TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Adaptívne modely pre detekciu anti-sociálneho správania na webe

Diplomová práca

2022 **Bc. Samuel Maťaš**

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Adaptívne modely pre detekciu anti-sociálneho správania na webe

Diplomová práca

Študijný program:	Hospodárska informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	Ing. Martin Sarnovský, PhD.

2022 Košice **Bc. Samuel Maťaš**

Abstrakt

Táto diplomová práca sa zaoberá problematikou týkajúcou sa detekcie antisociálneho správania na webe, konkrétne detekcie falošných správ a dezinformácii a problematikou spojenou s rozpoznávaním a klasifikovaním takýchto online príspevkov. Teoretická časť práce poukazuje na rôzne druhy antisociálneho správania na webe a ich dopady na spoločnosť v reálnom svete ale aj v online priestore, čím zdôrazňuje aktuálnosť riešenej témy. Taktiež sa v tejto časti popisujú prístupy, ktorými sa sociálne siete doteraz snažili obmedzovať dezinformácie a nedostatky týchto prístupov. Poslednou témou teoretickej časti tejto diplomovej práce je popis rôznych doteraz využívaných metód pre detekciu antisociálneho správania na webe, ich výhody a nedostatky v aktuálnej dobe a odôvodnenie, prečo nie sú dostatočne efektívne a je nutné ich zlepšiť resp. nahradiť. Praktická časť tejto diplomovej práce je zameraná na experimenty, ktoré ku problematike klasifikácie resp. detekcie falošných správ pristupujú pomocou prúdov dát, čím simulujú dynamickosť online priestoru. V experimentoch sú používané rôzne adaptívne a neadaptívne modely s cieľom porovnať ich efektívnosť v klasifikácii falošných správ z dátových prúdov. Cieľom tejto diplomovej práce je otestovať, či je predpoklad, že adaptívne metódy využívajúce detekciu zmien, takzvanú drift detekciu, budú v úlohe klasifikácie falošných správ z dátových prúdov efektívnejšie ako neadaptívne metódy bez detektoru driftov.

Kľúčové slová

detekcia falošných správ, dezinformácie, konceptový drift, dátové prúdy, klasifikácia, adaptívne modely, strojové učenie

Abstract

This diploma thesis focuses on the topic of detecting antisocial behavior on the web. The main focus is specifically directed at the detection of fake news and misinformation and also the underlying problems related to the recognition and classification of such online posts. The theoretical framework of this thesis describes various types of anti-social behavior on the web and their impact on society in the real world but also in online space, which emphasizes the relevance of the subject matter. This part of the thesis also describes the approaches that social networks have used so far to limit the spread of misinformation and the shortcomings of these approaches. It also provides a description of various methods used so far for the detection of antisocial behavior on the web, their advantages and disadvantages at present and the reason why they are not effective enough which points out the necessity to improve them or to replace them completely with a better approach. The main part of this thesis are the performed experiments in which fake news are classified using data streams to simulate the dynamic, everchanging nature of the online space. Various adaptive and non-adaptive models are used in these experiments in order to compare their effectiveness in classifying fake news from data streams. The aim of this thesis is to test whether the assumption that adaptive methods using change detection, so-called drift detection, will be more effective in the task of classifying fake news from data streams than non-adaptive methods without a drift detector.

Key words

fake news detection, misinformation, conceptual drift, data streams, classification, adaptive models, machine learning

64283

TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Katedra kybernetiky a umelej inteligencie

ZADANIE DIPLOMOVEJ PRÁCE

Študijný odbor: **Informatika**
Študijný program: **Hospodárska informatika**

Názov práce:

Adaptívne modely pre detekciu anti-sociálneho správania na webe

Adaptive models for detection of anti-social behavior on the web

Študent: **Bc. Samuel Maťaš**
Školiteľ: **Ing. Martin Sarnovský, PhD.**
Školiace pracovisko: **Katedra kybernetiky a umelej inteligencie**
Konzultant práce:
Pracovisko konzultanta:

Pokyny na vypracovanie diplomovej práce:
1. Podať teoretický prehľad oblasti vybraných druhov anti-sociálneho správania na webe.
2. Podať prehľad súčasného stavu v oblasti adaptívnych modelov strojového učenia a ich možného využitia v úlohách detekcie anti-sociálneho správania sa na webe.
3. Navrhnúť a implementovať vybrané adaptívne metódy na vybraných dátových prúdoch.
4. Experimentálne otestovať a vyhodnotiť implementované prístupy.
5. Vypracovať dokumentáciu podľa pokynov vedúceho práce.

