na základe daných atribútov nepodarí vytvoriť model, ktorý bude dosahovať relevantnú úspešnosť, je potrebné implementovať aj nejakú z foriem jeho validácie. Keďže ale explicitnej spätnej väzby je vo všeobecnosti málo, pravdepodobne nie je vhodným riešením ochudobnenie trénovacej množiny o inštancie, ktoré sa použijú výlučne ako testovacia množina pre zistenie úspešnosti modelu. Na spoločnom princípe však funguje krížová validácia (z angl. cross validation), kedy sa vyberie k inštancií z trénovacej množiny, použije sa algoritmus strojového učenia pre vytvorenie modelu a na daných k inštanciách je vypočítaná chyba. Toto sa zopakuje viac krát vždy s inou testovacou množinou, aby sa zvýšila presnosť odhadu chyby modelu. Zvoliť správnu hodnotu pre parameter k nie je triviálna úloha, no pre jednoduchosť nadobúda v tomto riešení celočíselnú časť 10% počtu inštancií v trénovacej množine.

Malý počet explicitne ohodnotených tweetov spôsobuje, že k najčastejšie nadobúda hodnoty len 1 alebo 2, preto aj rozhodnutie, či bude v danom sedení použitý odhad explicitnej spätnej väzby je vykonané zjednodušene; počas krížovej validácie sa zaznamenáva len binárna hodnota vyjadrujúca, či všetky inštancie v testovacej množine boli klasifikované správne. Validácia sa vykoná pre všetky inštancie a úspešnosť sa vyjadrí ako podiel správne klasifikovaných testovacích množín k počtu všetkých testovacích množín. Hranica úspešnosti bola pre testovacie účely volená v rozmedzí 60% až 80%. Úspešnosť uvedeného prístupu k modelovaniu vzťahu medzi implicitnou a explicitnou spätnou väzbou je overená praktickými testami a analyzovaná v časti 5.1.2.

4.1.4 Sledované charakteristiky obsahu, autorov a metadát tweetov

Pre účely predikovania zaujímavosti tweetov pre používateľa je potrebná množina atribútov, ktorými je možné tweety charakterizovať. Existujúcim riešeniam v tejto oblasti som sa venoval v častiach 2.3.1 a 2.3.2. Vzhľadom k zmene prístupu k riešeniu oproti experimentálnemu prototypu

z časti 3 a využitiu tematickej sumarizácie nie je pre charakterizovanie obsahu použité Hybrid TF-IDF [24].

Spomedzi základných charakteristík hovoriacich o jednotlivých tweetoch sú použité:

- počet retweetov (retweet count),
- počet URL,
- počet spomenutí používateľov (počet výskytov „@meno_pouzivatela“ v texte tweetu),
- počet hashtag-ov (počet výskytov „#znenie_hashtagu“ v texte tweetu),
- počet znakov v texte tweetu (čo umožňuje algoritmu strojového učenia nájsť aj súvislosť hodnôt iných atribútov s dĺžkou tweetu).

Pre autorov tweetov sú tiež sledované niektoré charakteristiky, ktoré tvoria ďalšie atribúty pre charakterizovanie tweetov, ktoré daní používatelia uverejnili. Patria k nim:

- počet uverejnených tweetov,
- počet uvedení používateľa v zoznamoch (listed count),
- FollowerRank [7] (pozri 2.3.1).

V analýze problematiky sú uvedené aj ďalšie ukazovatele s nádejnými výsledkami. Popularity Score [8] je však výpočtovo náročný a vzhľadom na jeho povahu ho ani nie je prakticky možné vypočítať vopred. TwitterRank [10] sa spolieha na odporúčania na základe aktivity používateľa na Twitteri, čo jednak nie je v súlade so zvoleným cieľovým používateľom a tiež nerieši situáciu s témami, s ktorými sa používateľ ešte na Twitteri nestretol.

Viacero analyzovaných prác sledovalo okrem spomínaných charakteristík, z ktorých veľká časť je poskytovaná priamo Twitterom, aj emócie v tweete. Napríklad v [9] autori sledovali výskyt emotikon, interpunkčných znamienok a opakovanie hlások v slovách (napr. „cooool“) . V [12] bol využitý prístup z práce ANEW [13], v rámci ktorej bol vytvorený slovník slov, ktoré boli ohodnotené podľa PAD modelu, čiže obsahovali vyjadrenie emócií podľa troch rozmerov. Keďže cieľom riešenia ja navrhnúť také atribúty, cez ktoré je možné sledovať rôznu motiváciu používateľa počas objavovania obsahu, je emocionálne zafarbenie považované za vhodný spôsob ako napríklad odlíšiť tweety vyjadrujúce silné názory a postoje či obsiahnuť situáciu, kedy sa používateľovi páčia tweety s vyjadreniami, ktoré korešpondujú s jeho náladou.

Pre tento účel bol použitý práve slovník ANEW, ktorý obsahuje 1030 slov. Pri implementácií a testovaní sa však ukázalo, že veľmi malé množstvo tweetov obsahovalo jedno alebo dokonca viac slov z tohto slovníka a pritom pri pohľade na text bolo zreteľné, že niektoré slová sú expresívne po emocionálnej stránke. Podarilo sa však nájsť novšiu prácu, ktorá tiež používa slovníkový prístup k počítaniu emocionálneho zafarbenia. AFINN [27] však už nepoužíva trojrozmerný model pre vyjadrenie emócie a opisuje ich len cez jeden rozmer tzv. „valence“, kedy vyjadrenia ako *vynikajúci*, *dych berúci* a *hurá* dosahujú maximum a negatívne expresívne

vyjadrenia naopak dosahujú minimum. Tento slovník obsahuje 2476 slov, ktoré boli vybraté zo slov vyskytujúcich sa v mikroblogoch [27]. Pre text tweetu sú spriemerované „emocionálne hodnoty“ slov, ktoré boli nájdené v slovníkoch. Stále však ostávajú niektoré tweety bez vyjadrenia emocionálneho zafarbenia, aj keď je prítomné vo forme emotikon.

Štúdia [9] neuvádza, akým spôsobom boli emotikony analyzované, či sa zaznamenával len ich výskyt alebo boli kategorizované do dvoch či viacerých skupín alebo akým spôsobom boli extrahované. Aby boli analyzované aj emócie z tweetov bez zhody slov v slovníkoch, ktoré ale obsahujú emotikony, vytvoril som na tento účel regulárne výrazy, ktoré extrahujú emotikony podľa toho, do akej skupiny patria. Skupiny boli vytvorené práve podľa hodnoty atribútu *valence*, čím sa dosiahne jednotné zaradenie emócií bez ohľadu na to, či sú vyjadrené slovne alebo emotikonami. Keďže v niektorých prípadoch mali na 10 stupňovej mierke slová rozdielnu hodnotu *valence* v ANEW a AFINN, bol pre emotikony použitý ich aritmetický priemer. Pokiaľ sa priemer nachádzal práve medzi dvoma celými číslami, hodnota bola zaokrúhlená smerom k AFINN, keďže tento slovník je novší a vytvorený s ohľadom na komunikáciu na Twitteri. Keďže riešenie sa zameriava na obsah v angličtine, analyzované sú emotikony tzv. západného štýlu ako `:)` a `8-D` a nie `^_^` alebo `(°Д°)`. Za základ zbierky emotikon slúžil zoznam z článku z Wikipedie[1].

Regulárne výrazy vždy začínali a končili výrazmi označujúcimi hranicu slov. Aby sa dosiahlo väčšie pokrytie a prehľadnejšie regulárne výrazy, boli emotikony rozdelené na časti, ktoré môžu byť spoločné pre väčšinu vyjadrených emócií. Ide o časť tváre nad očami, oči a nos, ktoré môžu byť vyjadrené rôznymi znakmi práve podľa emócie a časť nad očami a nos môže chýbať. Časti ako ústa či miesto medzi očami a nosom sa pre vyjadrenie rôznych pocitov a nálad líšia. Vzhľadom na použité prostredie (PHP) sú regulárne výrazy vyjadrené v PCRE formáte (Perl Compatible Regular Expressions). Pre bezproblémové spracovanie sú špeciálne znaky z tweetov konvertované do tzv. HTML entít, čiže napríklad znaky > a < sú vyjadrené ako `>` a `<`. Spoločné časti emotikon sú vyjadrené ako časti regulárnych výrazov:

- (nepovinná) časť nad očami je vyjadrená premennou `${aboveEyes}` ako `(?:(?:>)|[}3oO0])?`
- výraz pre oči v premennej `${eyes}` je `[;:8xX]`
- (nepovinný) nos je vyjadrený v premennej `${nose}` ako `[-=o\^\~]?`

Potom jednotlivé skupiny emócií so spoločnou hodnotou *valence* vyzerajú nasledujúco:

- pre vyjadrenie emócií s *valence* ekvivalentným so slovami *šťastný* a *láska* (*happy*, *love)*

```
/${wordBoundary1}(?:${aboveEyes}${eyes}\'?${nose}(?:[\)
}\]3]|(?:&gt;))+${wordBoundary2})|(?:&lt;3)/
```

[1] http://en.wikipedia.org/wiki/List_of_emoticons#Western

- emotikony ekvivalentné so slovami *hravý*, *bozk* alebo *„diablík“* (*playful*, *kiss*, *daredevil*)
  `/${wordBoundary1}${aboveEyes}${eyes}${nose}(?:[Ppb\*])+${wordBoundary2}/`
- *smiech* (*laugh*) je vyjadrený ako
  `/${wordBoundary1}${aboveEyes}${eyes}\'?${nose}D+${wordBoundary2}/`
- emotikony vyjadrujúce pocity opisované slovami *zahanbený*, *plačúci*, *šok*, *znudený* alebo *chorý/znechutený* (*embarassed*, *crying*, *shock*, *bored* alebo *sick*)
  `/${wordBoundary1}${aboveEyes}${eyes}\'?${nose}(?:[0Oo\$\/\\\\]|(?:#{2,}))+${wordBoundary2}/`
- *smutný* a *hnev* (*sad*, *angry*) zachytáva výraz
  `/${wordBoundary1}${aboveEyes}${eyes}${nose}(?:[\({Cc\[@]|(?:<)|(?:\|\|))+${wordBoundary2}/`

Pre charakterizovanie obsahu bol vybratý aj počet URL v príspevku, keďže tento ukazovateľ je dôležitý z viacerých aspektov pri modelovaní záujmu používateľa, ako bolo tiež uvedené v analýze prác [7, 8, 9]. Keďže je použité aj rozšírenie textu tweetu o názov odkazovanej stránky navrhnuté v časti 4.1.2, je k dispozícii aj plná adresa cieľovej stránky (jej rozšírená, pôvodná forma nie len skrátená forma). Pokiaľ však výskyt URL je tak dôležitý, môže byť užitočné vyťažiť z tohto aspektu ešte viac. Z pohľadu na rôzne tweety je vidieť, že obsah zdieľaný používateľmi je možné kategorizovať podľa typu média. Často sú odkazované články, samostatné fotografie, videá a v menšej miere aj zvukové záznamy. Práve kategórie podľa typu odkazovaného obsahu by mohli pomôcť vyjadriť aktuálne záujmy používateľa čiže to, aký obsah sa mu páči v súvislosti so zadaným dopytom.

Pre preukázanie tohto konceptu je implementovaná jednoduchá kategorizácia založená na analýze URL. Na zdieľanie obrázkov je veľmi často použitá jedna zo služieb Twitter, Twitpic, Facebook, Instagram, Pinterest, Tumblr a Flickr. Pre videá prevláda YouTube a menšia časť príspevkov obsahuje URL videí na Vimeo. Zdieľanie čisto audio obsahu je prakticky naviazané na službu SoundCloud. Čo sa týka článkov, nie je možné väčšiu časť URL klasifikovať tak jednoducho ako v prípade predchádzajúcich kategórií. Správy a články sa obvykle nachádzajú na spravodajských serveroch, ktorých je príliš veľké množstvo než aby bolo možné efektívne vytvoriť dostatočne obsiahly katalóg rôznych domén. Preto články momentálne spadajú do neznámej kategórie spolu s ostatným obsahom.

Samozrejme, presnejšia klasifikácia by mohla zahŕňať aj analýzu obsahu stránky, čím by bolo možné zisťovať, aký obsah na stránke prevláda. To je však príliš náročné na implementáciu

v súvislosti s riešením v tejto práci a taktiež si to vyžaduje väčšie množstvo času, čo predstavuje problém pre analýzu vykonávanú počas načítavania obsahu a jeho prezentovania používateľovi.

4.1.5 Model preferencií používateľa

Charakteristiky uvedené v predchádzajúcej časti tvoria atribúty pre model preferencií používateľa. Trénovacia množina je tvorená inštanciami (tweetmi), ktoré sú priradené k jednej z dvoch tried (zaujímavé/nezaujímavé). Klasifikácia do dvoch tried je pre trénovaciu množinu získaná buď priamo od používateľa, ktorý ohodnotí tweet hviezdičkami alebo je pre prečítané a používateľom neohodnotené tweety predikovaná modelom interpretácie implicitnej spätnej väzby. Ako bolo uvedené v časti 4.1.3, pri voľbe troch hviezdičiek z piatich nie je tweet klasifikovaný do žiadnej z tried, preto takéto tweety nie sú zaradené do trénovacej množiny.

Všetky z doteraz spomenutých atribútov sa týkajú povahy obsahu, no nie priamo toho, o čom je tweet z obsahového hľadiska. Čiže v ideálnom prípade je na základe uvedených charakteristík možné nájsť vzor, ktorý opisuje aké tweety používateľa zaujímajú. Napríklad ak používateľ skúma obsah Twitteru pre nejaké mesto, kam by chcel cestovať, môžu ho zaujímať názory ľudí resp. či sú ich reakcie a skúsenosti pozitívne alebo skôr negatívne. Okrem toho môže byť pre neho prínosom prezerať amatérske fotografie danej destinácie namiesto čítania PR článkov. V tomto prípade dobre poslúžia práve atribúty emocionálneho zafarbenia, atribút hovoriaci o výskyte URL v tweete a atribút typu obsahu na odkazovanej stránke. Čo však pri odlišnej motivácii používateľa, ktorý napríklad zadá veľmi široký dopyt ako „news“ čiže správy/novinky? Je pravdepodobné, že výskyt URL v tweete bude znova dôležitým ukazovateľom, no je tiež možné, že ho napríklad budú zaujímať správy a články o športe a nie obsah s politickou tematikou. V tomto prípade by na základe uvedených charakteristík nebolo možné tweety dostatočne diferencovať.

Pre riešenie tohto problému je samozrejme možné využiť prístup založený na TF-IDF miere. Pre krátke dokumenty už bol analyzovaný aj tzv. Hybrid TF-IDF ukazovateľ v časti 3.5. Je ale dôležité myslieť na zvolený prístup k riešeniu a vybrať správnu alternatívu. Uvedené miery dokážu identifikovať slová, ktoré najlepšie vystihujú nejakú množinu dokumentov, no v uvedenej modelovej situácii by sa mohlo stať, že pre veľmi krátky rozsah tweetov sa nenájdu žiadne významné slová, ktoré sa v zaujímavých tweetoch pre používateľa vyskytujú výrazne viac než iné slová (to aj naznačil praktický test z časti 3.6.3). Na dostatočné obsiahnutie toho, aké témy používateľa zaujímajú by aj tak bolo potrebné použiť napríklad prístup z [15] bližšie opísaný v časti 2.3.3, ktorý je ale výpočtovo náročný, keďže používa veľké množstvo predspracovania a komplexné algoritmy identifikácie významných slov a ich príslušnosti k témam. Keďže je však vykonávaná tematická sumarizácia, ktorej základom je identifikovanie významných fráz (n-gramov), na účel charakterizovania obsahu som sa rozhodol využiť informácie získané pri extrakcii fráz. Cieľom riešenia je vytvoriť model preferencií a na jeho základe odporúčať tweety. Pre povahu

tweetov sa teda použijú charakteristiky z predchádzajúcej časti a pre vystihnutie obsahu využívam predpoklad, že ak sa používateľovi páči väčšina tweetov priradených k nejakej téme, môže to byť spôsobené práve obsahom tweetov, keďže zahŕňajú aspoň jednu spoločnú frázu. Samozrejme tieto tweety môžu mať spoločné aj iné charakteristiky, no zistiť, ktoré atribúty sú dôležité a v akom vzťahu sú, je ponechané na algoritmus strojového učenia.

Pre tweet môže existovať priradenie k viacerým frázam. To znamená, že ak používateľa nezaujíma väčšina tweetov pre tému napr. „Bratislava“, mohli by byť nečítané tweety s touto témou považované za nezaujímavé. Zaujímavé tweety, ktoré sú v menšine nie je možné vystihnúť témou „Bratislava“, no pokiaľ všetky obsahujú napríklad frázu „hdr photography“ a zároveň väčšina prečítaných tweetov k tejto druhej téme bola pre používateľa zaujímavá, je možné predpokladať, že nečítané tweety s frázou „hdr photography“ sa budú používateľovi páčiť. Istota, s ktorou je možné dané rozhodnutie vysloviť sa môže meniť v závislosti od ďalších fráz, ktoré sa v tweetoch vyskytujú. Napríklad jeden tweet z témy „hdr photography“ môže obsahovať aj slovo „night“ a druhý slovné spojenie „Old Town“. Ak sa používateľovi páčila väčšina prečítaných tweetov s frázou „night“, ale len polovica prečítaných tweetov s témou „Old Town“, je intuitívne pravdepodobnejšie, že sa mu budú viac páčiť tweety s frázami „night“ a „hdr photography“ než tie, ktoré obsahujú „hdr photography“ a „Old Town“. Preto zavádzam atribúty vyjadrujúce obsah tweetu ako priemer časov pozornosti a priemer kliknutí na URL v tweetoch. Pre každý tweet sa načíta zoznam tém, ku ktorým prislúcha a potom sa v rámci každej témy vypočíta priemer týchto dvoch hodnôt čiže priemerný čas pozornosti na jeden tweet (v rámci tweetov k danej téme) a pomer kliknutí k počtu všetkých tweetov s URL (príslušných k danej téme). Z týchto priemerných hodnôt pre témy sa vypočítajú dve hodnoty, priemer časov pozornosti a priemer pomeru kliknutí na URL, čo tvorí výslednú hodnotu týchto dvoch atribútov pre daný tweet. Ako v prípade modelu interpretujúceho implicitnú spätnú väzbu, sú všetky atribúty normalizované min-max metódou.

Tvorba modelu preferencií používateľa je obdobná postupu pri modeli interpretácie implicitnej spätnej väzby. Trénovaciu množinu tvoria inštancie (tweety), ktorých atribúty boli opísané vyššie (citové zafarbenie, metadáta príspevkov a autorov a charakteristiky zaujímavosti podľa príslušných tém) a cieľovým atribútom je klasifikácia do tried zaujímavý a nezaujímavý. Použitý algoritmus ostáva SVM. Relevancia modelu je sledovaná rovnakou krížovou validáciou ako pri modeli z časti 4.1.3. Všetky nečítané tweety tvoria testovaciu množinu, čiže sú klasifikované vytvoreným modelom. Používateľovi je na základe tejto klasifikácie možné odporúčať buď samostatné tweety alebo témy. Pri témach je možné sledovať množstvo tweetov, ktoré boli označené za zaujímavé a tie, ktoré boli označené za nezaujímavé. Ak prevládajú zaujímavé tweety, téma môže byť odporučená a ak prevládajú tweety predikované ako nezaujímavé, tak téma môže byť označená za neodporúčanú.