Jazyk, v ktorom sa práca vypracuje: slovenský
Termín pre odovzdanie práce: 22.04.2022
Dátum zadania diplomovej práce: 29.10.2021

.....
prof. Ing. Liberios Vokorokos, PhD.
dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som celú diplomovú prácu vypracoval samostatne s použitím uvedenej odbornej literatúry.

Košice, 22. apríla 2022 .....

vlastnoručný podpis

Poďakovanie

Chcel by som sa poďakovať vedúcemu práce, Ing. Martinovi Sarnovskému, PhD. za odbornú pomoc, cenné rady a usmernenie pri vypracovávaní mojej diplomovej práce.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam symbolov a skratiek

ADWIN	Adaptive Windowing
CNN	Convolutional Neural Networks
DEF	Default
DT	Decission Tree
EDDM	Early Drift Detection Method
GRU	Gated recurrent unit
HAT	Hoeffding Adaptive Tree
HDDM_A	Hoeffding Drift Detection Method Average-test
HDDM_W	Hoeffding Drift Detection Method Weighted average-test
KNN	K-Nearest Neighbors
KSWIN	Kolmogorov-Smirnov Windowing
LDA	Latentná Dirichletova Alokácia
LPPNSE	Learn Plus Plus Non-Stationary Environments
LSTM	Long Short-Term Memory
NB	Naive Bayes
NBA	Naive Bayes Adaptive
NSE	Non-Stationary Environments
PH	Page Hinkley
RNN	Recurrent Neural Networks
SRP	Streaming Random Patches

Úvod

V súčasnosti predstavujú dezinformácie a falošné správy veľký problém, ktorý ovplyvňuje priamo či nepriamo každého človeka bez ohľadu na to, či je alebo nie je používateľom internetu a sociálnych sietí. V dobe internetu, v ktorej sa informácie šíria vo veľmi veľkom počte a veľmi veľkou rýchlosťou, je takmer nemožné identifikovať pri každom príspevku, zdieľaní, statuse, článku atď. či je alebo nie je pravdivý. Dezinformačné príspevky sa na sociálnych sieťach šíria mnohonásobne rýchlejšie ako faktické príspevky, keďže tieto príspevky mnohokrát využívajú šokujúco alebo kontroverzne znejúce informácie, ktoré pre mnohých používateľov pôsobia atraktívnejšie. To pochopiteľne predstavuje veľké riziko pre spoločnosť a demokraciu ako takú. Názory ľudí založené na nepravdivých informáciách sú veľmi silné a čoraz viac vedú k polarizácií spoločnosti, keďže aj keď sa presvedčenie niektorých ľudí preukáže ako nesprávne, svoj názor nezmenia. Okrem iného majú tieto príspevky na webe aj reálne dopady ako rasisticky alebo protinábožensky motivované útoky spôsobené dezinformáciami o istej etnicite resp. viere. Najaktuálnejšie sú dezinformácie spojené s pandémiou COVID-19, ktoré majú mnoho pod-tém napríklad nebezpečné dezinformačné rady ako ochorenie liečiť, dezinformácie o vakcínach, dezinformácie o pôvode ochorenia atď.

Diplomová práca sa delí na dve hlavné časti a to teoretickú časť a praktickú časť. V teoretickej časti sú popísané rôzne formy antisociálneho správania na webe, motivácia ich vzniku a ich dopady na spoločnosť a používateľov. Taktiež sú popísané rôzne prístupy, ktoré sa sociálne siete snažili použiť pre reguláciu a boj s dezinformáciami. Následne sú popísané aktuálne využívané metódy resp. experimenty, ktoré sa snažia efektívne bojovať proti dezinformáciám, či už s využitím ľudského faktora alebo použitím metód strojového učenia. Mnoho z týchto metód má ale nedostatky, na ktoré sa taktiež v teoretickej časti práce poukazuje. Jedným z hlavných nedostatkov mnohých experimentov je, že používajú statické prístupy pre klasifikáciu falošných správ aj keď je reálne online prostredie dynamické a rýchlo sa meniace.

V praktickej časti práce sú popísané experimenty, ktoré boli zamerané na detekciu resp. klasifikáciu antisociálneho správania z textových dát, najčastejšie reprezentovaných Twitter príspevkami. Pre experimenty boli využívané dátové prúdy, s cieľom nasimulovať pribúdanie príspevkov ako to je v online prostredí. Súčasťou experimentov je porovnávanie adaptívnych metód, ktoré disponujú detekciou zmien s neadaptívnymi metódami, ktoré detekciu zmien nemajú. Adaptácia na zmeny resp. na konceptové drifty spôsobené určitými zmenami je pri tejto klasifikačnej úlohe kľúčová, keďže príspevky v online priestore sa neustále menia rôznymi spôsobmi a je nutné aby boli využívané modely na tieto zmeny prispôsobené, ináč sa stanú postupom času nepoužiteľné.

1. Formulácia úlohy a cieľ práce

Táto diplomová práca je zameraná na tému detekcie antisociálneho správania na webe. Keďže je téma antisociálneho správania na webe veľmi aktuálnou témou, hlavne pri falošných správach ale aj iných kategóriách, je doležíte podať podrobné informácie o tejto problematike.

1.1. Teoretický prehlaď antisociálneho správania na webe

Prvou časťou tejto diplomovej práce je podanie teoretického prehľadu v oblasti antisociálneho správania na webe. Táto časť je zameraná na popis rôznych druhov antisociálneho správania na webe a ich dopadov na spoločnosť ale aj na skupiny a jednotlivcov. Pri jednotlivých druhoch tohto správania je dôležité popísať spôsoby akými negatívne vplývajú na svet, aké pojmy sú často krát spájané s jednotlivými druhmi antisociálneho správania a aký je ich vývoj resp. aktuálny stav. Táto diplomová práca sa primárne zameriava na tému detekcie falošných správ na webe, preto je potrebné popísať rôzne druhy falošných správ, motiváciu pre ich tvorbu a šírenie, rýchlosť akou sa šíria a hlavne poukázať na ich katastrofálne dopady, ktoré sa okrem online priestoru pretavujú aj do reálneho sveta, v ktorom žijeme. Keďže sa dezinformácie šíria primárne na sociálnych sieťach, je nutné popísať na akých konkrétne a aké sú s tým spojené následky. Taktiež je potrebné spomenúť z akého dôvodu sa to deje práve na týchto sociálnych sieťach a opísať resp. porovnať súčasné a historické pokusy daných sociálnych sieti s bojom proti dezinformáciám. Zhodnotiť relevanciu jednotlivých metód, keďže sú často krát na hranici s kontroverziou, napríklad regulácia príspevkov alebo blokovane šíriteľov poloprávd a falošných správ, čo potenciálne umlčuje slobodu slova. K tejto časti práce patrí aj preskúmanie a popísanie nedostatkov používaných metód ako napríklad nedôvera používateľov v tradičné metódy boja proti dezinformáciám, prečo nastáva a na základe čoho táto nedôvera vzniká, či má relevantné odôvodnenie alebo nie.

1.2. Súčasný stav modelov detekcie antisociálneho správania na webe

Ďalšou časťou práce je popis jednotlivých prístupov v detekcii antisociálneho správania na webe. Výhody týchto prístupov založených na strojovom učení oproti klasickým doteraz používaným metódam. Popis toho, ako tieto metódy k samotnej detekcii antisociálneho správania pristupujú, kedy vznikli a kedy boli implementované, aké sú ich výhody a nevýhody resp. nedostatky a ako by sa tieto nedostatky teoreticky dali vyriešiť. Poukázanie na hlavný problém s detekciou antisociálneho správania pri mnohých modeloch a to vysporiadanie sa s konceptovým driftom teda rôzne prístupy modelov ku konceptovému driftu a ich schopnosť resp. neschopnosť adaptácie. Taktiež popis nedostatkov experimentov a modelov v nich použitých, keďže je v nich mnohokrát online prostredie vnímané ako statické napriek tomu, že je dynamické, teda neustále sa meniace.