Model pracuje s triedami vyjadrenými ako čísla 1 a -1. Výstup klasifikácie je uvádzaný v podobe reálnych čísel. Aby sa nebrali do úvahy rozhodnutia s malou istotu, hodnoty medzi -0.1 a 0.1 vrátane týchto hraníc sú interpretované ako nemožnosť modelu rozhodnúť o zaujímavosti a takéto tweety nie sú označené ani za zaujímavé, ani za nezaujímavé.

4.2 Celkový opis riešenia a interakcie so systémom

Táto časť obsahuje opis súvislostí medzi jednotlivými časťami systému a vzťah, s akým navrhnuté čiastkové riešenia tvoria funkcionalitu systému. Zhrnutie tohto opisu zachytáva model použitia systému používateľom vyjadrený diagramom aktivít na Obr. 4.1.

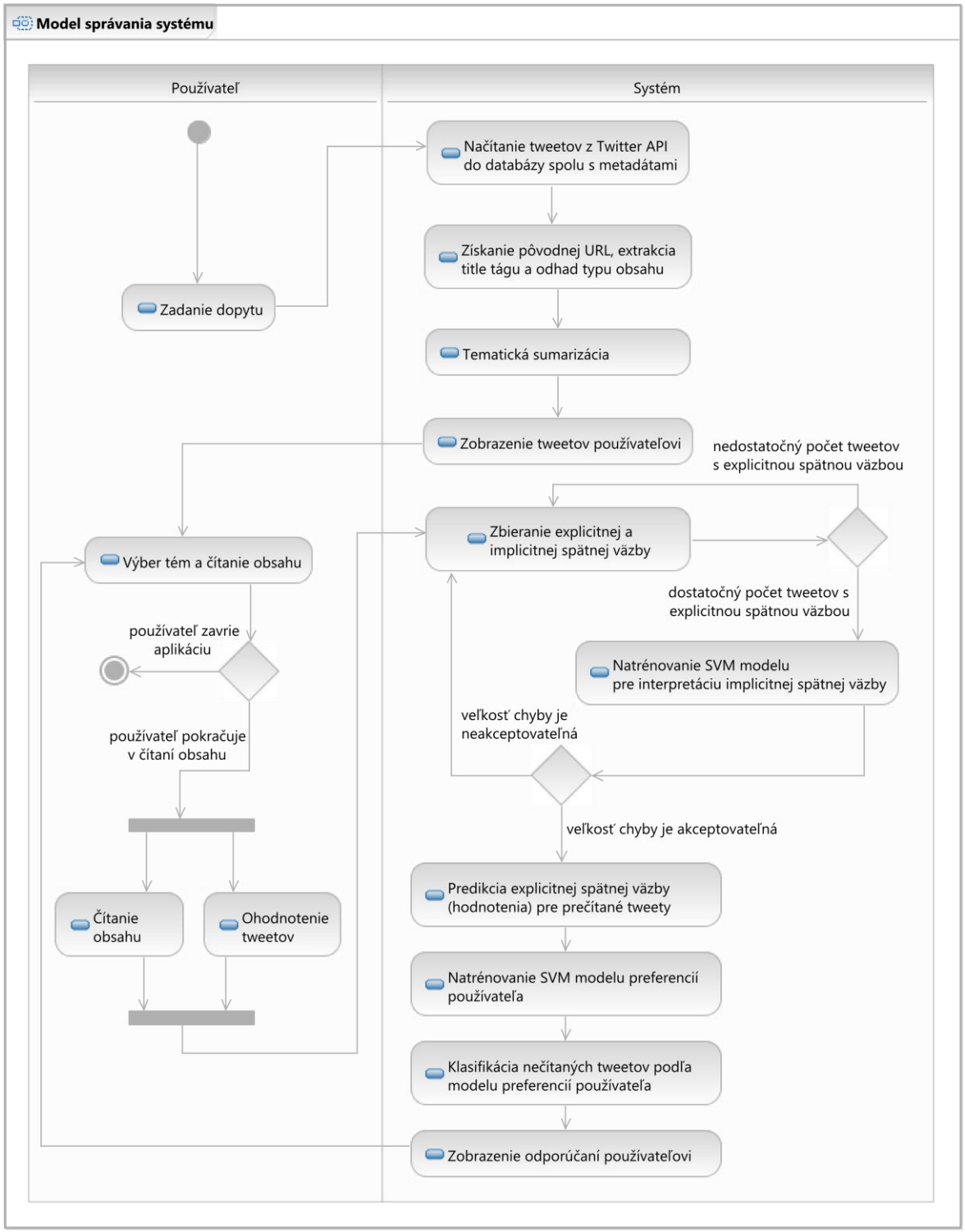
Ako bol už uvedené, navrhnuté riešenie počíta s cieľovým používateľom, ktorý nemusí byť aktívnym používateľom Twitteru, keďže riešenie počíta s modelom preferencií používateľa len v súvislosti s aktuálnym dopytom a len v kontexte jedného sedenia. Tým nie je potrebné mať k dispozícii napríklad informácie o iných aktivitách používateľa, ktoré napovedajú o jeho záujmoch. Vzhľadom na to, že Twitter oznámil, že v najbližšej dobe začne podporovať len autorizované požiadavky na Twitter API, bolo žiaľ nutné vyžadovať autentifikáciu používateľa cez jeho Twitter účet. Takáto autentifikácia je použitá len na zabezpečenie, že požiadavky navrhnutého systému cez Twitter API budú vybavené.

Prvým krokom pri použití navrhovanej webovej aplikácie je teda autentifikácia používateľa prostredníctvom Twitter účtu a autorizácia navrhovanej aplikácie vykonávať požiadavky v mene používateľa. Používateľ potom zadá dopyt (jedno alebo viac slov) a počká, kým aplikácia načíta a spracuje obsah. Táto fáza tvorí jediné významné časové oneskorenie pri interakcii s používateľom. Tento čas je využitý na to, aby sa používateľovi zobrazil jednoduchý návod na používanie aplikácie.

Spracovanie zahŕňa požiadavku na Twitter API pre načítanie tweetov, ktoré obsahujú dopyt zadaný používateľom. Použité Twitter REST API pracuje s neúplným indexom tweetov pre nešpecifikovaný počet predchádzajúcich dní. Je možné načítať populárne tweety, najnovšie tweety alebo zmiešané populárne a najnovšie tweety. Keďže navrhovaný prístup je potrebné overiť, načítavajú sa zmiešané populárne tweety s najnovšími tweetmi, aby výsledné overenie viac zachytávalo úspešnosť navrhovanej metódy než by to bolo v prípade načítania len populárnych tweetov. Načítané tweety sú uložené do databázy.

Nasleduje fáza spracovania tweetov, kedy je z databázy načítaných najnovších 200 tweetov, ktorých text obsahuje zadané kľúčové slová. Pre všetky tweety v databáze, ktoré obsahujú URL a tieto URL neboli spracované, sa vyšlú http požiadavky a z odpovedí sa extrahuje hodnota poľa, ktoré obsahuje cieľovú adresu presmerovania. Väčšina najpoužívanejších služieb skracovania URL totiž práve takto zabezpečuje svoju funkcionalitu. Načíta sa obsah z daných URL a extrahuje sa text z HTML tágu `title`. Na pôvodné, neskrátené URL sa aplikujú regulárne výrazy, ktoré podľa pravidiel navrhnutých v 4.1.5 klasifikujú typ obsahu stránky do jednej z kategórií audio,

video, obrázok a neznáme. Neskrátené URL, obsah `title` tágu a typ obsahu stránky sa uložia do databázy.



Obr. 4.1 Diagram aktivít zobrazujúci interakciu používateľa s navrhovaným systémom

Po tomto kroku nasleduje tematická sumarizácia bližšie opísaná v časti 4.1.1. V tweetoch sa však najprv URL nahradia názvom odkazovanej stránky (obsahom `title` tágu), potom sa vykoná

deduplikácia a až následne sa extrahujú n-gramy. Finálny zoznam tém (fráz) s priradenými množinami tweetov sa potom zobrazí používateľovi, ktorý môže začať v čítaní obsahu.

Používateľ kliknutím na frázu zobrazí tweety, ktoré danú frázu obsahujú. Ako bolo spomenuté, tweety sú čitateľné vtedy, keď je na nich umiestnení kurzor myši. Používateľ okrem čítania textu tweetu, môže kliknúť na URL a čítať externý obsah. Pre každý tweet, ktorý je čítaný (resp. je nad ním umiestnený kurzor myši) dlhšie ako je prahová hodnota (zvolených je 1000 ms) sa uložia informácie o implicitnej spätnej väzbe. Pre tweety, ktoré sa používateľ rozhodne ohodnotiť použitím škály s hviezdičkami sa zaznamená aj táto explicitná spätná väzba.

Po ohodnotení aspoň piatich tweetov ako nezaujímavých a aspoň piatich tweetov ako zaujímavých systém natrénuje model interpretujúci implicitnú spätnú väzbu. Pred spustením trénovania algoritmom SVM sa pre dané tweety načítajú atribúty zo štruktúr, v ktorých sú uložené a prebehne normalizácia ich hodnôt. Ak počas krížovej validácie bolo správne klasifikovaných aspoň 65% čiastkových testovacích množín (približne 10% inštancií z trénovacej množiny), použije sa tento model pre interpretovanie implicitnej spätnej väzby pri prečítaných neohodnotených tweetoch. Potom nasleduje trénovanie modelu preferencií používateľa. Natrénovaný model je použitý, ak je rovnako ako pri predchádzajúcom prípade úspešne klasifikovaných aspoň 65% čiastkových testovacích množín v krížovej validácii. Predikované triedy pre nečítané tweety sú zobrazené používateľovi farebným odlíšením indikátora pri tweetoch. Ak témy s neprečítanými tweetmi obsahujú väčší počet tweetov predikovaných ako zaujímavých, zobrazí sa rovnakým spôsobom indikátor odporúčania. Analogicky sa zobrazuje aj indikátor neodporúčaných tém, ak prevládajú neodporúčané tweety. Témy sú zároveň nanovo zoradené, kde sa najprv zobrazujú odporúčané témy, potom témy, pre ktoré neprevažuje žiadna z tried (veľmi podobný počet odporúčaných aj neodporúčaných tweetov alebo veľký počet tweetov, ktoré model preferencií klasifikoval na rozhraní zaujímavých a nezaujímavých), nasledujú témy neodporúčané a nakoniec sú zobrazené témy, pre ktoré používateľ prečítal všetky tweety. V rámci týchto štyroch skupín sú témy zoradené zostupne podľa skóre priradeného v tematickej sumarizácii.

Oba z modelov sú iteratívne trénované po tom, ako pribudnú inštancie v trénovacích množinách, čiže po prečítaní ďalších tweetov. Aby serverová časť systému nebola príliš zahltená požiadavkami, vykonáva sa opakované zostavovanie modelov vždy až po piatich nových vzorkách v trénovacej množine, čiže v prípade modelu interpretujúceho implicitnú spätnú väzbu ide o explicitne ohodnotené tweety, pre model preferencií to môžu byť aj neohodnotené tweety, pretože ich klasifikácia je vykonaná práve modelom pre implicitnú spätnú väzbu (ak tento spĺňa požiadavku na úspešnosť uvedenú v predchádzajúcom odstavci).

Náhľad na používateľské rozhranie počas používania aplikácie sa vo forme dvoch obrázkov a stručného vysvetlenia nachádza v Prílohe 3.

4.3 Overenie riešenia

Účelom navrhovaného prístupu je poskytnúť predstavu o témach vyskytujúcich sa na Twitteri súvisiacich s dopytom, ktorý zadal používateľ. Hlavným prínosom je však podpora takéhoto skúmania obsahu vytvorením modelu preferencií, ktorý pri menšom množstve explicitnej spätnej väzby umožní odporúčanie obsahu podľa typu a tém používateľovi, čím mu zjednoduší navigáciu v rôznorodom obsahu. Zhodnotenie prístupu sa preto sústredí predovšetkým na úspešnosť modelu preferencií. Model je určený pre klasifikáciu tweetov, ktoré boli načítané k zadanému dopytu a boli súčasťou tematickej sumarizácie. Výstupom je predpoveď toho, či je tweet pre používateľa zaujímavý, nezaujímavý alebo informácia, že o tomto priradení nie je možné rozhodnúť na základe aktuálneho modelu. Úspešnosť modelu preferencií používateľa je teda možné opísať podielom správne klasifikovaných neprečítaných tweetov k celej množine neprečítaných tweetov. Prakticky ale nie je možné poznať správnu triedu pre všetky neprečítané tweety. Táto informácia totiž závisí na názore používateľa, ktorý je potrebné získať. Aby sa zvýšila pravdepodobnosť, že používatelia, ktorí sa zúčastnia praktického testu boli ochotní poskytnúť (explicitnú) spätnú väzbu, je spätná väzba vyžiadaná pre menšie množstvo tweetov než je počet všetkých neprečítaných príspevkov.

4.3.1 Postup praktického testovania

V časti 4.2 sa nachádza opis práce so systémom v bežnom režime. Mód overenia je veľmi podobný, no odlišuje sa v dvoch podstatných aspektoch. Od používateľa sa po prečítaní 40 tweetov vyžaduje explicitná väzba k vybratým tweetom. Aby bolo možné sledovať úspešnosť oboch zostavovaných modelov: model interpretujúci implicitnú spätnú väzbu (skrátene MIISV) a model preferencií používateľa (skrátene MPP), vyberie sa:

- 15 tweetov, ktoré používateľ prečítal a MIISV odhadol, že sú pre používateľa nezaujímavé,
- 15 tweetov, ktoré používateľ prečítal a MIISV odhadol, že sú pre používateľa zaujímavé,
- 15 tweetov, ktoré používateľ nečítal a MPP predikoval, že budú pre používateľa nezaujímavé
- a 15 tweetov, ktoré používateľ nečítal a MPP predikoval, že budú pre používateľa zaujímavé.

Tieto tweety sú vybraté náhodne. K výberu stanoveného počtu tweetov pre obe možné triedy pre oba modely som pristúpil z dôvodu, že počet inštancií pre jednotlivé triedy nemusí byť rovnomerný a mohlo by tak dôjsť k situáciám, kedy je počet vzoriek pre niektorú z tried výrazne menší. Potom by výsledky overenia hovorili skôr o úspešnosti klasifikátorov (modelov) na triedach, pre ktoré prevládali náhodne vybraté tweety. Pri testoch však viackrát došlo k situácii, v ktorej neexistoval aspoň stanovený počet vzoriek a bolo tak od používateľa vyžadované ohodnotiť menej než 60 tweetov.

Druhým rozdielom oproti bežnému správaniu navrhovaného systému bolo nezobrazovanie odporúčaní skôr než používateľ poskytol spätnú väzbu v rámci overenia a ani počas neho. Tým som chcel docieliť odstránenie ovplyvnenia názoru používateľa v súvislosti s tým, čo mu systém odporúča alebo neodporúča.

4.3.2 Charakteristika výstupov praktického testovania

V rámci overenia a zhodnotenia navrhovaného prístupu je sledovaná úspešnosť modelu preferencií používateľa a modelu interpretujúceho implicitnú spätnú väzbu. Preto sa pred vyžiadaním spätnej väzby od používateľa vykoná natrénovanie overovaných modelov a pre vybraté tweety sa uloží predikcia tried z oboch modelov do CSV súborov (z angl. comma-separated values). Po ohodnotení všetkých tweetov používateľom sa rovnako uloží klasifikácia do tretieho CSV súboru. Zhodnotenie je potom vykonané na základe údajov z týchto troch súborov.

Pre analyzovanie alternatív sú pre interpretovanie implicitnej spätnej väzby natrénované dva modely. Prvý tak, ako bolo doteraz opísané a druhý, ktorého trénovacia množina neobsahuje veľmi podobné inštancie, ktoré patria do rôznych tried. Trénovaciu množinu v tomto prípade tvoria vektory v trojrozmernom priestore (charakteristiky implicitnej spätnej väzby bez atribútu určujúceho príslušnosť k triede), ktoré nadobúdajú hodnoty od 0 do 1. Podobnosť je vyhodnocovaná kosínusovou mierou podobnosti s hranicou stanovenou na 0,99999.

Na úspešnosť modelu preferencií používateľa pri klasifikácii ešte neprečítaných tweetov vplýva mnoho aspektov. Keďže množina charakteristík bola navrhnutá tak, aby pokrývala viaceré možné motivácie používateľa (resp. rôzne aspekty jeho preferencií) a rešpektovala požiadavku na nízky výpočtový výkon, pri možnostiach vylepšenia úspešnosti som sa zameral na ostatné hľadiská než overovanie veľkého množstva kombinácií atribútov. Na výslednú úspešnosť môže mať samozrejme vplyv aj algoritmus strojového učenia, či už ide o to, ktorý algoritmus je zvolený alebo o jeho konkrétnu implementáciu či voľbu hodnôt parametrov. Trénovacia množina pozostáva z inštancií, ktoré tvorí 16 atribútov a zaradenie do triedy zaujímavý/nezaujímavý. To je relatívne veľký počet vzhľadom k počtu inštancií, ktoré túto množinu tvoria. Načítaných je 200 tweetov, z ktorých po odstránení podobných tweetov ostáva menej (počet sa pohybuje obvykle medzi 120 a 180), z čoho bola stanovená spodná hranica pre overenie 40 prečítaných tweetov. Som presvedčený, že na základe 40 inštancií je veľmi obťažné zostrojiť model rešpektujúci 16 vstupných atribútov.

Pri priaznivej situácii, kedy tweety tvoria zopár zhlukov v tomto 16 rozmernom priestore potom pre algoritmus SVM predstavujú zložitejšie zostavenie modelu tie inštancie, ktoré sa nachádzajú blízko seba a patria do rôznych tried. V konečnom dôsledku môže byť úspešnejší model, ktorý nebude trénovaný na takýchto vzorkách. Preto je zostavených aj niekoľko alternatívnych modelov bez podobných inštancií, kde je postup rovnaký ako v prípade modelu interpretujúceho implicitnú spätnú väzbu.

Vo všeobecnosti môže byť klasifikátor ovplyvnený aj veľkým rozdielom v počte inštancií pre jednotlivé triedy. Aby som sa uistil, že takéto nerovnomerné rozdelenie počtu inštancií pre jednotlivé triedy nemá vplyv na úspešnosť klasifikácie, sledujem úspešnosť pôvodného modelu a modelu, ktorý je trénovaný na upravenej trénovacej množine. Táto upravená množina obsahuje rovnaký počet inštancií pre obe triedy, čo je dosiahnuté tak, že pre menej početnú triedu sa zduplikuje potrebný počet náhodných inštancií.

Používateľa môžu zaujímať tweety podľa typu obsahu alebo podľa obsahu samotného. Typ obsahu je vystihnutý charakteristikami ako typ odkazovaného obsahu, emocionálne zafarbenie či výskyt spomenutí používateľov. Obsah je opísaný poslednými dvoma atribútmi, ktoré pomáhajú vyjadriť, či sú tweety príslušné témam, ktoré používateľa zaujímajú alebo nezaujímajú. Čas pozornosti a podiel kliknutí na URL nemusia byť vždy interpretované rovnako, preto je táto interpretácia ponechaná na SVM algoritmus, ktorý tak rozhodne na základe príslušností tweetov k triedam podľa explicitnej spätnej väzby a interpretovanej implicitnej spätnej väzby. Môže byť zaujímavé sledovať, či prevláda príslušnosť k témam alebo typ obsahu alebo sú v rovnováhe. Preto je pozorovaná úspešnosť na trénovacích množinách, ktoré berú do úvahy obidve skupiny atribútov alebo iba jednu z nich.