1.3. Implementácia adaptívnych metód na dátové prúdy

Tento bod práce popisuje praktickú časť, v ktorej boli použité rôzne adaptívne modely na rôznych prúdoch dát s cieľom klasifikovať falošné správy resp. dezinformácie. K problematike sa pristupovalo použitím adaptívnych modelov a dátových prúdov s cieľom nasimulovať dynamiku online prostredia a otestovať ako dobre budú vybrané modely klasifikovať falošné správy rôzneho typu.

1.4. Experimentálne testovanie

Pre experiment boli použité rôzne datasety resp. ich kombinácie a taktiež rôzne modely s rôznymi nastaveniami, ktoré sa obmieňali a porovnávali s cieľom nájsť ideálne a najefektívnejšie riešenie pre detekciu falošných správ pomocou týchto modelov. Pre jednotlivé datasety boli zakaždým použité viaceré modely a následne boli porovnané ich výsledné metriky, teda ako efektívne klasifikujú príspevky.

Jedným z hlavných cieľov práce je otestovať, či sa adaptívne modely zo softvérovej knižnice „scikit-multiflow“ implementované na prúdy dát dokážu dostatočne prispôsobovať na zmeny nastávajúce v daných prúdoch a teda či sú tieto modely potenciálne použiteľné v dynamickom online prostredí pre detekciu antisociálneho správania.

2. Teoretický rozbor zvolenej témy

V súčasnosti sa stal internet takmer neoddeliteľnou súčasťou života pre väčšinu ľudí. Internet má nespočetné množstvo benefitov ako napríklad jednoduché zabezpečenie komunikácie v rôznych formách, získavanie a zdieľanie informácií a celkovo prináša rôzne benefity, ktoré ľuďom uľahčujú prácu, zabezpečujú zábavu a zjednodušujú život. Internet má ale ako každá vymoženosť okrem benefitov aj negatívne aspekty a jedným z nich je pravé antisociálne správanie na webe. Antisociálne správanie na webe je pomerne rozsiahly pojem, ktorý zahŕňa široké spektrum činností ale aj skupín a ľudí. Antisociálne správanie na webe je možné podľa obsahu rozdeliť na dve kategórie a to antisociálne správanie sa používateľov a generovanie dezinformačného obsahu.

2.1. Antisociálne správanie sa používateľov na webe

Do tejto kategórie patrí obsah na webe, ktorý môže mať rôzne formy, najčastejšie sa ale vyskytuje v textovej forme. Cieľom používateľov zapadajúcich do tejto kategórie je negatívne ovplyvniť online konverzácie alebo uraziť iných používateľov, skupiny, menšiny atď. Niektoré príspevky týchto používateľov sú písané s jednoznačným zámerom napádať a uraziť, zatiaľ čo iné príspevky sú písané len ako provokácia pre získanie reakcií.

2.1.1. Online trolling

K antisociálnemu správaniu na webe patria napríklad rôzne formy antisociálnych používateľov, ktorí sú často krát označovaní ako trollovia. Títo používatelia uverejňujú komentáre so zámerom uraziť iných používateľov, vyvolať chaos alebo priamo útočiť na ostatných komentujúcich resp. autorov príspevku, pri ktorom svoje komentáre uverejňujú. Vo všeobecnosti by sa dalo povedať, že správanie týchto používateľov je v plnej miere zámerné, čiže sa v žiadnom prípade nejedná o konštruktívnu kritiku alebo príspevky so satirickým významom. Trollovia častokrát komentujú na sociálnych sieťach, diskusných fórach a ďalších online platformách, kde začínajú hádky a píšu provokačné príspevky. Tieto provokačné príspevky nemusia byť pre komentujúceho provokatéra témou vôbec zaujímavé a ide mu často krát iba o vyprovokovanie reakcie. Témy, pri ktorých spomenutí požívatelia komentujú sú častokrát citlivého charakteru, napríklad prírodné katastrofy alebo úmrtia známych osobností, keďže pri týchto témach je veľká šanca, že ich provokácia bude mať vyžadovanú odpoveď.

Existujú ale aj názory, ktorých tvrdenie je, že online trollovia ale aj činnosť, ktorú vykonávajú, taktiež známa ako „trolling“, nie je jasne definovaná ani jasne odlišujúca sa od iných podobných antisociálnych aktivít na webe. Dôvodom týchto nejasností je napríklad vývoj samotného pojmu troll ale aj nejednoznačný význam činnosti, ktorú títo používatelia vykonávajú, keďže je slovo „trolling“ v súčasnosti často krát spájané aj s nevinnými humornými príspevkami ale na druhej

strane aj so záväznými šikanujúcimi príspevkami, ktoré boli spomenuté. Za smerodajnú sa dá považovať definícia [7], opisujúca online trolling ako „Opakujúce sa online rušivé, deviantné správanie sa jednotlivca voči iným jednotlivcom a skupinám, pričom má toto správanie širokú škálu významov a kontextov“.

2.1.2. Nenávistné prejavy na webe

Väčšina bežných používateľov na sociálnych sieťach má uvedené svoje úradné meno, prípadne mesto, v ktorom žijú alebo ďalšie osobné informácie, ktoré umožňujú aspoň čiastočne prepojiť tieto virtuálne profily s reálnou osobou. Tieto údaje ale nie sú v drvivej väčšine sociálnych sietí a fór vyžadované ani žiadnym spôsobom overované. Toto prostredie umožňuje používateľom v online priestore uverejňovať rôzny obsah napríklad komentáre bez toho, aby sa museli akokoľvek identifikovať. Anonymita takéhoto typu má za následok, že mnoho používateľov, napríklad s extrémistickými sklonmi, začne na internete vytvárať ofenzívne resp. nenávistné príspevky, keďže pri zachovaní anonymity sa prakticky za svoje vyjadrenia nemusia obávať takmer žiadnych následkov pretavujúcich sa do reality.

Medzi ofenzívne prejavy na webe patria napríklad príspevky podnecujúce k nenávisti a násiliu voči určitým skupinám ľudí na základe ich rasy, pohlavia, sexuálnej orientácie, náboženstva alebo rôznych iných charakteristických vlastností. Prvá extrémistická webstránka s ofenzívnym nenávistným obsahom bola vytvorená už v roku 1995 [29] a do roku 2000 bol počet webstránok tohto typu približne 400. V roku 2010 bol odhadovaný počet týchto stránok okolo 8 000. Jedná sa teda o pomerne veľký problém, keďže okrem samotných webstránok zameraných na nenávistný obsah existuje mnoho sociálnych sietí resp. skupín na sociálnych sieťach, zameraných na publikovanie obsahu tohto typu, pričom majú oveľa väčší dosah aj na bežných, priamo nezainteresovaných používateľov danej sociálnej siete.

Sociálne siete majú striktné pravidlá týkajúce sa ofenzívnych a nenávistných príspevkov. Twitter nepovoľuje akúkoľvek formu propagácie nenávisti voči iným používateľom alebo skupinám [30] a taktiež nepovoľuje existenciu účtov, ktorých cieľom je propagácia takéhoto nenávistného obsahu. Odstránenie týchto príspevkov môže ale niekedy trvať dlhšiu dobu, čo má za následok, že sa s nenávistnými príspevkami na sociálnych sieťach stretne väčšie množstvo používateľov. Pri niektorých príspevkoch je náročné automaticky určiť či sa jedná o nenávistný príspevok alebo nie, keďže môže ísť o satiru resp. príspevky členov v rámci „chránenej kategórie“. Napríklad členovia chránenej kategórie na sociálnej sieti Twitter [30] sa môžu navzájom označovať pomocou výrazov, ktoré sa zvyčajne považujú za nadávky, no v kontexte, v ktorom nie je ich cieľom uraziť nie sú tieto

príspevky regulované, jedná sa pritom napríklad o členov afroamerickej komunity používajúcich výraz, ktorý je historicky spojený s utláčaním tejto komunity.