Tab. 4.1 Prehľad alternatív modelov preferencií používateľa podľa atribútov, ktoré využívajú, použitých metód predspracovania a podľa použitia modelu interpretujúceho implicitnú spätnú väzbu.

atribúty a metódy / označenie modelu	charakteristiky typu obsahu (m)	charakteristiky témy (t)	odstránenie veľmi podobných inštancií patriacich do rôznych tried (s)	kompenzácia rozdielneho počtu inštancií v triedach (c)	použitie modelu interpretujúceho implicitnú spätnú väzbu (f)
mtf	x	x			x
mtsf	x	x	x		x
tf		x			x
mf	x				x
mtcf	x	x		x	x
mtscf	x	x	x	x	x
tcf		x		x	x
mcf	x			x	x
mt	x	x			
mts	x	x	x		
t		x			
m	x				
tsf		x	x		x
msf	x		x		x

Na úspešnosť modelu preferencií používateľa môže mať tiež vplyv aj úspešnosť modelu interpretácie implicitnej spätnej väzby, ktorý do trénovacej množiny modelu preferencií zavádza zle klasifikované inštancie. Samozrejme je možné nepoužiť model interpretujúci implicitnú spätnú

väzbu, pokiaľ počas krížovej validácie dosahuje príliš veľkú chybu. No ak je jeho chyba menšia než stanovená hranica, je pre čítané no neohodnotené tweety odhadnutá klasifikácia a okrem korektne klasifikovaných vzoriek pribúdajú aj nekorektne klasifikované. Model implicitnej spätnej väzby môže však ľahko zdvojnásobiť či strojnásobiť počet inštancií v trénovacej množine pre model preferencií. Je teda otázne, do akej miery je možné tolerovať nepresnosť interpretácie implicitnej spätnej väzby. Preto je pozorovaný aj rozdiel v úspešnosti klasifikácie s použitím modelu interpretujúceho implicitnú spätnú väzbu a bez neho.

Keďže trénovanie modelu preferencií zaberá istý čas (typicky niekoľko sekúnd), nebolo praktické použiť všetky kombinácie spomínaných alternatív. Celkovo pripadalo do úvahy 24 zmysluplných možností (použitie modelu interpretácie spätnej väzby a metódy predspracovania je možné ľubovoľne kombinovať, no vždy musí byť použitá aspoň jedna množina atribútov). Tab. 4.1 zachytáva 14 vybratých alternatív modelu preferencií používateľa.

5 Zhodnotenie

5.1 Výsledky validácie riešenia v praxi

Systém v režime pre overenie riešenia vyžadoval od používateľa po 40 prečítaných tweetoch pre ľubovoľný dopyt explicitnú spätnú väzbu pre maximálne 60 tweetov. Polovica slúžila pre validáciu modelu interpretujúceho implicitnú spätnú väzbu a druhá polovica pre model preferencií používateľa. Pre obe predikované triedy (zaujímavé / nezaujímavé) systém vždy vybral maximálne 15 tweetov. Pokiaľ modely predikovali pre jednu z tried menej než 15 tweetov, bol vybratý dostupný počet tweetov (menší než 15). Predikovaná trieda nebola pre tweety nijak naznačovaná používateľovi a poradie tweetov bolo náhodné (vzhľadom na predikovanú triedu), čím mala byť zabezpečená eliminácia ovplyvnenia poskytnutej explicitnej spätnej väzby. Podrobnejšie informácie o prístupe sa nachádzajú v časti 4.3.1.

Celkový počet korektne ukončených sedení v režime overenia bol 25. Niektorí používatelia nerešpektovali pokyn nehodnotiť všetky tweety hviezdičkami, čo znamenalo, že pre overenie modelu pre implicitnú spätnú väzbu neostali tweety, na ktorých by bolo možné overiť predikciu explicitnej spätnej väzby na základe implicitnej. Okrem toho sa model interpretujúci spätnú väzbu nepoužíval pre model preferencií používateľa, ak pri krížovej validácií nedosiahol úspešnosť aspoň 65% (presný postup je opísaný v časti 4.1.3). Preto sú tiež pre overenie sledované len sedenia, kde bola dosiahnutá táto hranica. V konečnom dôsledku tak zostalo 12 sedení použitých na vyhodnotenie modelu interpretujúceho implicitnú spätnú väzbu. Model preferencií bol sledovaný vo všetkých 25 sedeniach.

5.1.1 Ukazovatele pre vyhodnotenie

Úspešnosť klasifikátorov je vypočítaná ako definuje Vzorec 5.1, čiže ako podiel počtu vzoriek so správne predikovanými triedami k počtu všetkých vzoriek v testovacej množine čiže pre tie vzorky, pre ktoré je predikcia vykonávaná.

$$\text{úspešnosť} = \frac{\textbf{\# správne klasifikovaných vzoriek}}{\textbf{\# všetkých vzoriek}} * \mathbf{100}\ \%$$

Vzorec 5.1 Výpočet úspešnosti klasifikátora.

Z testovacej množiny sú však vypustené tie vzorky, ku ktorým používateľ uviedol neutrálny postoj (hodnotenie troma hviezdičkami) alebo ku ktorým klasifikátor zaujal nejednoznačné stanovisko (hodnoty medzi -0,1 a 0,1 na stupnici, kde hodnoty -1 a 1 predstavujú dve triedy). Keďže nie je možné jednoducho vyhodnotiť úspešnosť v týchto prípadoch, boli takéto vzorky úplne vypustené z výpočtu úspešnosti.

Keďže rozloženie tried pre tweety v testovacej množine nemusí byť rovnomerné, úspešnosť nehovorí jasne o tom, do akej miery je klasifikátor lepší než náhodný klasifikátor.

Napríklad úspešnosť 80% znamená, že 80% tweetov je odporúčaných správne, no v prípade, že 80% tweetov sa používateľovi nepáči, klasifikátor predikujúci triedu „nezaujímavý“ pre všetky tweety síce dosiahne túto úspešnosť, no rovnakú by dosiahla aj náhodná zhoda.

Preto okrem úspešnosti jednotlivých modelov (resp. klasifikátorov či prediktorov) uvádzam aj hodnotu kappa štatistiky [28], ktorá meria úspešnosť ako proporciu z úspešnosti perfektného prediktoru. Zhoda medzi skutočnými a predikovanými triedami je upravovaná aj zhodou danou náhodou, ako zachytáva Vzorec 5.2. Zhoda je v tomto prípade počítaná ako úspešnosť uvedená vyššie. Hodnota 0% znamená, že overovaný klasifikátor je zhodný s náhodným klasifikátorom. 100% znamená, že overovaný klasifikátor je perfektným klasifikátorom. Náhodný klasifikátor je v tomto prípade náhodné priradenie tried s rovnakým rozložením (rovnaký počet priradených tried) ako u overovaného klasifikátora [28].

$$\kappa = \frac{\text{pozorovaná zhoda} - \text{náhodná zhoda}}{1 - \text{náhodná zhoda}} * 100\%$$

Vzorec 5.2 Kappa štatistika pre vyhodnotenie klasifikátora.

Pri viacerých vygenerovaniach náhodného rozloženia pre náhodný klasifikátor môže dôjsť k nadpriemerným odchýlkam, čo je možné kompenzovať tak, že sa náhodné rozloženie vygeneruje viackrát. Pri overení sú preto všetky sedenia brané do úvahy trikrát, kedy výsledky implementovaných klasifikátorov zostávajú rovnaké, ale náhodné rozloženie pre náhodný klasifikátor je vždy nanovo vygenerované.

Vyhodnotenie alternatív modelu preferencií používateľa porovnáva aj úspešnosť pre obe triedy osobitne, na čo je použitá tzv. F-measure čiže miera, ktorá kombinuje presnosť (z angl. precision) a pokrytie (z angl. recall). Na vyjadrenie týchto ukazovateľov boli použité hodnoty správneho zaradenia do triedy (true positive), nesprávneho zaradenia do triedy (false positive) a nesprávneho nezaradenia do triedy (false negative). Presnosť vyjadruje Vzorec 5.3, pokrytie Vzorec 5.4 a F-measure uvádza Vzorec 5.5.

$$presnosť = \frac{\text{správne zaradenie do triedy}}{\text{správne zaradenie do triedy} + \text{nesprávne zaradenie do triedy}}$$

Vzorec 5.3 Presnosť.

$$pokrytie = \frac{\text{správne zaradenie do triedy}}{\text{správne zaradenie do triedy} + \text{nesprávne nezaradenie do triedy}}$$

Vzorec 5.4 Pokrytie.

$$F - measure = 2 * \frac{\text{presnosť} * \text{pokrytie}}{\text{presnosť} + \text{pokrytie}}$$

Vzorec 5.5 F-measure.

5.1.2 Vyhodnotenie modelu interpretujúceho implicitnú spätnú väzbu

Ako bolo už uvedené, z celkového počtu 25 dokončených sedení bolo len 12 takých, v ktorých používatelia podľa pokynov nehodnotili všetky prečítané tweety hviezdičkami a súčasne model

dosiahol pri krížovej validácií pri trénovaní úspešnosť aspoň 65%. Sledované boli dve alternatívy modelov, kedy prvý je základný model opísaný v opise riešenia a druhý model je trénovaný na inštanciách, ktoré nie sú veľmi podobné (podľa kosínusovej miery podobnosti, detaily sú uvedené v časti 4.3.2). Okrem toho boli sledované aj výsledky hlasovania modelov; keďže ale ide len o dva modely, výsledok je braný do úvahy len vtedy, keď sa modely zhodli, inak hlasovanie bolo nerozhodné a tieto situácie neboli do sledovaných štatistík započítané.

V Tab. 5.1 sa nachádza vypočítaná úspešnosť pre overované klasifikátory (čiže implementované klasifikátory podľa návrhu riešenia) a náhodné klasifikátory. Kappa štatistika je vypočítaná pre každú alternatívu modelu, ako uvádza Vzorec 5.2. Záporné hodnoty kappa štatistiky znamenajú, že overovaný model je horší ako náhodný prediktor. Z pohľadu na kappa štatistiku pre model trénovaný bez veľmi podobných inštancií je vidieť, že v 9 prípadoch je implementovaný klasifikátor zhodný s náhodným klasifikátorom. Základný model trénovaný na všetkých inštanciách dosahuje lepšie výsledky. Na základe troch ukazovateľov implicitnej spätnej väzby bola správne predikovaná explicitná spätná väzba v približne 68% prípadoch oproti približne 59% prípadov správnej klasifikácie náhodným prediktorom. Hlasovanie modelov dosahuje horšie výsledky, keďže berie do úvahy aj model trénovaný bez veľmi podobných inštancií, ktorý je ale podľa výsledkov prakticky nepoužiteľný.

Tab. 5.1 Výsledky overenia modelu interpretujúceho spätnú väzbu.

sedenie	základný model			bez podobných inštancií			hlasovanie		
	úspešnosť		kappa	úspešnosť		kappa	úspešnosť		kappa
	overovaný prediktor	náhodný prediktor		overovaný prediktor	náhodný prediktor		overovaný prediktor	náhodný prediktor	
1	62,50%	50%	25%	37,50%	37,50%	0%	50%	33,33%	25%
2	75%	62,50%	33,33%	25%	25%	0%	50%	33,33%	25%
3	100%	100%	0%	100%	100%	0%	100%	100%	0%
4	75%	44,44%	55%	69,23%	69,23%	0%	87,50%	62,50%	66,67%
5	70%	70%	0%	0%	0%	0%	70%	70%	0%
6	65%	48,33%	32,26%	52,63%	50,88%	3,57%	83,33%	50%	66,67%
7	55%	56,67%	-3,85%	40%	40%	0%	42,86%	45%	-3,90%
8	85,71%	85,71%	0%	14,29%	14,29%	0%	0%	0%	0%
9	23,08%	23,08%	0%	0%	0%	0%	23,08%	23,08%	0%
10	81,82%	84,85%	-20%	0%	0%	0%	81,82%	84,85%	-20%
11	80%	46,67%	62,50%	60%	46,67%	25%	70%	36,36%	52,86%
12	65%	55%	22,22%	42,11%	52,63%	-22,22%	60%	59,09%	2,22%
celkovo	**68,02%**	**58,91%**	**22,17%**	**45,99%**	**46,23%**	**-0,45%**	**64,95%**	**56,20%**	**19,97%**

5.1.3 Vyhodnotenie modelu preferencií používateľa

Model preferencií používateľa bol natrénovaný a overovaný vo všetkých 25 sedeniach. V 12 prípadoch bol použitý aj model interpretujúci implicitnú spätnú väzbu na zväčšenie počtu inštancií

v trénovacej množine pre model preferencií. Vo zvyšných 13 prípadoch preto sledované alternatívy modelov odlišujúcich sa len v použití alebo nepoužití modelu pre implicitnú spätnú väzbu dosahujú rovnaké výsledky.

Keďže výsledky pre 16 alternatívnych modelov pre všetkých 25 sedení obsahujú príliš veľké množstvo údajov, Tab. 5.2 zachytáva už celkové výsledky pre alternatívne metódy (bez uvedenia výsledkov pre jednotlivé sedenia). Alternatívne modely sú označené skratkami, ktoré vyjadrujú, ktoré skupiny atribútov a ktoré metódy predspracovania boli použité a ich význam uvádza Tab. 4.1 v časti 4.3.2.

Okrem úspešnosti implementovaného prediktoru, úspešnosti náhodného prediktoru a kappa štatistiky uvádzam aj orientačný ukazovateľ tvorený súčinom úspešnosti overovaného prediktora a kappa štatistiky. Tento ukazovateľ by mal umožniť jednoduchší prehľad o celkovej vhodnosti daného modelu, keďže závisí nielen od nárastu úspešnosti klasifikátora oproti náhodnému klasifikátoru, ale aj celkovej úspešnosti klasifikácie, čiže korektnosti odporúčania pre používateľa.

Tab. 5.2 Úspešnost a kappa štatistika alternatívnych modelov preferencií používateľa.

model	úspešnosť		kappa	úspešnosť overovaného prediktora * kappa
	overovaný prediktor	náhodný prediktor		
mtf	54,21%	51,88%	4,84%	2,62%
mtsf	61,14%	53,87%	15,75%	9,63%
tf	53,52%	50,89%	5,35%	2,86%
mf	56,43%	54,91%	3,37%	1,90%
mtcf	54,01%	51,73%	4,73%	2,55%
mtscf	60,99%	51,98%	18,76%	11,44%
tcf	54,37%	50,60%	7,64%	4,15%
mcf	62,73%	59,88%	7,12%	4,47%
mt	54,87%	48,20%	12,88%	7,07%
mts	59,65%	53,60%	13,05%	7,78%
t	53,53%	50,72%	5,70%	3,05%
m	55,94%	54,37%	3,46%	1,94%
tsf	78,95%	78,95%	0,00%	0,00%
msf	63,34%	57,68%	13,37%	8,47%
hlasovanie (všetky)	57,83%	56,43%	3,22%	1,86%
hlasovanie 2 (mtsf + mtscf + msf)	60,39%	54,37%	13,19%	7,97%

Na základe 14 alternatív modelu preferencií sú na konci Tab. 5.2 uvedené aj ďalšie dva modely, z ktorých prvý používa na klasifikáciu hlasovanie všetkých 14 alternatívnych modelov a druhý len tri vybraté modely. Tieto tri modely boli vybraté tak, aby dosahovali čo najlepšie výsledky na základe viacerých kritérií, čiže okrem celkovej úspešnosti a kappa štatistiky aj čo najúspešnejšiu

klasifikáciu pre obe triedy osobitne. Na tento účel bola použitá štatistika F-measure (opísaná v časti 5.1.1) počítaná osobitne pre obe triedy. Súčet F-measure pre obe triedy bol vynásobený celkovou úspešnosťou a kappa štatistikou. Výsledky sa nachádzajú v Tab. 5.3.

Tab. 5.3 Výber troch najlepších alternatívnych modelov podľa úspešností klasifikácií pre obe triedy osobitne, celkovej úspešnosti a kappa štatistiky.

model	F-measure		celková úspešnosť	kappa štatistika	obe F-measure * úspešnosť * kappa
	trieda nezaujímavý	trieda zaujímavý			
mtf	59,76%	46,88%	54,21%	4,84%	**2,80%**
mtsf	66,53%	53,67%	61,14%	15,75%	**11,58%**
tf	56,50%	50,09%	53,52%	5,35%	**3,05%**
mf	65,13%	41,34%	56,43%	3,37%	**2,02%**
mtcf	59,50%	46,81%	54,01%	4,73%	**2,72%**
mtscf	66,40%	53,52%	60,99%	18,76%	**13,72%**
tcf	57,51%	50,74%	54,37%	7,64%	**4,50%**
mcf	73,53%	37,04%	62,73%	7,12%	**4,94%**
mt	60,48%	47,41%	54,87%	12,88%	**7,62%**
mts	62,70%	56,06%	59,65%	13,05%	**9,25%**
t	56,31%	50,37%	53,53%	5,70%	**3,25%**
m	61,93%	47,72%	55,94%	3,46%	**2,12%**
tsf	88,24%	0,00%	78,95%	0,00%	**0,00%**
msf	71,79%	47,68%	63,34%	13,37%	**10,12%**

Cieľom overenia bolo tiež vyhodnotiť vplyv použitia implicitnej spätnej väzby na celkovú úspešnosť a tiež porovnať vplyv použitia príslušnosti k téme a charakteristík na základe metadát tweetov a typu obsahu v tweetoch. V Tab. 5.4 sa nachádza takýto prehľad, v ktorom sú zvýraznené hodnoty kappa štatistiky, keďže dáva smerodajný údaj o naozajstnom výkone klasifikátorov.

Ako vidieť z vyhodnotenia v časti 5.1.2, model interpretujúci spätnú väzbu síce určil explicitnú spätnú väzbu lepšie než náhodný klasifikátor, stále je jeho úspešnosť ale pomerne nízka. Z Tab. 5.4 vyplýva, že bez použitia metód predspracovania je rozdiel medzi modelom trénovaným na tweetoch s explicitnou spätnou väzbou a modelom trénovaným aj na tweetoch s predikovanou spätnou väzbou zanedbateľný pri použití len jednej z dvoch skupín atribútov a pri oboch skupinách bolo dokonca kontraproduktívne používať interpretáciu implicitnej spätnej väzby. V prípade použitia uvedených dvoch metód predspracovania boli však výsledky modelov trénovaných aj na tweetoch s odhadovanou explicitnou spätnou väzbou celkovo lepšie (s výnimkou modelu s kappa štatistikou s hodnotou 0%, kedy bolo odstránených príliš veľa podobných inštancií). Na základe týchto ďalších dvoch metód však nie je možné rozhodnúť o užitočnosti použitia implicitnej spätnej väzby, keďže neboli trénované aj modely bez implicitnej spätnej väzby s rovnakými metódami predspracovania.

V prípade skupín charakteristík boli atribúty rozdelené na typ obsahu a metadáta tweetov a ich autorov do jednej skupiny a do druhej skupiny priemerné hodnoty časov pozornosti a pomer kliknutí na URL pre tweety v príslušnej téme. Vidieť, že zaujímavosť témy pre používateľa (vyjadrená ako implicitná spätná väzba) nesie pre určenie zaujímavosti obsahu pravdepodobne viac informácie než charakteristiky opisujúce autorov, tweety a typ obsahu v nich. Je ale veľmi dôležité vyzdvihnúť fakt, že ukazovatele tém sú len dva, zatiaľ čo charakteristík v druhej skupine je 14. Trénovacích vzoriek je v prípade použitia implicitnej spätnej väzby 40, čo je príliš malý počet na to, aby mohol trénovaný SVM model dosiahnuť lepšie výsledky.

Tab. 5.4 Porovnanie úspešností a kappa štatistík podľa použitia implicitnej spätnej väzby (SV) a podľa typu použitých charakteristík.

metódy predspracovania a použitie implicitnej SV / charakteristiky		implicitná SV	implicitná SV + kompenzácia nerovnomerného počtu vzoriek pre triedy	implicitná SV + odstránenie veľmi podobných vzoriek	žiadna metóda predspracovania a bez implicitnej SV
ukazovatele tém (t)	úspešnosť	53,52%	54,37%	78,95%	53,53%
	kappa	**5,35%**	**7,64%**	**0,00%**	**5,70%**
ukazovatele typu obsahu a metadáta (m)	úspešnosť	56,43%	62,73%	63,34%	55,94%
	kappa	**3,37%**	**7,12%**	**13,37%**	**3,46%**
všetky charakteristiky (mt)	úspešnosť	54,21%	54,01%	61,14%	54,87%
	kappa	**4,84%**	**4,73%**	**15,75%**	**12,88%**

5.2 Zhodnotenie riešenia a budúca práca

Cieľom tejto diplomovej práce bolo navrhnúť metódu prieskumného vyhľadávania na sociálnej sieti Twitter, ktorá bude využívať viacero aspektov príspevkov, bude kombinovať zvolené metódy vyhľadávania a pokúsi sa dynamicky upravovať ich kritériá podľa skupín tém. Toto riešenie malo byť overené implementáciou experimentálneho prototypu.

V časti 4.1.1 bola zhrnutá definícia pojmu prieskumného vyhľadávania a vymedzená oblasť, ktorú rieši táto práca. Prieskumné vyhľadávanie je v širšom kontexte skúmanie informácií a navigácia nimi, čo umožňuje získať bližší pohľad na povahu informácií v skúmanej oblasti a cieľ toho, kto toto prieskumné vyhľadávanie uskutočňuje sa často formuje práve počas tohto procesu. Existuje mnoho možností, ako pristupovať k návrhu metódy pre prieskumné vyhľadávanie. Pre Twitter však podľa uskutočnenej analýzy súvisiacich prác bolo navrhnutých a skúmaných veľmi málo metód prieskumného vyhľadávania.

Keďže Twitter je miestom, kde rýchlo pribúda obrovské množstvo krátkych príspevkov najrôznejšej povahy, považujem za dôležité poskytnúť používateľovi pri prieskumnom

vyhľadávaní predstavu o tom, o čom ľudia píšu v súvislosti so zadaným dopytom. Používateľ potom prirodzene môže byť viac zaujatý istým typom príspevkov alebo istými témami v príspevkoch. Tweetov je vo všeobecnosti veľké množstvo a líšia sa v povahe zdieľaných informácií, preto pripisujem vyššiu dôležitosť aj podpore používateľa pri navigácií obsahom vo forme odporúčaní. V tejto práci sa preto sústredím na návrh metódy prieskumného vyhľadávania, ktorá umožňuje používateľovi získať prehľad o témach v tweetoch, ktoré sú relevantné zadanému dopytu a zároveň táto metóda používateľovi pomáha nestratiť sa v množstve pre neho nezaujímavého obsahu a naopak podporuje používateľa v objavovaní príspevkov, ktoré ho zaujímajú.

Analyzovaných bolo viacero prác, ktoré sa zaoberali metadátami tweetov či už v kontexte jednotlivých tweetov alebo ich autorov. Existuje veľké množstvo prác, ktoré sledujú tieto charakteristiky a používajú ich ako atribúty pre účel klasifikácie tweetov podľa toho, aké tweety sa používateľom páčia a aké nie. Bolo definovaných viacero ukazovateľov poskytovaných priamo Twitterom a odvodených, vypočítaných charakteristík, ktoré sa ukázali ako vhodné pre takúto klasifikáciu. Vo všeobecnosti tieto charakteristiky vyjadrujú rôzne aspekty tweetov, často ide o popularitu tweetov na Twitteri alebo odvodenú popularitu autorov tweetov na Twitteri. Skúmané sú tiež ukazovatele ako využitie spomenutí používateľa, čo sa dá interpretovať ako odpoveď na iný tweet alebo ako časť konverzácie medzi používateľmi Twitteru. Tiež je tu možnosť použiť hashtagy ako istý spôsob vyjadrenia témy príspevku. Analyzované bolo aj emocionálne zafarbenie mikroblogov podľa použitých slov či emotikon a mnoho ďalších aspektov príspevkov. V práci bolo vybratých niekoľko overených charakteristík tak, aby bolo možné charakterizovať obsah a jeho typ z čo najväčšieho množstva hľadísk, keďže aj motivácia používateľa a jeho záujem môžu byť ovplyvňované množstvom rôznych aspektov.

Bolo tiež navrhnutých niekoľko nových prístupov a charakteristík, aby bol dosiahnutý spomenutý účel. K tomu patrila vlastná extrakcia emotikon, ktorých emocionálne zafarbenie bolo definované prostredníctvom existujúcich prístupov, čiže v tomto prípade boli použité hodnoty zo slovníkov s uvedeným emocionálnym zafarbením pre slová vystihujúce emóciu zachytenú emotikonom. Analyzované výskumy často vyhodnotili výskyt URL v tweete ako veľmi silnú charakteristiku pre určenie zaujímavosti tweetu, preto bol navrhnutý aj ukazovateľ, ktorý kategorizoval typ odkazovaného obsahu medzi triedy ako obrázok, video či audio. Bolo tak urobené na základe predpokladu, že aj typ zdieľaného obsahu odkazovaného prostredníctvom URL môže ovplyvňovať zaujímavosť príspevku pre používateľa.

Oproti väčšine analyzovaných existujúcich prístupov je riešenie navrhnuté v tejto práci zamerané výhradne na aktuálne sedenie používateľa, čiže neberie do úvahy žiadne ďalšie informácie o používateľovi (ako napríklad jeho dlhodobé záujmy, históriu vyhľadávania, aktivitu na Twitteri a pod.). Preto bolo nevyhnutné vybrať a navrhnúť také charakteristiky, ktoré nemusia byť vypočítané vopred čiže môžu byť vypočítané počas práce používateľa s aplikáciou

implementujúcou navrhované riešenie. Toto vymedzenie ovplyvňovalo prakticky všetky aspekty návrhu.

Boli skúmané aj práce, ktoré sa snažia kompenzovať relatívne malé množstvo informácie vyskytujúcej sa v jednom tweete. Práve pre spomínané obmedzenie výpočtov výlučne na jedno sedenie bol však navrhnutý spôsob obohacujúci text tweetov o názov odkazovanej stránky.

Pre zostavenie modelu preferencií používateľa čiže modelu opisujúceho, o čom by mali byť tweety a aké by mali byť, aby boli zaujímavé pre používateľa je nevyhnutné získať vyjadrenie jeho názoru na prečítané príspevky. Získavanie explicitnej spätnej väzby nie je jednoduché, či už z pohľadu presnosti alebo ochoty používateľa poskytnúť ju. Tejto problematike boli venované predovšetkým časti 3.4 a 3.6, ktoré zahŕňali experiment s rôznymi typmi používateľského rozhrania z pohľadu spôsobu prezentovania obsahu a spôsobu poskytovania explicitnej spätnej väzby. Ukázalo sa, že používatelia zúčastnení v experimente nemali na žiadnu z troch diametriálne odlišných spôsob prezentovania tweetov a poskytovania spätnej väzby výrazne horší či lepší názor. Experiment zároveň využíval odlišný prístup k prieskumnému vyhľadávaniu než finálne riešenie. Používatelia totiž k zadanému dopytu videli vždy len niekoľko tweetov a aby im boli zobrazené ďalšie, museli existujúce tweety ohodnotiť z pohľadu zaujímavosti. V pozitívne ohodnotených tweetoch boli prostredníctvom upravenej TF-IDF charakteristiky (upravenej pre použitie vo veľmi krátkych dokumentoch) vyhľadané slová, ktoré sa použili na rozšírenie pôvodného dopytu, čím sa ďalšie zobrazené tweety už týkali aj ďalších podtém zaujímavých pre používateľa. Výsledok tohto experimentu ukázal, že dôraz je potrebné klásť na väčšie množstvo charakteristík opisujúcich tweety a že konkrétny spôsob poskytovania explicitnej spätnej väzby pravdepodobne nebude mať výraznejší dopad na dosahované výsledky. Preto bol vybratý zaužívaný spôsob hodnotenia piatimi hviezdičkami a všetky vyššie spomenuté charakteristiky tweetov.

Aby bola splnená požiadavka zadania diplomovej práce na sumarizáciu príspevkov podľa tém a zároveň tak aj dôležitý aspekt prieskumného vyhľadávania na poskytovanie prehľadu o existujúcich informáciách k danému dopytu, bol použitý existujúci prístup tzv. tematickej sumarizácie, ktorá bola validovaná aj v praxi. Výsledky dopytu z Twitteru sú zaradené do tém, ktoré tvoria frázy (n-gramy) vyskytujúce sa v tweetoch. Používateľ si môže vybrať, ktoré takto vytvorené témy bude čítať, ak nechce čítať všetky príspevky relevantné dopytu. Témy čiže frázy sú zoradené podľa frekvencie výskytu a zároveň sú zvýhodňované také slovné spojenia, ktoré sa bežne v tweetoch nevyskytujú, čiže majú predpoklad čo najlepšie vystihnúť témy. Tento prevzatý prístup zároveň umožňuje opísať tweet práve z obsahového hľadiska podľa príslušnosti k témam, čo tvorí doplnok k charakteristikám, ktoré opisujú skôr štýl príspevku a spôsob a typ zdieľanej informácie.

Pre posilnenie vhodnosti navrhovanej metódy pre použitie v praxi bola zvážená aj obťažnosť získavania explicitnej spätnej väzby. Používateľa totiž zaťažuje a v prípade tweetov by v ideálnom prípade pre model preferencií poskytol používateľ explicitnú spätnú väzbu za

prečítaním každých približne 140 znakov. Nie je preto možné spoliehať sa na to, že explicitnej spätnej väzby bude dostatok alebo si ju je možné od používateľa vynucovať. Analyzované preto boli aj metódy získavania a interpretácie implicitnej spätnej väzby, aby bolo možné kompenzovať explicitnú spätnú väzbu tam, kde chýba. Pre overenie konceptu využívania implicitnej spätnej väzby pri prieskumnom vyhľadávaní na Twitteri boli použité tri základné charakteristiky: čas pozornosti pre text tweetu, kliknutie na URL v tweete a čas pozornosti pre tento externý obsah odkazovaný cez URL.

Práve čas pozornosti pre text tweetu a podiel kliknutí na URL v tweetoch slúžili tiež ako charakteristiky súvisiace s príslušnosťou tweetov k témam. Pre každý tweet boli vypočítané priemery oboch hodnôt na jeden tweet pre témy, ku ktorým daný tweet prislúchal.

Výsledkom je navrhnutá a implementovaná metóda, ktorá pre zadaný dopyt používateľom načíta tweety, pripojí k textu tweetu aj názov odkazovanej stránky cez URL, ak sa v tweete nachádza, jednoduchým spôsobom odstráni tweety s veľmi zhodujúcim sa textom, použije prevzatý algoritmus pre sumarizáciu tweetov do tém a tie zobrazí používateľovi. Používateľovi sa po kliknutí na tému zobrazia príslušné tweety, ktoré môže čítať a v prípade záujmu tiež kliknúť na URL a prezerať aj externý obsah. Počas toho je zaznamenávaná implicitná spätná väzba a používateľ má tiež možnosť ohodnotiť tweety podľa zaujímavosti. Na základe toho sa vytvárajú dva modely metódou strojového učenia SVM (Support Vector Machines). Prvým je model interpretujúci implicitnú spätnú väzbu vo forme explicitnej spätnej väzby a druhým model preferencií používateľa, ktorý opisuje tweety celkovo 16 atribútmi. Trénovaciu množinu modelu preferencií tvoria ohodnotené tweety používateľom a aj tweety, ktorým je hodnotenie (čiže explicitná spätná väzba) predikované práve modelom interpretujúcim implicitnú spätnú väzbu, ak dosiahne istú úspešnosť v krížovej validácií. Neprečítaným tweetom je tak odhadované, či sú pre používateľa zaujímavé alebo nezaujímavé a toto hodnotenie je naznačené používateľovi pre tweety osobitne a aj pre témy práve podľa toho, či obsahujú prevažne odporúčané alebo neodporúčané tweety.

Finálne riešenie spĺňa prakticky všetky povinné požiadavky zo špecifikácie požiadaviek z časti 3.1. Praktické overenie riešenia bolo uskutočnené tak, že predikované odporúčanie (klasifikácia do tried zaujímavý a nezaujímavý) nebolo používateľom zobrazované počas čítania tweetov a po prečítaní 40 tweetov bola od nich vyžiadaná spätná väzba na tweety. Pre model interpretujúci implicitnú spätnú väzbu tak bolo sledované ohodnotenie používateľom pre tweety, ktoré čítal a neohodnotil a pre model preferencií ohodnotenie tweetov, ktoré čítal po prvý krát. Model preferencií bol zostavovaný v 16 alternatívach, aby bolo možné porovnať osobitne prínos charakteristík typu obsahu a charakteristík príslušnosti k témam (čiže samotného obsahu), osobitne aj užitočnosť použitia modelu interpretujúceho spätnú väzbu na obohatenie trénovacej množiny a aj prínos dvoch metód predspracovania. Výsledky sú sledované predovšetkým podľa celkovej úspešnosti modelov (klasifikátorov) čiže podielu správne klasifikovaných tweetov voči všetkým

klasifikovaným tweetom a podľa kappa štatistiky, ktorá uvádza úspešnosť v miere zlepšenia sa oproti náhodnému klasifikátoru.

Sledovanie implicitnej spätnej väzby umožnilo správne odhadnúť zaujímavosť pre používateľa pri približne 68% tweetov oproti približne 59%, ktoré by odhadol náhodný klasifikátor používajúci rovnaký počet zaujímavých a nezaujímavých tweetov. Sledované však boli len tri ukazovatele pre implicitnú spätnú väzbu súvisiace s čítaním textu tweetu a odkazovaného obsahu. Z výsledkov je vidieť, že sledovanie a interpretovanie implicitnej spätnej väzby má prínos aj pri dokumentoch, akým sú tweety. Z dôvodu obsahového a časového vymedzenia riešenia však nebol skúmaný vplyv aktivít ako vykonanie tzv. retweetu (znovu uverejnenia konkrétneho tweetu používateľom) alebo zobrazenie ďalších tweetov autora čítaného tweetu, ktoré majú svojou povahou dobré predpoklady na ukazovatele implicitnej spätnej väzby. Domnievam sa, že práve umožnenie a sledovanie týchto úkonov by výrazne zlepšilo model interpretujúci spätnú väzbu.

V prípade troch najlepších alternatív modelu preferencií používateľa a alternatívy pozostávajúcej z hlasovania vybratých troch alternatív bola dosiahnutá úspešnosť klasifikácie nad 60%. Kappa štatistika dosahovala v týchto prípadoch približne od 13% do 20%, kde 0% znamená úplnú zhodu s náhodným klasifikátorom a 100% úplnú zhodu s perfektným klasifikátorom. Tieto výsledky síce neznamenajú dosiahnutie výrazne dobrej úspešnosti, no všetky alternatívy modelov preferencií dosiahli lepšie výsledky než náhodné klasifikátory. Algoritmus strojového učenia SVM bol v každom sedení pri overení trénovaný len na 40 vzorkách v trénovacej množine pre klasifikáciu na základe 16 atribútov. Takáto množina je veľmi malá pre priestor toľkých rozmerov, no vždy bolo dosiahnuté zlepšenie v úspešnosti, z čoho usudzujem, že navrhovaný prístup k prieskumnému vyhľadávaniu má predpoklady pre použitie v praxi a je vhodný na ďalší výskum a zlepšovanie v rámci budúcej práce.

Finálne riešenie je v čase písania práce dostupné online na adrese `http://dp.zilincik.eu/` a v prípade záujmu pristúpiť na systém v režime overenia, je možné použiť adresu `http://dp.zilincik.eu/?page=exploreTest`. Rozdiel je však len v nezobrazovaní odporúčaní od prvého modelu, ale až po vyžiadaní spätnej väzby, kedy je potrebné ohodnotiť maximálne 60 ďalších tweetov. Výstupy z testov je potrebné ďalej spracovať, takže systém v tomto režime nezobrazuje dosiahnuté miery, ktoré boli využité na overenie riešenia.

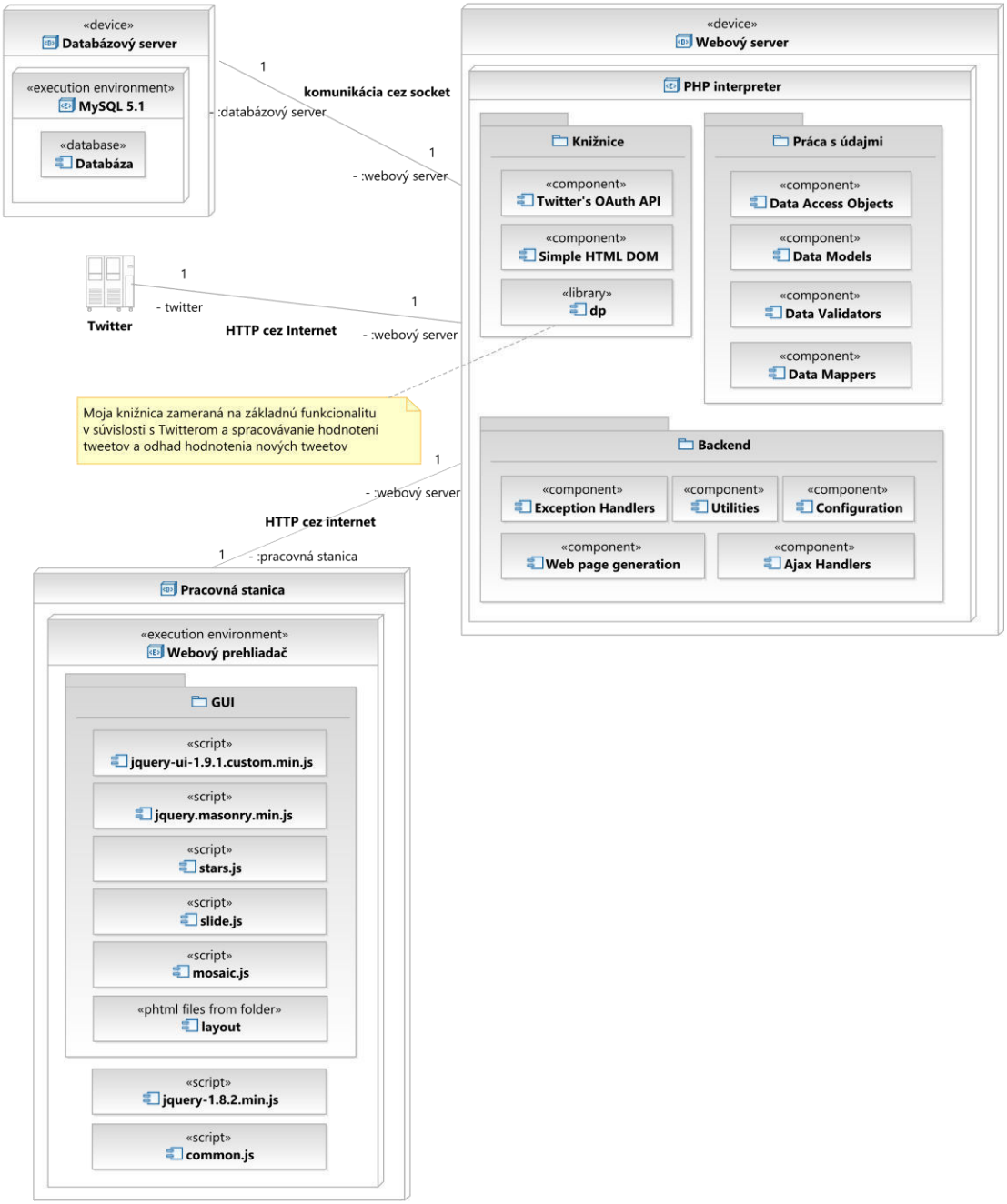
Pre budúcu prácu založenú na riešení predostretom v tejto diplomovej práci vidím základ v charakteristikách pre implicitnú spätnú väzbu, ktorá je vo všeobecnosti veľmi dôležitá pre praktické použitie vzhľadom na náročnosť a vysokú cenu získavania explicitnej spätnej väzby. Ako bolo spomenuté, bolo by vhodné skúmať vplyv akcií súvisiacich s Twitterom (vykonanie retweetu, zobrazenie profilu autora, zobrazenie ďalších tweetov autora a pod.) ako ukazovateľov implicitnej spätnej väzby. Z výsledkov overenia pre model preferencií používateľa je tiež vidieť vplyv metód predspracovania údajov použitých pre algoritmus strojového učenia. Použité boli dve základné metódy: duplikovanie vzoriek v trénovacej množine tak, aby obe sledované triedy boli

reprezentované rovnakým počtom inštancií a tiež odstraňovanie príliš podobných inštancií, ktoré patria do rôznych tried. Samozrejme, metódy predspracovania súvisia aj s vybratým algoritmom strojového učenia. V práci bol pre implementáciu experimentálneho prototypu použitý jeden algoritmus, preto je tiež vhodné skúmať aj vplyv výberu algoritmu na celkovú úspešnosť modelu. Na zvýšenie úspešnosti môžu mať vplyv aj použité charakteristiky, preto sa otvára priestor aj na použitie nových a zlepšenie tých charakteristík, ktoré boli v tejto práce navrhnuté.