2.2. Generovanie dezinformačného obsahu

Do tejto kategórie patrí obsah, ktorý je generovaný s cieľom zmiasť alebo ovplyvniť ostatných používateľov. Dezinformačný obsah na webe sa vyskytuje v rôznych formách ako napríklad falošné správy v textovej podobe, videá s cielene mätúcim názvom alebo obsahom alebo falošné hodnotenia služieb a produktov s cieľom zmiasť a ovplyvniť zákazníkov.

2.2.1. Falošné recenzie

Ďalšou skupinou spájanou s antisociálnym správaním na webe sú falošné hodnotenia resp. recenzie produktov a služieb známe ako „fake reviews“. Tieto hodnotenia existujú vo forme rankingových systémov, komentárov známych ako spätná väzba alebo recenzia a nakoniec v kombinácii obidvoch foriem. Zaujímavým faktom pri falošných hodnoteniach je, že ich je možné nájsť ako v negatívne pôsobiacej, tak aj v pozitívne pôsobiacej podobe. V konečnom dôsledku ale obe praktiky patria ku antisociálnemu správaniu.

V nedávnej minulosti, konkrétne pred rokom 2017, bolo pri veľmi známom internetovom obchode zistené, že jeho zákazníci mnohokrát obdržali „darček“ najčastejšie vo forme online zľavovej poukážky, za podmienok, že zanechajú 5 hviezdičkové hodnotenie na produkt, ktorý si predtým na danom internetovom obchode zakúpili. Daná spoločnosť tieto hodnotenia nazývala ako takzvané „stimulované recenzie“, keďže sa jednalo o reálnych zákazníkov, ktorých chceli takto motivovať k hodnoteniu produktov. Problémom pri tejto praktike je, že spomenuté hodnotenia nie sú vôbec založené na samostatne sformovanom názore používateľov ale len na umelo vytvorenom názore s účelom získať odmenu od danej spoločnosti. Spomínaná spoločnosť od roku 2016 diametrálne zmenila názor na tieto praktiky a aktívne blokuje účty predajcov používajúcich ich online priestor, u ktorých zistia pokusy motivovať zákazníkov k podobným falošným hodnoteniam. Táto spoločnosť sa taktiež vyjadrila, že v roku 2020 bolo na ich webstránke zablokovaných 200 miliónov recenzií s podozrením, že sú falošné [8], pričom ku zablokovaniu došlo ešte predtým než boli recenzie uverejnené, takže v podstate nemali žiaden negatívny vplyv na iných zákazníkov, keďže sa k ním nedostali.

Okrem recenzií, ktoré sa snažia používateľov presvedčiť, že produkty alebo služby sú reálne hodnotené veľmi pozitívne, existujú samozrejme aj cielené recenzie s úplne opačným zámerom, teda odradiť používateľov od istých produktov resp. značiek. Tieto recenzie môžu byť uverejňované s rôznymi motívmi. Napríklad môžu byť úzko spojené so spomenutým fenoménom online trollingu, kde môže a nemusí byť konkrétny úmysel poškodenia práve dotknutej značky, keďže trolling je pri

tejto problematike ešte nepredvídateľnejší a môže jednoducho ísť len o náhodnú obeť nenávistnej aktivity trollov. Zároveň ale môže dôjsť k cielenému útoku vo forme veľkého počtu negatívnych recenzií jasne zameraných na daný produkt alebo značku s cieľom odradiť nič netušiacich zákazníkov a naviesť ich takto ku kúpe podobného alebo identického produktu resp. služby u konkurencie. Z uvedených skutočností sa teda dá konštatovať, že ako falošné negatívne, tak aj falošné pozitívne hodnotenia majú nevhodné, v niektorých prípadoch až devastujúce následky nie len na zákazníkov ale aj na predajcov v online priestore a sú preto považované za formu antisociálneho správania na webe.

2.2.2. Falošné správy

Najviac známym a zároveň najaktuálnejším druhom antisociálneho správania na webe sú dezinformácie taktiež známe ako hoaxy, falošné správy alebo „fake news“. Táto diplomová práca je zameraná hlavne na túto kategóriu antisociálneho správania. Používateľ internetu a primárne sociálnych sietí sa počas bežného dňa stretne s obrovským množstvom informácií či už sú to správy, statusy a rôzne príspevky od iných ľudí na sociálnych sieťach alebo reklamy a mnohé iné kategórie informácií. Mnoho z týchto informácií je ale nesprávnych, nepravdivých a zavádzajúcich.