6 Technická dokumentácia

6.1 Dokumentácia k návrhu základného a testovacieho systému

6.1.1 Nasadenie a štruktúra systému



Obr. 6.1 Diagram nasadenia pre testovací systém.

Na Obr. 6.1 sa nachádza diagram nasadenia (z angl. Deployment Diagram) pre navrhovaný systém. Zachytáva najdôležitejšie časti systému ako databáza, ktorá je obsluhovaná serverom MySQL vo verzii 5.1 na databázovom serveri, grafické používateľské rozhranie a pridruženú funkcionalitu

prostredníctvom HTML, CSS a JavaScript-u interpretovanom webovým prehliadačom používateľa a zvyšnú časť systému v jazyku PHP, ktorý je obsluhovaný webovým/aplikačným serverom. Je využitá knižnica jQuery[2] pre jazyk JavaScript spolu s doplnkom jQuery UI[3] pre rozšírenie možnosti práce v rámci grafického používateľského rozhrania.

Táto serverová časť riešenia je v diagrame rozvrhnutá do troch balíkov: knižnice, práca s údajmi a backend. V testovacom systéme sa nachádzajú dve prevzaté knižnice alebo moduly, ide o PHP knižnicu pre Twitter's OAuth API[4] od Abrahama Williamsa a Simple HTML DOM modul[5]. Rovnakým spôsobom je vyčlenená aj ďalšia funkcionalita v knižnici s pracovným názvom „dp". Ide o časť, ktorú som naprogramoval pre podporu hodnotenia tweetov a spracovávania tohto hodnotenia spolu s odhadom zaujímavosti pre používateľa. Obsahuje tiež aj všetko potrebné pre načítavanie tweetov z Twitter API, ich predspracovanie a uloženie do databázy.

Práca s údajmi je časť skladajúca sa z modelov údajov, abstrakčných objektov, mapovacích tried a prípadne validátorov údajov. Zabezpečujú používanie údajov v systéme s čo najvyššou možnou mierou abstrakcie ich fyzického umiestnenia a spôsobu uloženia.

Backend časť obsahuje ďalšiu funkcionalitu zabezpečujúci chod systému. Ide napríklad o základný spôsob vyberania a spúšťania skriptov a spôsob zakomponovávania ich výstupu do stanovených šablón pre zobrazenie alebo o asynchrónnu komunikáciu s prehliadačom používateľa. Nastavenie systému, ošetrovanie chýb či pomocná funkcionalita je tiež sústredená v tejto časti.

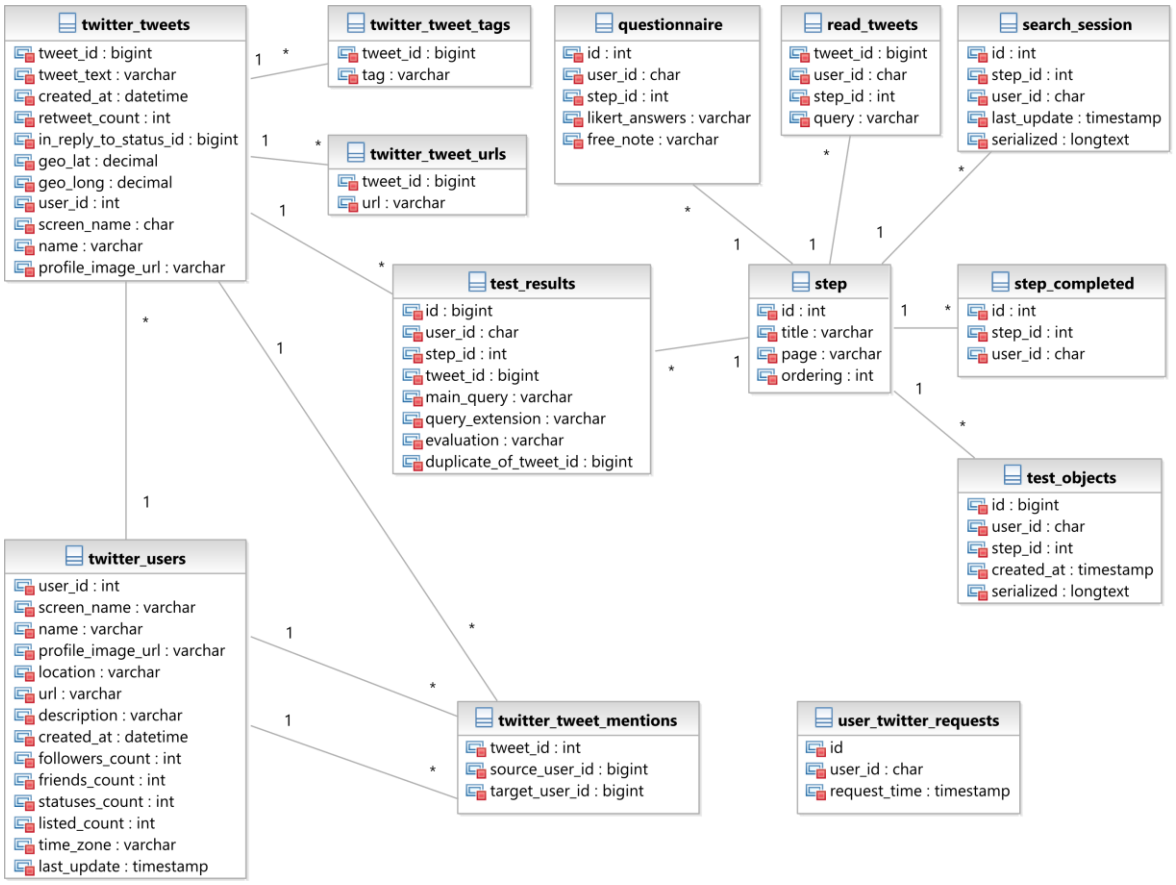
6.1.2 Fyzická štruktúra údajov

Diagram z Obr. 6.2 zobrazuje fyzickú štruktúru údajov v databáze testovacieho systému. Táto štruktúra je z veľkej časti použiteľná aj pre finálne riešenie. Základ tvoria entity tweetov ako tabuľka `twitter_tweets`, ktorá združuje obsah i metadáta tweetu. Pre optimalizáciu do budúcna duplikuje aj niektoré atribúty entity `twitter_users`, aby sa znížil počet nutných vyhľadávaní príslušných záznamov v tabuľke `twitter_users` pri bežnom zobrazení tweetu so základným informáciami o autorovi tweetu. Napríklad hashtagy alebo odkazy sa nachádzajú priamo v tele tweetu, ale pri ich analýze vo finálnom riešení bude efektívnejšie mať ich extrahované. Na to slúžia tabuľky `twitter_tweet_tags` a `twitter_tweet_urls`. Tabuľka `twitter_tweet_mentions` obsahuje informácie o tom, v ktorom tweete spomína ktorý autor akého iného používateľa Twitteru. Tento vzťah je tiež možné extrahovať z textu tweetu a informácii o autorovi, ale z rovnakého dôvodu ako vyššie je aj táto informácia uložená osobitne pre efektívnejšiu prácu aplikácie.

Pre účely kontroly limitu Twitter Search API sa zaznamenávajú časy požiadaviek na toto API pre jednotlivých používateľov resp. sedenia do tabuľky `user_twitter_requests`.

[2] `http://jquery.com/`
[3] `http://jqueryui.com/`
[4] `https://github.com/abraham/twitteroauth/tree/master/twitteroauth`
[5] `http://simplehtmldom.sourceforge.net/`

Jednotlivé časti resp. kroky testu sú reprezentované entitou `step`. K nej sú potom priradené aj záznamy z tabuliek `step_completed`, ktoré hovoria o tom, ktorý krok testu už bol vykonaný používateľom. Aby boli zachované všetky potrebné údaje o aktuálnom stave počas sedenia, ukladá sa rovnako ku kroku testu a identifikácii používateľa do `search_session` aj serializovaný objekt zodpovedný za hlavnú funkcionalitu. Pre zabezpečenie nezobrazovania duplikátov sa v tabuľke `read_tweets` uchováva informácia o tom, ktoré tweety, v akom kroku a ktorý používateľ už čítal. Ten istý tweet však môže byť dobré zobraziť, ak bol predtým už čítaný, ale predstavoval výsledok pre iný dopyt, preto táto tabuľka obsahuje aj znenie dopytu.



Obr. 6.2 Entitno-relačný diagram zachytávajúci fyzickú štruktúru údajov v testovacom systéme.

Aby bolo možné využiť údaje z testovacích sedení aj neskôr, do tabuľky `test_results` sú ukladané aj údaje ako identifikácia prečítaného tweetu, hodnotení od používateľa, znenie originálneho dopytu ako i rozšírenie dopytu, pre ktorý bol tweet načítaný. `Test_objects` pre prípadné potreby v budúcnosti uchováva serializovanú inštanciu objektu zodpovedného za hlavnú funkcionalitu. Údaje z vyplneného dotazníka po každom kroku sa nachádzajú v tabuľke `questionnaire`.

Je dôležité poukázať na fakt, že v uvedenom diagrame sa nenachádza tabuľka pre informácie o používateľovi. Pôvodná implementácia používala takúto tabuľku, avšak pre dôvody

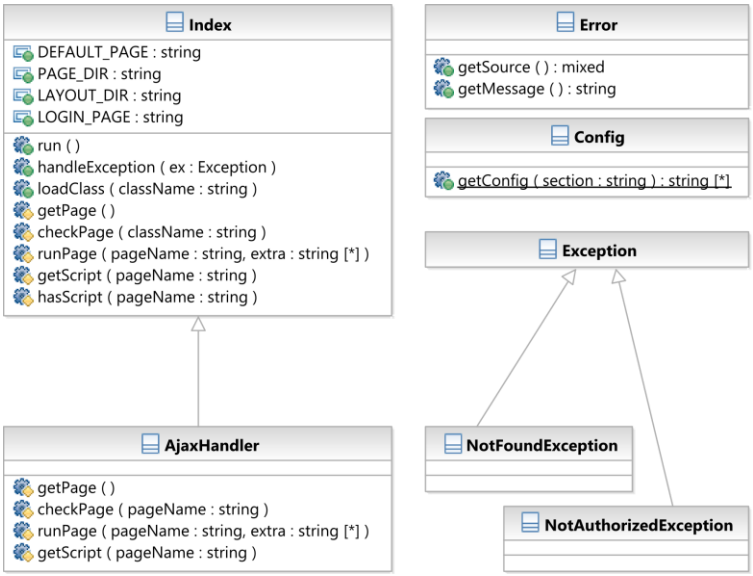
uvedené v časti 3.6.3 bolo nutné prispôsobiť aj fyzickú štruktúru údajov. Preto je atribút `user_id` reprezentovaný reťazcom, ktorý predstavuje identifikáciu sedenia, nie používateľa. Tým sa stráca prepojenie rôznych sedení toho istého používateľa, čo však nie je v rozpore so špecifikáciou požiadaviek.

6.1.3 Podrobná štruktúra systému

Táto časť opisuje podrobnejšiu štruktúru systému, konkrétne jeho časti umiestnenej na webovom serveri (pozri 6.1.1).

K triedam sú uvedené len také atribúty a metódy, ktoré sú viditeľné bez obmedzenia (public) alebo pre dediace triedy (protected). Niektoré triedy nemajú uvedené žiadne metódy či atribúty, ak si to daný kontext nevyžaduje. Triedy, ktoré sú špecifikáciou iných tried majú uvedené tie metódy, ktorých implementáciu menia. Pri typoch vstupných alebo návratových hodnôt je uvedené „`array [*]`“, pokiaľ ide o dvoj- alebo viacrozmerné pole, označenie typu „`mixed`“ pojednáva o možnosti použitia rôznych typov údajov.

Základná časť systému pozostáva z triedy `Index` zodpovednej za výber skriptu, ktorý sa vykoná, za jeho spustenie a za umiestnenie jeho výstupu do správnej šablóny. Samozrejme aj z pohľadu na obmedzenia napríklad podľa povolení spustiť či nespustiť skript. Rozšírením tejto triedy je trieda `AjaxHandler`, ktorá pracuje podobne pre požiadavky vykonané asynchrónne z prehliadača používateľa a upravuje metódy, ktoré sú naznačené na diagrame na Obr. 6.3. Použité sú ďalej tieto triedy: `Error` pre spracovanie chýb, `Config` načítava základné konfiguračné údaje z konfiguračného súboru a `NotFoundException` a `NotAuthorizedException` pre spracovávanie výnimiek.



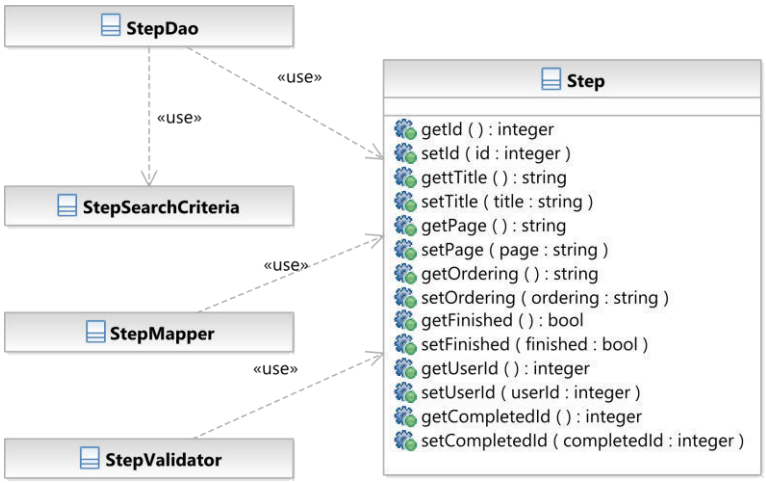
Obr. 6.3 Diagram tried pre základnú časť systému.

Pomocné triedy poskytujúce ďalšiu všeobecnú funkcionalitu sú znázornené na diagrame na Obr. 6.4. Trieda `Authentication` slúžila pred zmenami spomínanými v časti 3.6.3 na registrovanie, prihlasovanie a kontrolu autentifikácie, jej funkcionalita však bola v danom testovacom systéme pozmenená tak, aby udržiavala potrebné informácie pre vytvorenie a identifikáciu sedenia. `Utils` poskytuje rôzne funkcie zjednotené pre celý systém. Napríklad vytváranie odkazov, presmerovania rôznym spôsobom či získanie objektu pre prácu s databázovým systémom.

Fungovanie konceptu testovacieho systému a jeho priebeh pomáhajú zabezpečovať triedy `Step`, `StepDao`, `StepSearchCriteria`, `StepMapper` a `StepValidator`, ktoré sú znázornené diagramom na Obr. 6.5. Trieda `Step` je model pre entitu jedného kroku testu. `StepDao` slúži na prístup k uloženým krokom a realizáciu zmien v nich. Vyhľadávanie podľa jednotných kritérií je umožnené prostredníctvom `StepSearchCriteria` triedy. `StepMapper` slúži na naplnenie objektu `Step` údajmi, ktoré boli načítané z databázy prostredníctvom `StepDao`. `StepValidator` slúži na overenie korektnosti obsahu objektu `Step` pred jeho uložením do databázy.

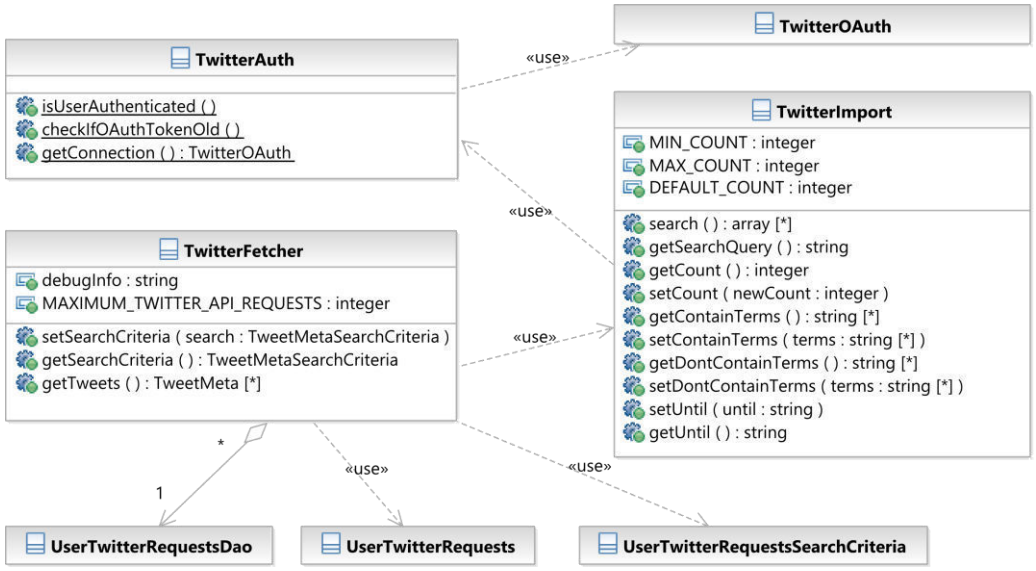
Uvedený prístup k abstrakcii údajov je v malých obmenách použitý v celom systéme, preto funkcie jednotlivých typov tried pre prácu údajov nie sú ďalej takto podrobne vysvetľované.

Obr. 6.4 Diagram tried pre pomocnú funkcionalitu základnej časti systému.



Obr. 6.5 Diagram tried pre testovaciu časť systému.

Načítavanie tweetov pre použitie v systéme je realizované prostredníctvom triedy `TwitterFetcher`. Táto načítava tweety podľa zadaných kritérií prostredníctvom triedy `MetaSearchCriteria` z databázy. Pokiaľ sa v databáze nenachádza dostatočný počet tweetov, ktoré vyhovujú daným kritériám, sú tweety načítané z Twitter API prostredníctvom triedy `TwitterImport`. Tá pre svoje fungovanie používa `TwitterAuth` triedu, ktorá zjednodušuje prístup k `TwitterOAuth` z prevzatej knižnice[6]. `TwitterFetcher` nemusí tweety načítať z Twitter API, ak zistí, že používateľ je blízko k vyčerpaniu limitu počtu požiadaviek. Ich zaznamenávanie a kontrolu limitu zabezpečujú triedy `UserTwitterRequests`, `UserTwitterRequestsDao` a `UserTwitterRequestSearchCriteria`. Tieto vzťahy je možné vidieť aj vo forme diagramu na Obr. 6.6.

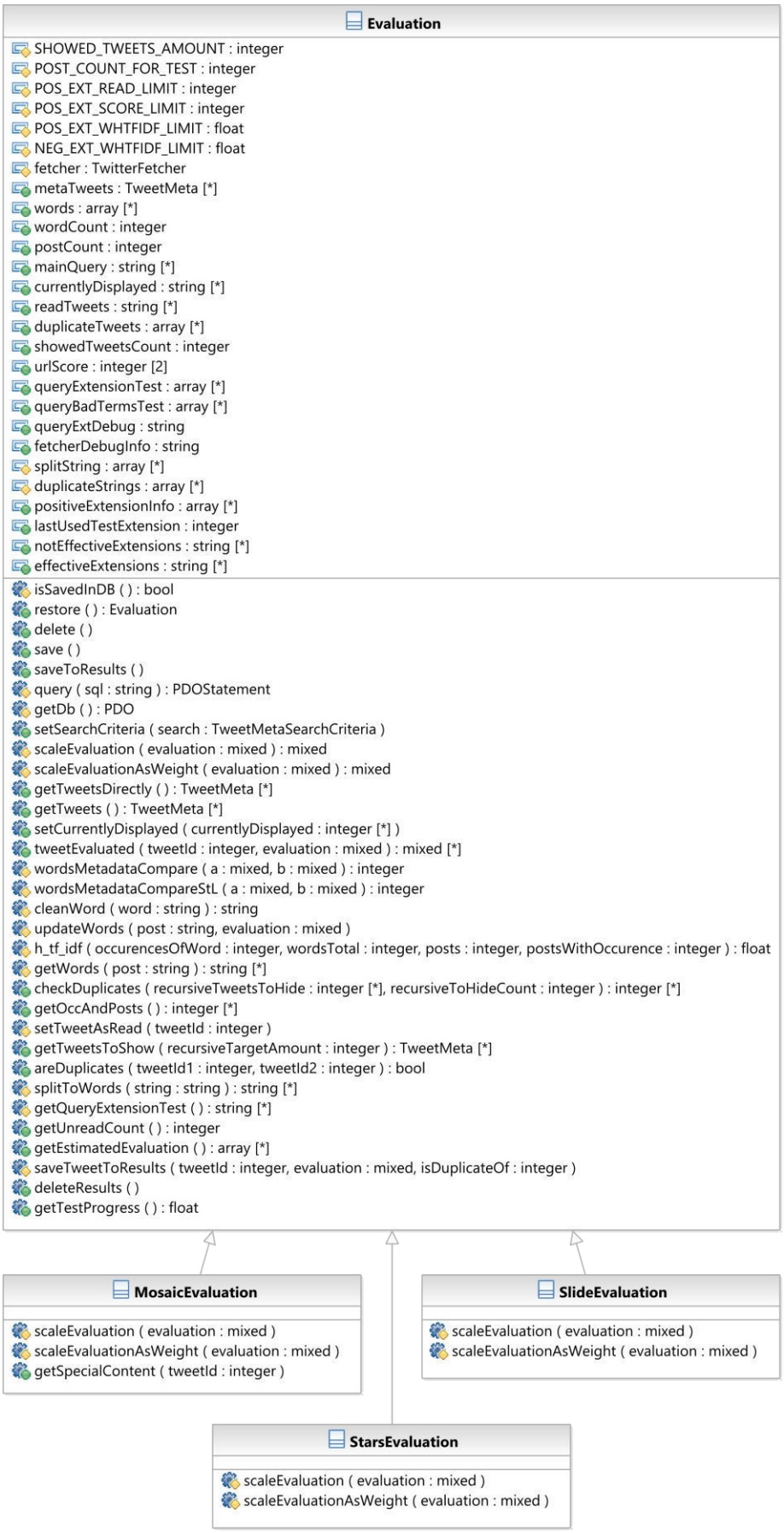


Obr. 6.6 Diagram tried pre funkcionalitu komunikácie s Twitter API.

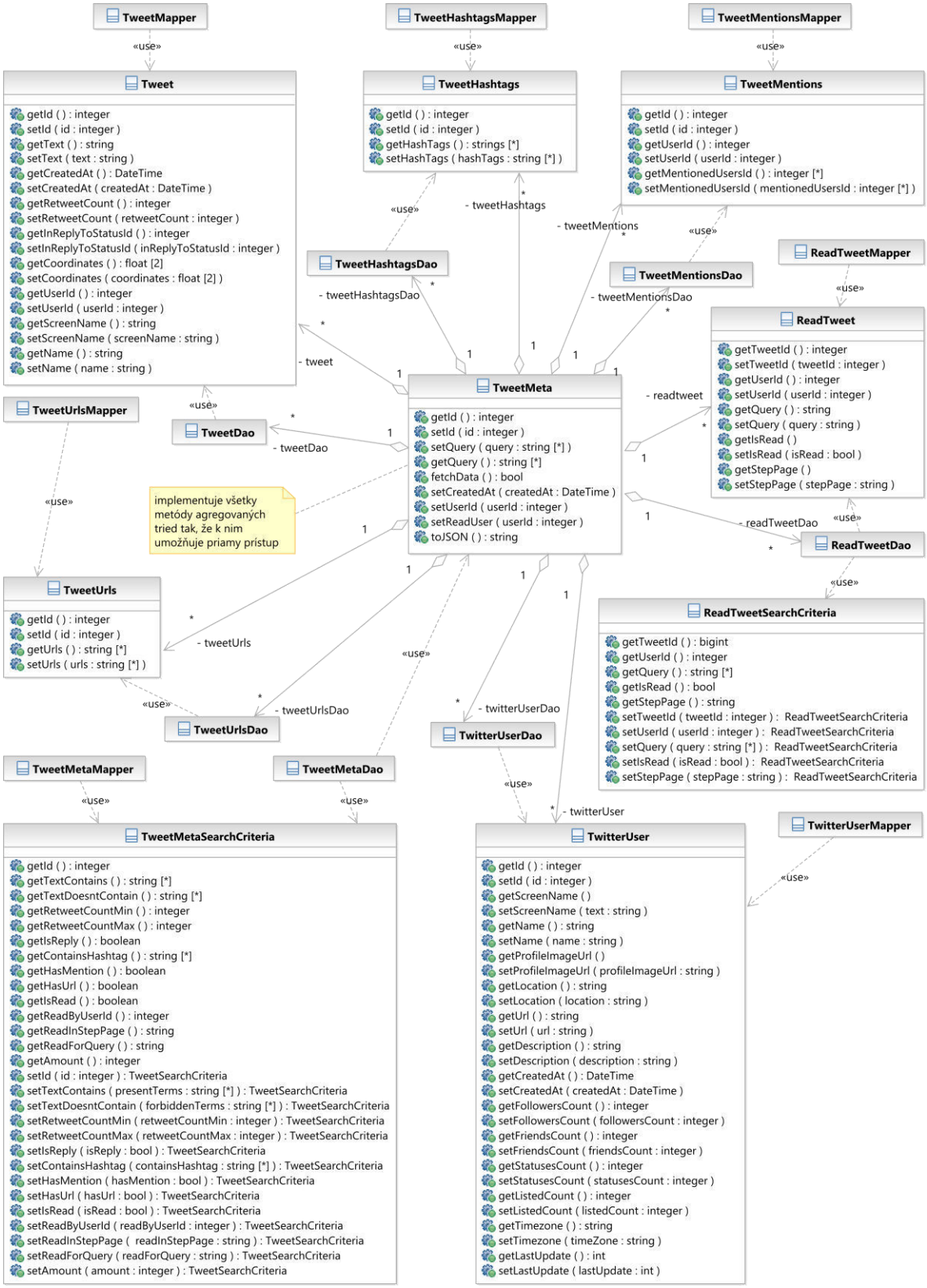
[6] `https://github.com/abraham/twitteroauth/tree/master/twitteroauth`

Hlavná funkcionalita je implementovaná triedou `Evaluation` respektíve odvodenými triedami `StarsEvaluation`, `SlideEvaluation` a `MosaicEvaluation`. Hlavnou funkcionalitou sa myslí prinášanie výsledkov v podobe tweetov pre dopyt používateľa, spracovávanie používateľského hodnotenia, analýza tweetov, snaha o odhad preferencií používateľa a modifikácia kritérií pre načítavanie nových tweetov, ktoré sa zobrazia používateľovi. Pomerne veľká trieda `Evaluation` spolu s odvodenými triedami sa nachádza na Obr. 6.7. Vysoký počet viditeľných atribútov a metód je spôsobený vysokou náročnosťou testovania a analýzy stavu inštancií týchto tried.

Na Obr. 6.8 sa nachádza diagram, ktorý znázorňuje triedy, ktoré súvisia s prácou s tweetmi, ich metadátami, informáciami o ich autoroch a s inými entitami ako hashtagy a odkazy. S tweetom súvisí aj informácia o tom, či už bol čítaný v danom sedení pre uvedený dopyt. Významnou triedou tu však je `TweetMeta`, ktorá agreguje triedy `Tweet`, `TweetDao`, `TweetUrls`, `TweetUrlsDao`, `TweetHashtags`, `TweetHashtagsDao`, `TweetMentions`, `TweetMentionsDao`, `ReadTweet`, `ReadTweetDao`, `TwitterUser` a `TwitterUserDao`. Reprezentuje tak všetky informácie súvisiace s tweetom, ktoré sú uložené v databáze. Trieda združujúca kritéria vyhľadávania tweetov `TweetMetaSearchCriteria` obsahuje široký zoznam týchto kritérií a je určená na vyhľadávanie, ktoré zahŕňa podmienky pre rôzne entity súvisiace s tweetom. `TweetMeta` implementuje metódy agregovaných tried tak, že ich priamo volá s výnimkou tých metód, ktoré sú znázornené v diagrame (tie sú obsluhované modifikovaným spôsobom).



Obr. 6.7 Diagram tried pre hlavnú funkcionalitu systému vrátane hodnotenia tweetov a odhad hodnotenia systémom.



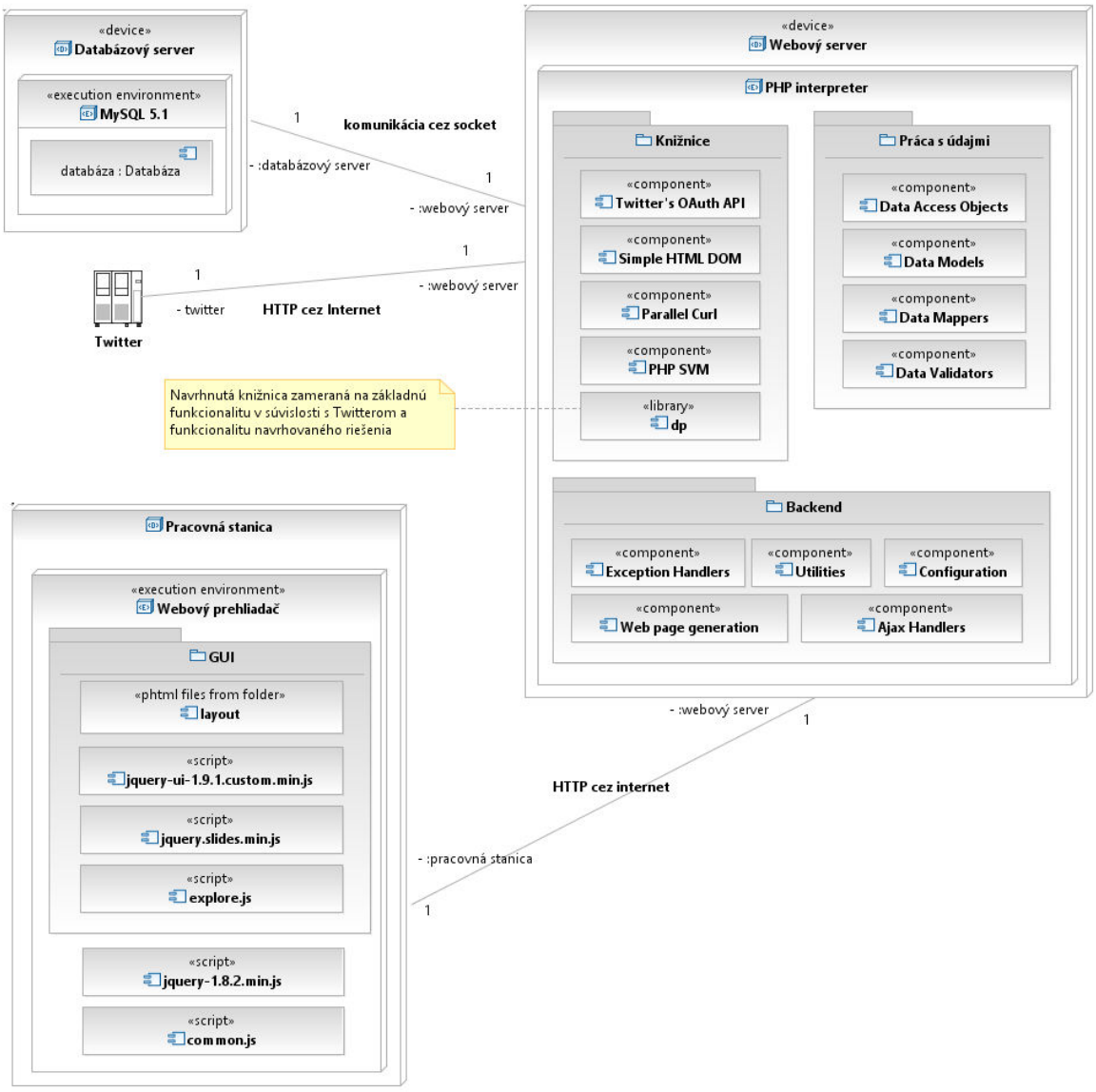
Obr. 6.8 Diagram tried určených pre prácu s tweetmi, ich metadátami a autormi.

6.2 Revízia dokumentácie podľa finálneho systému

V tejto často sa nachádza doplnenie technickej dokumentácie z časti 6.1 a aktualizované časti tejto dokumentácie, ktoré súvisia so zmenami, ktoré nastali pri implementácii finálneho riešenia oproti implementácii základného a testovacieho systému.

6.2.1 Nasadenie a štruktúra systému

Na Obr. 6.9 sa nachádza diagram nasadenia, ktorý zachytáva štruktúru systému, ktorý je implementáciou finálneho riešenia. Oproti pôvodnej štruktúre z časti 6.1.1 sa zmenilo zostavenie klientskej časti aplikácie vo webovom prehliadači, kedy hlavnú funkcionalitu obsluhuje skript `explore.js`. Pre zobrazenie návodu na prácu so systémom po zadaní dopytu používateľom sa využíva prevzatý modul `jquery.slides.min.js`[7].



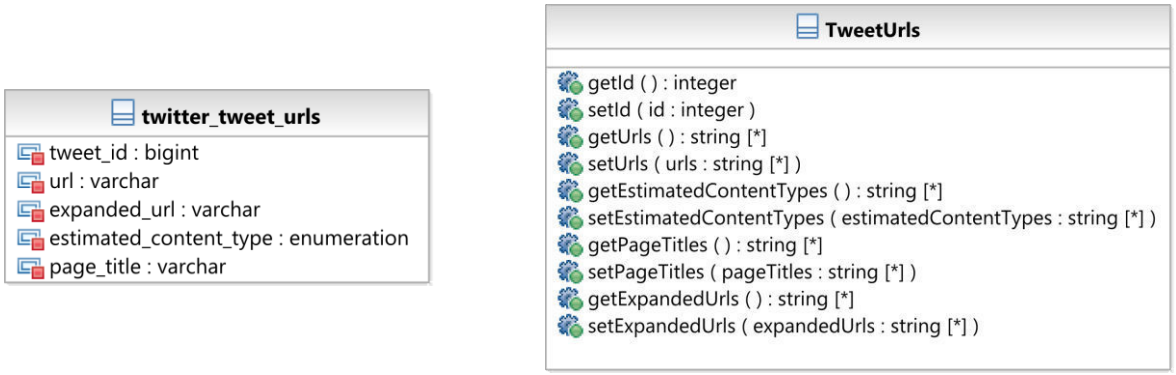
Obr. 6.9 Diagram nasadenia pre finálne riešenie.

[7] `http://slidesjs.com/`

Na serverovej strane riešenia pribudli dve prevzaté knižnice. Parallel Curl[8] od autora menom Pete Warden je určené pre vytváranie paralelných http požiadaviek, čo je využité pri práci s URL v tweetoch (zistenie originálnej podoby URL a načítanie obsahu odkazovanej web stránky). PHP SVM[9] vytvoril Ian Barber ako implementáciu SVM algoritmu pre jazyk PHP. Toto riešenie sa používa pre vytvorenie a používanie modelu interpretujúceho implicitnú spätnú väzbu a modelu preferencií používateľa.

6.2.2 Fyzická štruktúra údajov a podrobná štruktúra systému

Pre rozšírenie systému o odhad typu obsahu referencovaného prostredníctvom URL v tweete a extrakciu textov z HTML `title` tágu stránok došlo k zmene štruktúry ukladaných údajov v databáze. Na Obr. 6.10 sa nachádza nová štruktúra databázovej entity nesúcej informáciu o URL z tweetov spolu s aktualizovanou triedou abstrakcie údajov pre prácu s týmito dátami v prostredí PHP.



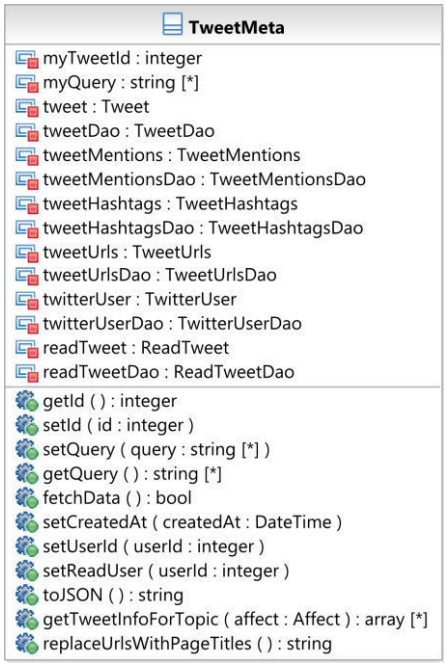
Obr. 6.10 Štruktúra databázovej entity a triedy pre abstrakciu informácií týkajúcich sa URL v tweetoch.