Falošné správy sa stali zdanlivo neoddeliteľnou súčasťou našej doby a dominujú v online priestore ako najškodlivejší druh antisociálneho správania. Vedci z Massachusetts Institute of Technology (MIT) [1] prišli svojou štúdiou na to, že falošné správy sa šíria rýchlejšie ako fakty, ktoré sú dokázateľne založené na pravde a to bez ohľadu na to o akú kategóriu príspevkov ide. Falošné správy sa šíria obzvlášť rýchlo na sociálnych sieťach.

Pri sociálnych sieťach je za týmto účelom veľakrát používaný napríklad Facebook, ale ešte viac badateľný je tento fenomén pri doméne s názvom Twitter. Twitter je mikroblogovacia sociálna sieť, ktorá umožňuje svojim používateľom uverejňovať krátke príspevky v textovej podobe s maximálnym rozsahom jednej správy obmedzeným na 280 znakov, takýto príspevok je známy ako „tweet“. Tento systém krátkych textových príspevkov je veľmi populárnym pre jeho jednoduchosť a nenáročnosť, keďže pri každom príspevku sa dá očakávať rýchle prečítanie a prísun informácií resp. ako sa čoraz častejšie preukazuje dezinformácií. Twitter taktiež obsahuje možnosť zdieľania príspevkov, takzvané „retweetovanie“, ktoré umožňuje príspevok zdieľať v jeho pôvodnej podobe alebo ku nemu pridať vlastné vyjadrenie, takzvaný „quote tweet“. Navyše pri použití populárnych kľúčových slov, ktoré sú na sociálnych sieťach známe ako „hešteg“ a spravidla majú označenie mriežky „#“ , je možné príspevky výrazne zviditeľniť aj keď účet, ktorý stoji za príspevkom nemá veľa sledovateľov. Mnoho používateľov sociálnych sietí sleduje okrem konkrétnych účtov aj heštegy, ktoré reprezentujú rôzne témy ako napríklad šport, politiku, ale reflektujú aj aktuálne

dianie napríklad voľby a vojnové konflikty. Preto je pre účty šíriace dezinformácie pomerne jednoduché dostať sa k bežnému používateľovi, keďže im stačí v uverejnenom príspevku použiť populárny hešteg a tým sa im rapídnym spôsobom zvýši šanca na zhliadnutie ich dezinformácií. Falošné správy majú oproti reálnym správam podľa štúdie o 70% väčšiu pravdepodobnosť na to, že budú retweetované, čo dokazuje akútnosť tejto problematiky .

Takéto falošné príspevky obsahujú mnohokrát šokujúco alebo neuveriteľne znejúcu správu napríklad v názve článku, ktorý zdieľajú, kde cieľom je hlavne prilákať pozornosť ľudí pomocou takzvaného „clickbaitu“ a následne ich naviesť k retweetovaniu takejto správy, ktorá pôsobí zaujímavo alebo šokujúco.

Dezinformácie sa šíria aj pomocou automatizovaných účtov nazývaných ako „boti“, ktorí slúžia na rozšírenie týchto dezinformácií medzi čo najviac používateľov. Takáto automatizácia uverejňovania dezinformácií ešte viac urýchľuje rozšírenie týchto príspevkov a komplikuje boj, ktorý je proti nim vedený. V štúdii od vedcov z MIT [1] sa ale ukázalo, že hlavnými šíriteľmi dezinformácií ostávajú naďalej ľudia, teda reálne neautomatizované účty, keďže pri odstránení botov z dátovej množiny príspevkov, ktorú skúmali sa pomer šírenia falošných správ ku faktom nezmenil. To značí, že šírenie zapríčiňujú hlavne používatelia sami, či už vedome samotným vytváraním a zdieľaním príspevkov tohto charakteru alebo nevedomým zdieľaním bežnými používateľmi, ktorí sú v danej téme nedostatočne informovaní.