Pre prácu s tweetmi slúži entita TweetMeta, ktorej pri finálnom riešení boli pridané dve metódy. Metóda s názvom getTweetInfoForTopic poskytuje informácie potrebné pre zostavenie niektorých hodnôt atribútov sledovaných v modeli preferencií a samotný text tweetu. Pre zefektívnenie zisťovania atribútov súvisiacich s emocionálnym zafarbením textu berie ako argument objekt triedy Affect, ktorý napríklad obsahuje slovníky ANEW [13] a AFINN [27] načítané z databázy. Okrem toho tiež poskytuje informácie o odhadovanom type webstránky, ak je v tweete uvedené URL. Aktualizovaný diagram triedy TweetMeta sa nachádza na Obr. 6.11.

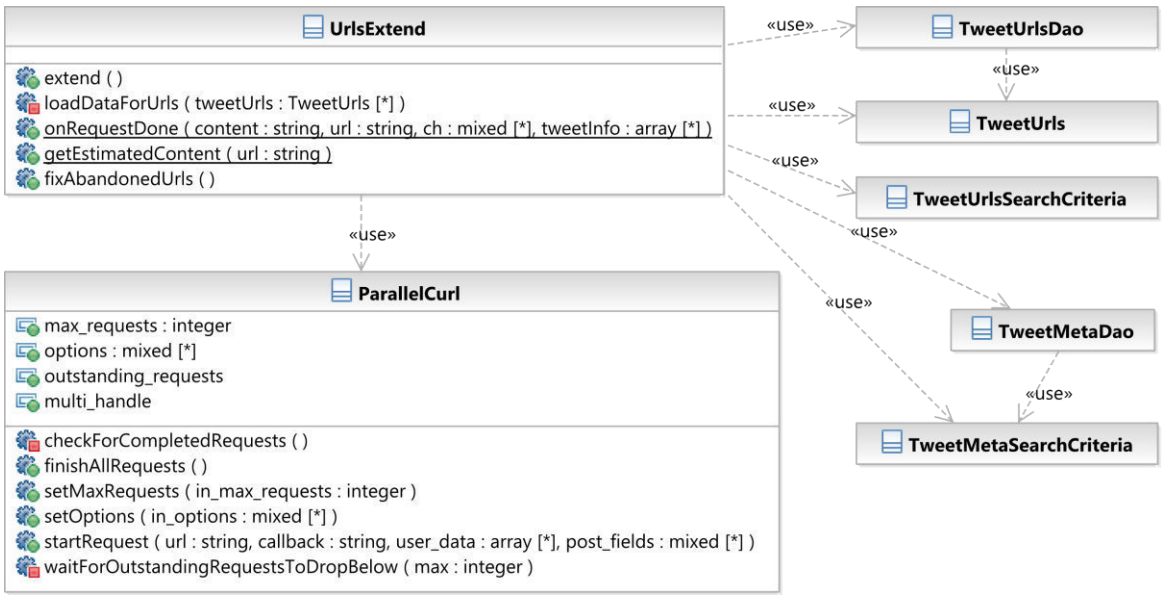
Diagram z Obr. 6.12 zobrazuje triedu `UrlsExtend` spolu so vzťahmi s inými triedami. Je určená pre zistenie a uloženie rozšírených informácií o URL z tweetov. Pre zistenie originálnej URL, ktorá bola skrátená a pre načítanie obsahu stránky používa už spomínanú prevzatú triedu `ParallelCurl`. Z obsahu potom extrahuje HTML `title` tág. Na základe regulárnych výrazov je tiež odhadovaný typ obsahu stránky. Tieto informácie sú potom použitím objektu triedy manipulujúcej s údajmi o URL (`TweetUrlsDao`) uložené do databázy. V databáze sú informácie

[8] `https://github.com/petewarden/ParallelCurl`
[9] `http://phpir.com/svm`

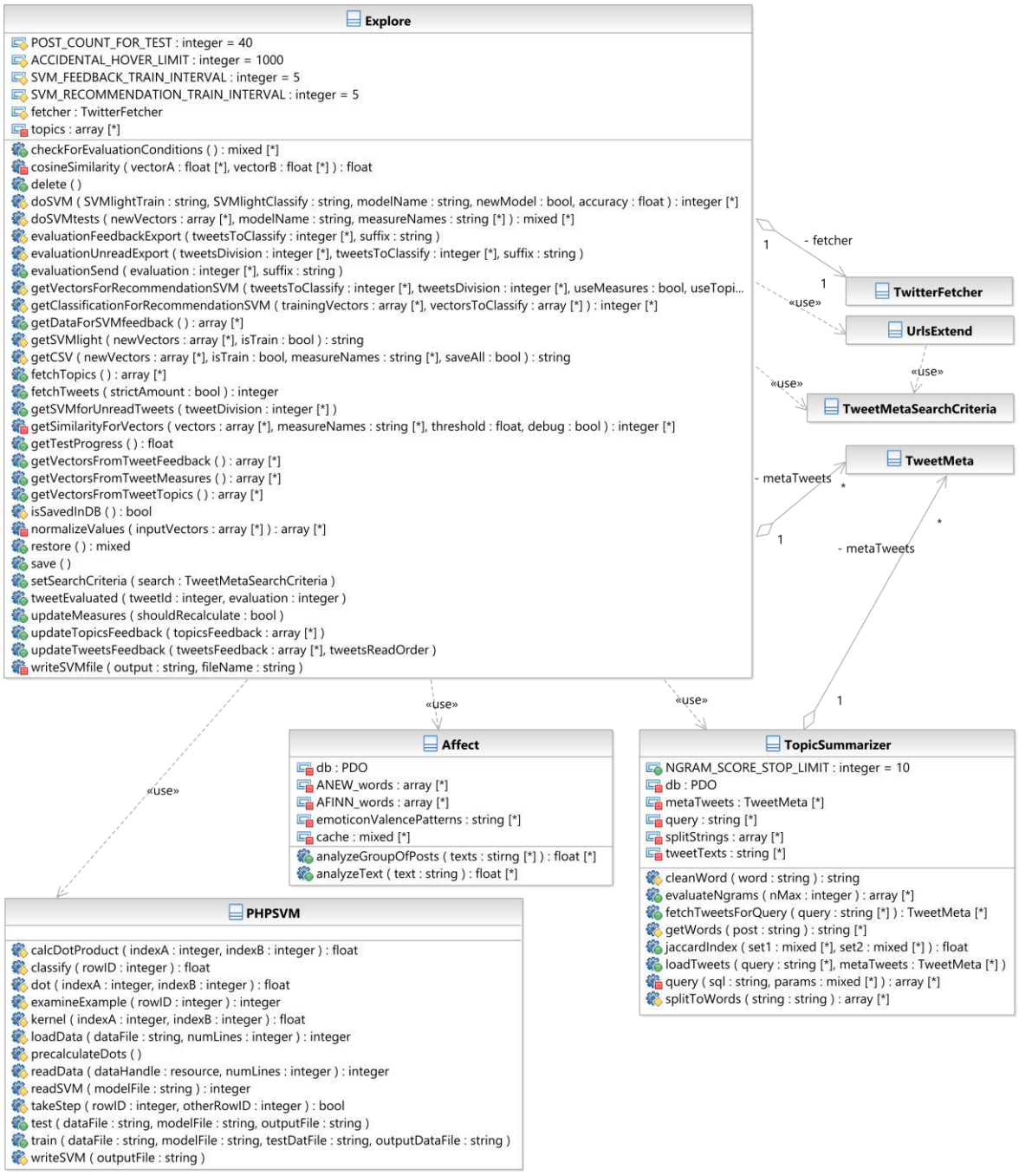
o URL v tweetoch uložené v osobitnej entite (pozri Obr. 6.2 a Obr. 6.10). Vyskytovali sa však prípady, kedy po načítaní tweetov cez Twitter API neboli počas ich ukladania do databázy extrahované všetky URL. Preto je implementovaná aj metóda `fixAbandonedUrls`, ktorá pre entity tweetov v databáze, pre ktoré neexistuje entita typu `twitter_tweet_urls`, ale obsahujú reťazec znakov „http“ znova extrahuje regulárnym výrazom platné http a https adresy. Tie potom uloží do príslušnej databázovej entity.



Obr. 6.11 Finálna podoba triedy TweetMeta, ktorá agreguje čiastkové triedy pre abstrakciu a prácu s čiastkovými údajmi o tweetoch.



Obr. 6.12 Štruktúra tried UrlsExtend a prevzatej triedy ParallelCurl spolu s ich vzťahmi s ostatnými triedami v systéme.



Obr. 6.13 Diagram tried, ktorý zachytáva vzťahy a štruktúru tried, ktoré implementujú funkcionalitu načítavania a zobrazovania obsahu a budovanie modelov interpretácie implicitnej spätnej väzby a modelu preferencií používateľa v rámci finálneho riešenia systému.

Obr. 6.13 zobrazuje diagram tried pre triedy, ktoré implementujú funkcionalitu zobrazovania a načítavania obsahu, jeho spracovanie pre tematickú sumarizáciu, počítanie atribútov pre modely strojového učenia a trénovanie a použitie modelu interpretujúceho implicitnú spätnú väzbu a modelu preferencií používateľa. Trieda `PHPSVM` je už spomínaným prevzatým riešením pre SVM algoritmus. Pre použitie v tejto diplomovej práci bola upravená metóda `train` tak, aby v prípade poskytnutia argumentu `testDataFile` použila pridaný argument `outputDataFile` tak, že je zavolaná aj metóda `test`, ktorá vytvorí súbor s názvom z argumentu `outputDataFile`

s výstupom klasifikácie pre testovací súbor. Objekt triedy `Affect` je zodpovedný za výpočet atribútov súvisiacich s emocionálnym zafarbením textového obsahu tweetov. Z databázy načíta slovníky do polí `ANEW_words` a `AFINN_words`, ktoré pozostávajú zo slov s priradenými hodnotami vystihujúcimi emocionálny podtón a metóda `analyzeText` vypočíta priemerné hodnoty pre zadaný text tak, že vypočíta priemery pre nájdené slová z textu. Okrem toho je vypočítaný aj priemer podľa regulárnych výrazov z `emoticonValencePatterns` pre nájdené emotikony. Metóda `analyzeGroupOfPosts` spojí zadané reťazce znakov a použije metódu `analyzeText` pre výpočet už uvedených hodnôt. Metóda nie je vo finálnom riešení použitá, no umožňuje podporu jednoduchej analýzy emócií napríklad pre všetky tweety pre tému. Keďže načítavanie slovníkov z databázy je výkonovo náročnejšie než práca s údajmi z operačnej pamäte, objekt tejto triedy je vhodný na znovupoužitie, keďže si slovníky pamätá a tiež používa pole `cache`, aby rovnaké texty neboli viackrát spracovávané, ak je výstup analýzy emócií známy.

Pre tematickú sumarizáciu slúži trieda `TopicSummarizer`, ktorá si do poľa `metaTweets` ukladá objekty typu `TweetMeta`. Tie sú načítané buď z databázy podľa zadaného dopytu `query`, alebo zadané ako argument metódy `loadTweets`. Texty tweetov rozdelené na slová poskytuje metóda `getWords`, ktorá využíva metódu `splitToWords` na samotné rozdelenie na slová a `cleanWords` slová konvertuje na malé znaky. Metóda `evaluateNgrams` sústreďuje funkcionalitu tematickej sumarizácie a argument `nMax` slúži na obmedzenie počtu slov v n-gramoch. Využíva metódu `jaccardIndex`, ktorá počíta Jaccardovu mieru podobnosti dvoch množín.

Trieda `Explore` sústreďuje veľké množstvo funkcionality a aj kvôli tejto komplexnosti nie sú v tejto dokumentácii uvedené niektoré metódy, ktoré nesúvisia priamo s implementáciou princípov finálneho riešenia. Rovnako nie sú uvedené atribúty triedy, keďže pre pochopenie fungovania objektu tejto triedy sú dôležitejšie samotné metódy. Keďže PHP skripty sú vykonávané na základe volania, pre udržanie potrebných informácií medzi rôznymi volania skriptov pracujúcich s objektom triedy `Explore` je objekt pre každé sedenie serializovaný a ukladaný do databázy. Túto funkcionalitu implementujú metódy `isSavedInDB`, `save`, `restore` a `delete`. Načítanie tweetov podľa dopytu používateľa, ich odoslanie na tematickú sumarizáciu a následné uloženie všetkých informácií slúžia metódy `setSearchCriteria` a `fetchTweets`, ktoré využívajú objekt typu `TwitterFetcher` uložený v atribúte `fetcher` a `fetchTopics`, ktorá informácie o témach ukladá v atribúte `topics`. Mód slúžiaci pre overenie riešenia využíva aj metódu `getTestProgress`, ktorá vracia podiel prečítaných tweetov k potrebnému počtu pre overenie, ktorý je stanovený atribútom `POST_COUNT_FOR_TEST`. Po prečítaní tweetu sa aktualizujú informácie o implicitnej spätnej väzbe prostredníctvom `updateTweetsFeedback` metódy pre jednotlivé tweety a agregované údaje podľa tém prostredníctvom `updateTopicsFeedback`. Metóda `updateTweetsFeedback` je implementovaná tak, že

z používateľského rozhrania prijíma aj poradie prečítaných tweetov (argument `tweetsReadOrder`), aby bolo možné trénovať modely na stanovenom počte posledných vzoriek. Tento postup ale nie je vo finálnom riešení využívaný. Ohodnotenie tweetov čiže explicitná spätná väzba je ukladaná prostredníctvom metódy `tweetEvaluated`. Po uložení ktoréhokoľvek typu spätnej väzby sú aktualizované aj atribúty tweetov (bez atribútov v súvislosti s témami) metódou `updateMeasures`. V prípade hodnoty `true` argumentu `shouldRecalculate` sú znovu vypočítané atribúty pre všetky tweety namiesto vypočítania atribútov len pre novo-prečítané tweety.

Za zostavenie modelu interpretujúceho implicitnú spätnú väzbu zodpovedá metóda `getDataForSVMfeedback`, ktorá načíta atribúty pre trénovaciu a testovaciu množinu prostredníctvom `getVectorsFromTweetFeedback`. Hodnoty sú vždy normalizované využitím metódy `normalizeValues`. Vstupy sú objektu triedy `PHPSVM` posúvané prostredníctvom súborov vo formáte SVMlight[10], konverziu do tohto formátu vykonáva metóda `getSVMlight` a vytvorenie súboru `writeSVMfile`. S pripravenými vstupmi pracuje najprv metóda `doSVMtests`, ktorá vykonáva trénovanie a krížovú validáciu a poskytuje dosiahnuté výsledky. Finálny model je natrénovaný a využitý na klasifikáciu vzoriek v metóde `doSVM`.

Model preferencií používateľa je zostavovaný podobne, atribúty preň poskytujú metódy `getVectorsFromTweetMeasures` a `getVectorsFromTweetTopics`, normalizované už spomínanou metódou `normalizeValues`. Funkcionalitu zastrešuje metóda `getSVMforUnreadTweets`.

Pre overenie je implementovaná metóda `checkForEvaluationConditions`, ktorá v prípade dosiahnutia požadovaného počtu prečítaných tweetov vyberie náhodné tweety pre overenie pre oba modely. Klasifikácia pre model interpretujúci implicitnú spätnú väzbu je exportovaná metódou `evaluationFeedbackExport` do formátu CSV so zabezpečením konverzie v metóde `getCSV`. Alternatívne modely preferencií používateľa sú trénované a použité na klasifikáciou v metóde `evaluationUnreadExport`, ktorá napríklad tiež prijíma v argumente `tweetsDivision` odhady ohodnotenia z predchádzajúceho modelu. Jednotlivé modely sú trénované na množinách poskytnutých metódou `getVectorsForRecommendationSVM`. Tieto množiny zostavuje podľa argumentov `useMeasures`, `useSimilarityThreshold`, `useClassAmountCompensation`, `useFeedbackEstimation`, a `useTopics` ktoré definujú použitie skupín atribútov a metód predspracovania opísaných v časti 4.3.2. Zoznam príliš podobných tweetov poskytuje metóda `getSimilarityForVectors`, ktorá na výpočet kosínusovej miery podobnosti volá metódu `cosineSimilarity`. Natrénovanie SVM modelu a klasifikáciu zadaných vektorov vykonáva

[10] Stručný opis formátu údajov v angličtine sa nachádza za nadpisom „Using The Classifier“ na adrese `http://phpir.com/support-vector-machines-in-php`, aktuálne k 3.5.2013

metóda `getClassificationForRecommendationSVM`. Aj tieto výstupy sú exportované do CSV súboru. Keď používateľ vykoná ohodnotenie tweetov určených na overenie, metóda `evaluationSend` uloží tieto výsledky do tretieho CSV súboru.

Zdrojové kódy obsahujú niekoľko ďalších tried a metód, ktoré nie sú explicitne dokumentované v tejto práci. Časť z nich tvorí implementáciu funkcionality, ktorá nie je priamo spätá s princípmi finálneho riešenia diplomovej práce, ale slúži len na celkovú podporu fungovania systému. Zvyšok tvorí implementácia alternatív, ktoré boli skúmané počas iteratívneho procesu hľadania finálnej implementácie. Síce nie sú potrebné pre fungovanie systému a nepodporujú opis riešenia z časti 4, môžu pomôcť pri možnom rozširovaní systému v rámci budúcej práce na projekte. Na priloženom digitálnom médiu sa nachádza aj dokumentácia zdrojových súborov vytvorená generátorom ApiGen[11] a obsahuje aj ďalšie informácie o použitých triedach, metódach a ich parametroch.

6.3 Dokumentácia k nasadeniu systému

Základný testovací systém opísaný v častiach 3.6 a 6.1 je súčasťou finálneho riešenia, preto jeho nasadenie je vykonané nasadením finálneho riešenia. Pre jeho používanie je ale potrebné vykonať niekoľko úprav opísaných v časti 6.3.2.

6.3.1 Nasadenie systému implementujúceho finálne riešenie

Pre nasadenie a spustenie systému je potrebný webový server s PHP modulom. Odporúčaná verzia PHP interpretera ja 5.3. Rovnako je potrebný server MySQL odporúčanej verzie 5.

Pre komunikáciu s Twitter API je potrebný kľúč zákazníka (z angl. customer key) a tajomstvo zákazníka (z angl. customer secret) pre použitie s protokolom OAuth. Tieto je možné získať na Twitter stránke pre vývojárov[12]. (Pre vylúčenie zablokovania týchto údajov ich v práci neposkytujem, ale poskytnem ich na opýtanie.) Tieto údaje je potrebné zadať do konfiguračného súboru (pozri nižšie).

Do priečinka obsluhovaného web serverom je potrebné skopírovať obsah priečinku `src`, ktorý je možné nájsť v digitálnej prílohe. Následne je nutné upraviť konfiguračný súbor `config/config.ini`.

Časť `db` konfiguračného súboru `config/config.ini` musí obsahovať platný `dsn` reťazec. Tento je v prílohe vyplnený len dočasnými údajmi. V ňom musí byť špecifikovaný spôsob pripojenia k databázovému serveru napr. adresou soketu a názov databázy. Ďalšie dva údaje sú `username` a `password`, čo zodpovedá prihlasovaciemu menu a heslu do databázy.

[11] `http://apigen.org/`
[12] `https://dev.twitter.com/`

Časť server súboru `config.ini` obsahuje len jeden údaj `path`, ktorý označuje adresu, ktorá ukazuje na koreňový adresár s prekopírovanými súbormi z priečinku `src`. Adresa musí byť absolútna a musí ju byť možné zadať do webového prehliadača.

V časti `twitter` súboru `config.ini` sa okrem `CONSUMER_KEY` a `CONSUMER_SECRET` spomínaných vyššie nachádza aj `OAUTH_CALLBACK` údaj, ktorý musí obsahovať adresu, ktorej použitie vyvolá `twitterCallback` skript. Stačí ale zmeniť tú časť adresy, ktorá odkazuje na konkrétny webový server a prípadne priečinok, v ktorom sa nachádzajú súbory z priečinku `src`. Je samozrejme potrebné vyplniť aj prvé dva údaje.

Príprava databázy pozostáva zo spustenia SQL príkazov, ktoré sa nachádzajú v podobe skriptov na priloženom digitálnom médiu. Prvým krkom je vytvorenie databázy na databázovom serveri MySQL s názvom, ktorý bol zadaný do konfiguračného. Z priečinku `db` je potom potrebné na databázovom serveri spustiť SQL skript `structure.sql` pre vytvorenie potrebnej štruktúry. Aby bolo možné využiť finálne riešenie, je ešte potrebné naplniť tabuľky so slovníkmi pre emocionálne zafarbenie slov spustením skriptov `data_ANEW_words.sql` a `data_AFINN_words.sql`. Keďže tematická sumarizácia používa aj všeobecný korpus tweetov pre získanie frekvencie výskytov n-gramov, aby nebolo potrebné tento korpus znova zostavovať, je priložený aj skript s korpusom načítaným vo februári 2013 s názvom `data_tweets_corpus.sql` a skripty s extrahovanými n-gramami spolu s informáciami o ich výskyte. Extrahované boli n-gramy do dĺžky 5, aj keď finálne riešenie používa maximálnu dĺžku 3. Množstvo údajov je už výrazne väčšie, preto sa tieto údaje vkladajú viacerými skriptami:

- `data_tweets_corpus_ngrams_1.sql`
- `data_tweets_corpus_ngrams_2.sql`
- `data_tweets_corpus_ngrams_3.sql`
- `data_tweets_corpus_ngrams_4.sql`

Počas experimentov s používateľským rozhraním či práce v systéme implementujúceho finálne riešenie sa do databázy ukladajú aj načítané tweety spolu s pridruženými metadátami a tiež sú uložené napríklad aj serializované objekty s informáciami potrebnými pre správne fungovanie pre jednotlivé sedenia. Tieto údaje nie sú potrebné pre funkcionalitu nasadeného systému, no je ich možné vložiť do databáza skriptom `data_other.sql`.

Týmto je systém pripravený na použitie.

6.3.2 Sprístupnenie systému pre experimentálne skúmanie používateľského rozhrania

Keďže vo finálnom riešení sú prístupné len skripty, ktoré implementujú toto finálne riešenie, pre sprístupnenie časti systému, ktorá implementovala rôzne rozhrania prezentácie a hodnotenia tweetov je potrebné vykonať nasledujúce úpravy.

V súbore `indexClass.php` je vo funkcii zodpovednej za výber súborov, ktoré sa vykonajú zodpovedná metóda `getPage`. Štandardne sú povolené len súbory potrebné pre implementáciu finálneho riešenia. Súbory implementujúce funkcionalitu experimentálneho skúmania používateľského rozhrania ako výstup DP II. je možné povoliť pridaním riadkov, ktoré uvádza Zdrojový kód 6.1 do príkazu `switch`, napr. medzi riadky `case 'test':` a `$pageName = $_GET['page'];`.

```
case 'home':
case 'query':
case 'twitterLogin':
case 'stars':
case 'q-stars':
case 'slide':
case 'q-slide':
case 'mosaic':
case 'q-mosaic':
```

Zdrojový kód 6.1 Kód na doplnenie umožňujúci prístup k systému experimentálneho skúmania používateľského rozhrania.

Pristúpiť k tejto časti systému je potom možné prostredníctvom adresy `http://dp.zilincik.eu/index.php?page=home` resp. je potrebné nahradiť doménu a prípadne doplniť cestu ku koreňovému adresáru finálneho riešenia. Test pozostával z troch alternatívnych používateľských rozhraní a po prečítaní stanoveného počtu tweetov bol používateľ presmerovaný na vyplnenie dotazníku. Prezrieť tieto časti (používateľské rozhrania a dotazníky) je ale možné aj bez nutnosti korektne ukončiť jednotlivé časti experimentu, a to priamym zadaním týchto adries (v prípade vlastného nasadenie systému je opäť potrebné nahradiť doménu a prípadne doplniť cestu ku koreňovému adresáru finálneho riešenia):

- `http://dp.zilincik.eu/index.php?page=stars`
- `http://dp.zilincik.eu/index.php?page=q-stars`
- `http://dp.zilincik.eu/index.php?page=slide`
- `http://dp.zilincik.eu/index.php?page=q-slide`
- `http://dp.zilincik.eu/index.php?page=mosaic`
- `http://dp.zilincik.eu/index.php?page=q-mosaic`

7 Použitá literatúra

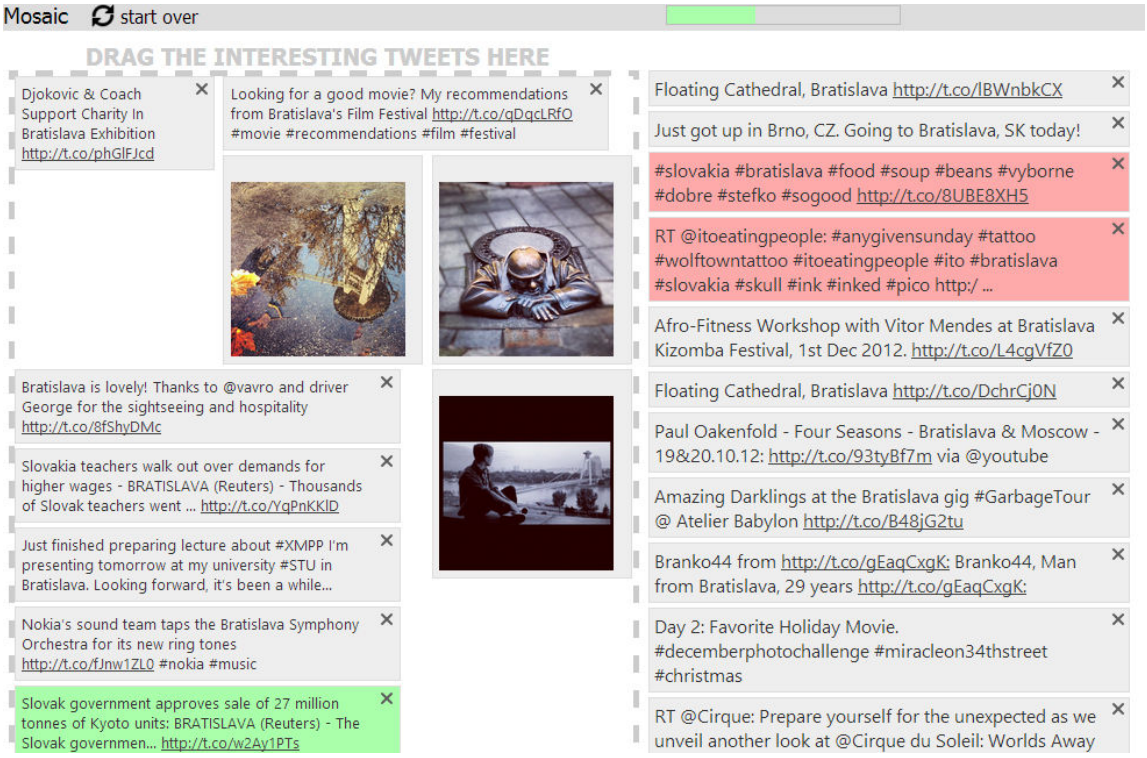
1. Twitter basics. 2012. https://support.twitter.com/groups/31-twitter-basics
2. Documentation. 2012. https://dev.twitter.com/docs/
3. Po-Hsiang, W., Jung-Ying, W., Hahn-Ming, L.: QueryFind: Search Ranking Based on Users' Feedback and Expert's Agreement. In: e-Technology, e-Commerce and e-Service, 2004 IEEE International Conference on, 2004, s. 299 –304.
4. Liu, Y.: CRRA: A Collaborative Approach to Re-Ranking Search Results. In: Educational and Information Technology (ICEIT), 2010 International Conference on, 2010, s. V2-214 – V2-218.
5. Pohorec, S., Čeh, I., Zorman, M. et al.: Automated parametrization of custom search ranking functions. In: Electronics and Information Engineering (ICEIE), 2010 International Conference on, 2010, s. V1-374 – V1-378.
6. Zhuang, Z., Cucerzan, S.: Exploiting Semantic Query Context to Improve Search Ranking. In: Semantic Computing, 2008 IEEE International Conference on, 2008, s. 50 – 57.
7. Nagmoti, R., Teredesai, A., De Cock, M.: Ranking Approaches for Microblog Search. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010, s. 153 – 157.
8. Duan, Y., Jiang, L., Qin, T. et al.: An Empirical Study on Learning to Rank of Tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, 2010, s. 295 – 303.
9. Uysal, I., Bruce Croft, W.: User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In: Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, s. 2261 – 2264.
10. Weng, J., Lim, E., Jiang, J. et al.: TwitterRank: Finding Topic-sensitive Influential Twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, 2010, s. 261 – 270.
11. Trifan, M., Ionescu, D.: A New Search Method for Ranking Short Text Messages Using Semantic Features and Cluster Coherence. In: Computational Cybernetics and Technical Informatics (ICCC-CONTI), 2010 International Joint Conference on, 2010, s. 643 – 648.
12. Naveed, N., Gottron, T., Kunegis, J. et al.: Searching Microblogs: Coping with Sparsity and Document Quality. In: Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, s. 183 – 188.

13. Bradley, M., Lang, P.: Affective norms for English words (anew): Stimuli, instruction manual and affective ratings. Technical report c-1, Gainesville, FL: University of Florida, (1999).
14. De Choudhury, M., Counts, S., Czerwinski, M.: Identifying Relevant Social Media Content: Leveraging Information Diversity and User Cognition. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia, 2011, s. 161 – 170.
15. Kamath, K., Caverlee, J.: Expert-Driven Topical Classification of Short Message Streams. In: Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom), 2011, s. 388 – 393.
16. Han, B., Baldwin, T.: Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, s. 368 – 378.
17. Zanzotto, F., Pennacchiotti, M., Tsioutsiouliklis, K.: Linguistic Redundancy in Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, s. 659 – 669.
18. Chen, Q., Shipper, T., Khan, L.: Tweets Mining Using WIKIPEDIA and Impurity Cluster Measurement. In: Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on, 2010, s. 141 – 143.
19. Zhang, B., Guan, Y., Sun, H. et al.: Survey of user behaviors as implicit feedback. In: Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on, 2010, s. 345 – 348.
20. Oh, J., Lee, S., Lee, E.: A User Modeling using Implicit Feedback for Effective Recommender System. In: Convergence and Hybrid Information Technology, 2008. ICHIT '08. International Conference on, 2008, s. 155 – 158.
21. Nobarany, S., Oram, L., Rajendran, V. K. et al.: The Design Space of Opinion Measurement Interfaces: Exploring Recall Support for Rating and Ranking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, s. 2035 – 2044.
22. Broekens, J., Brinkman, W.: AffectButton: Towards a Standard for Dynamic Affective User Feedback. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, 2009, s. 1 – 8.
23. Feinberg, M., Geisler, G., Whitworth, E. et al.: Understanding Personal Digital Collections: An Interdisciplinary Exploration. In: Proceedings of the Designing Interactive Systems Conference, 2012, s. 200 – 209.
24. Inouye, D., Kalita, J. K.: Comparing Twitter Summarization Algorithms for Multiple Post Summaries. In: Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third

International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011, s. 298 – 306.

25. Kang, R., Fu W., Kannampallil, T.: Exploiting knowledge-in-the-head and knowledge-in-the-social-web: effects of domain expertise on exploratory search in individual and social search environments. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, 2010, s. 939 – 402.
26. O'Connor, B., Krieger, M., Ahn, D.: TweetMotif: Exploratory Search and Topic Summarization for Twitter. In: Proc. of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, s. 384 – 385.
27. Nielsen, F.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings, 2011, s. 93 – 98.
28. Kosková, Gabriela. 2007. Vyhodnocovanie metód dolovania znalostí. http://www2.fiit.stuba.sk/~polcicova/ZZ/prednasky/08_vyhodnocovanie.pdf

PRÍLOHA 1: Ukážka implementovaného hodnotiaceho rozhrania



PRÍLOHA 2: Ukážka dotazníka po práci s rozhraním pre testovanie spôsobov hodnotenia

Survey / Slide

Please answer few questions regarding evaluation using **Slide**.

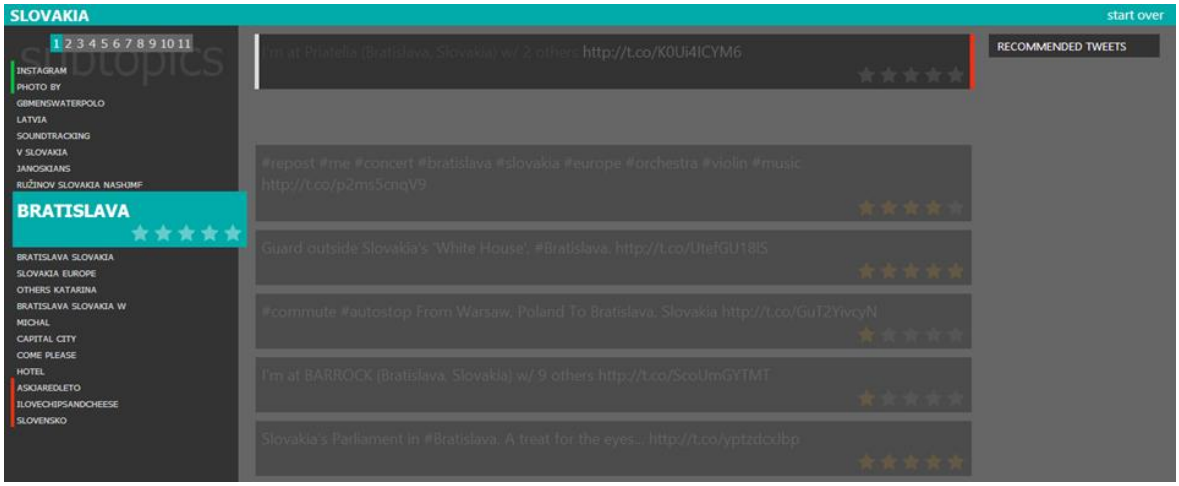
	totally disagree				totally agree
It was **too hard for me to chose** the right option (I could not decide easily where to drop the tweet)	○	○	○	○	○
It **took me too much time** to chose the right option.	○	○	○	○	○
I would definitely like to have **more options to choose from**, while rating.	○	○	○	○	○
I was **able to express** my **opinion**.	○	○	○	○	○
I don't like that I **must chose rating in order to hide the tweet**.	○	○	○	○	○
It was **fun to use the slide interface**.	○	○	○	○	○
I would **recommend to use this interface**.	○	○	○	○	○

If you have anything to add, please comment the slide rating interface.

Save answers

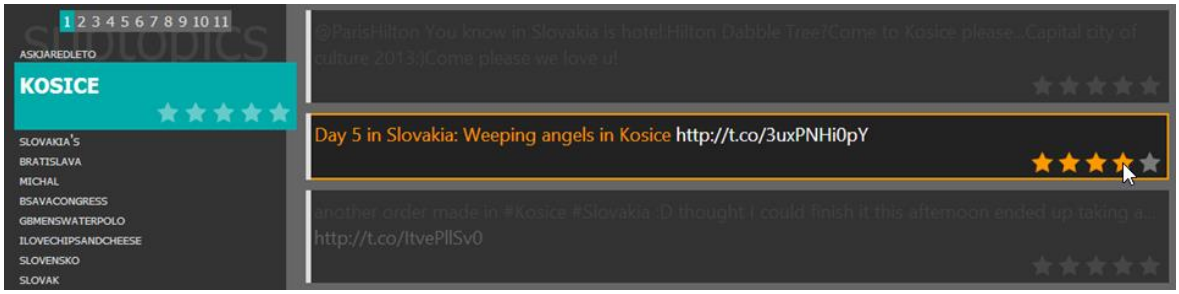
PRÍLOHA 3: Používateľské rozhranie finálneho riešenia

Na obrázku nižšie je vidieť používateľské rozhranie v stave po tom, ako používateľ zadal dopyt „Slovakia“, ktorý sa zobrazuje vľavo hore. Časť pod ním zobrazuje témy, vždy 20 na jednu stranu (ďalšie témy je možné zobraziť kliknutím na číslo strany nad zoznamom tém). Používateľ vybral tému „Bratislava“. Stredná časť používateľského rozhrania zobrazuje tweety, ktorých text nie je možné ľahko prečítať, pokiaľ nad ne nie je umiestnený kurzor myši. Tým je možné získavať spoľahlivejšiu implicitnú spätnú väzbu. Tweety s bledším pozadím sú už prečítané a je tiež vidieť, ako boli ohodnotené používateľom (podľa počtu zvýraznených hviezdičiek). Prvý zobrazený tweet bol práve prečítaný používateľom a červený pásik napravo naznačuje, že podľa implicitnej spätnej väzby používateľa nezaujal. Pri niektorých témach v zozname vľavo sa tiež nachádza farebný indikátor. Zelený označuje témy, ktoré model preferencií vyhodnotil ako zaujímavé pre používateľa a červený práve označuje predikciu tém za nezaujímavé.



Obrázok vyššie tiež zobrazuje biely pásik na ľavej strane tweetu, čo znamená, že model preferencií nevedel rozhodnúť o zaujímavosti tohto tweetu pre používateľa. V prípade tweetov vyhodnotených za zaujímavé sa namiesto bieleho indikátora zobrazuje zelený, v prípade nezaujímavých červený.

Nad tweet v strede bol umiestnený kurzor myši, čím je tweet ľahšie čitateľný než ostatné. Táto funkcionalita slúži pre zabezpečenie získavania času pozornosti pre text tweetu. Je tiež vidieť, že používateľ vyberá hodnotenie štyrmi hviezdičkami z piatich.



PRÍLOHA 4: Obsah priloženého digitálneho média

Priložené digitálne médium obsahuje nasledujúce súbory alebo priečinky s uvedeným obsahom.

readme.txt – obsah tejto prílohy

Prieskumné vyhľadávanie na sociálnych sieťach so zameraním na dynamické kritéria a vzťahy metadát k obsahu.pdf – Diplomová práca

Anotácia SK.pdf – anotácia v slovenskom jazyku

Annotation EN.pdf – anotácia v anglickom jazyku

product – priečinok s implementovaným systémom

db – priečinok so SQL skriptami pre vytvorenie štruktúry databázy a jej naplnenie údajmi

doc – priečinok s dokumentáciou k systému (spustenie súborom index.html)

src – zdrojové súbory systému

PRÍLOHA 5: Návrh článku do časopisu PLOS ONE

Exploratory Search on Twitter Utilizing User Feedback and Multi-perspective Microblog Analysis

Michal Zilincik, Pavol Navrat, Gabriela Koskova

ABSTRACT

In recent years, besides typical information retrieval, a broader concept of information exploration – *exploratory search* is getting into foreground. Moreover, more and more valuable information is presented in microblogs on *social networks*. We propose a new method for supporting the exploratory search on Twitter social network. The method copes with several challenges, namely brevity of microblogs called tweets, limited number of available ratings and need to process the recommendations online. To tackle the first challenge, the representation of microblogs is enriched by information from the referenced links, topic summarization and affect analysis is adopted. Small number of available ratings is raised by interpreting implicit feedback trained by *feedback model* during browsing. Recommendations are made by *preference model* that models user's preferences over tweets. The evaluation shows promising results even when navigating in the space of brief pieces of information, making recommendations based only on small number of ratings, and by optimizing the models to process in real time.