2.3. Dôvody tvorby dezinformácii na webe

Pri existujúcom veľkom množstve dezinformácií na webe nastáva otázka, čo je samotnou motiváciou za tvorbou dezinformácií a prečo ich tvorcovia tak veľmi potrebujú virálne šírenie. Motiváciu na ich tvorbu a šírenie môžu tvoriť rôzne dôvody.

2.3.1. Monetizácia dezinformačných webov

Pri dezinformáciách vo forme článkov ide veľakrát o monetizáciu, čo znamená, že autori článku majú na stránkach kde sa tento článok nachádza reklamy, ktoré pre nich znamenajú zisk. Odtiaľ prúdi ich motivácia, keďže čím viac ľudí klikne na článok a uvidí uverejnenú reklamu tým viac peňazí to znamená pre autorov článku. Prepojenia na dezinformačné weby mnohokrát kopírujú názvy od tradičných overených správ s cieľom ešte viac zmiasť používateľov a pôsobiť ako dôveryhodné weby s faktickými správami a dobrou reputáciou, ktoré sú verejne známe.

2.3.2. Dezinformácie s politickou agendou

Dezinformácie majú mnohokrát aj politické ciele, konkrétne môže ísť o vytváranie dezinformácii o politikoch, politických stranách alebo celej politickej skupine s cieľom ich diskreditácie. V štúdii

[2] zameranej na skúmanie šírenia článkov s nízkou dôveryhodnosťou pomocou botov na Twitteri bolo analyzovaných 14 miliónov správ z roku 2016, pričom dokopy obsahovali prepojenia na 400 tisíc článkov. Zistenie tejto analýzy bolo, že počas prezidentských volieb v USA, ktoré prebiehali pravé v roku 2016 boli automatizované účty na Twitteri, čiže boti, v začiatočných fázach volieb zodpovední za šírenie dezinformácií a to hlavne tým, že sa snažili dezinformácie šíriť medzi vplyvných ľudí tým, že ich na Twitteri označovali vo svojich príspevkoch s prepojením na články s dezinformačným charakterom. Je samozrejmé, že označovanie týchto populárnych účtov malo podobný efekt ako použitie heštegov a teda ešte viac pomáhalo so šírením týchto správ, keďže príspevok, ktorý označuje verejne známu osobu sa zobrazí väčšiemu množstvu používateľov.

Pri podobnej štúdií [3] bol skúmaný dataset, ktorý obsahoval 171 miliónov Tweetov zhromaždených počas trvania volieb v USA v roku 2016. Vyselektované boli Tweety, ktoré obsahovali odkaz na správy bez ohľadu na to či sa jedná o fakt alebo dezinformáciu. Medzi takmer 31 miliónmi Tweetov, ktoré obsahovali odkaz na správy sa zistilo, že približne 10% z nich obsahuje odkaz na dezinformačné weby alebo konšpiračné teórie a ďalších 15% z nich obsahuje odkaz na správy, ktoré sú extrémne zaujaté, niektoré by sa dali označiť aj ako propaganda. Zo spomenutého experimentu je taktiež možné vydedukovať, že tradičné overené správy na Twitteri uverejňujú overení autori, teda účty so symbolom overeného účtu, zatiaľ čo dezinformácie sa šíria hlavne cez neoverené resp. nové účty, ktoré sú s veľkou pravdepodobnosťou vytvárané iba pre účel zdieľania a vytvárania dezinformácií, veľa z týchto účtov je taktiež automatizovaných.

2.4. Následky dezinformácii

Veľkým problémom dezinformácií je ich vplyv a ich dopady na jednotlivcov ale aj spoločnosť ako takú. Falošné správy vplývajú veľmi výrazným spôsobom na svet, v ktorom žijeme. Mnoho používateľov si neoveruje informácie a všetko s čím sa na internete stretnú považujú automaticky za fakty. To vedie ku vzniku rôznych hnutí ako napríklad „Flat Earth Society“ (FES) alebo „Anti-vax movement“ (AVM). Pri FES spočíva problém v tom, že šíria nepravdivé informácie, ktoré sa maskujú ako vedecké fakty. Dôsledkom toho sa znižuje dôvera v média, vládu a inštitúcie ako je napríklad NASA a vznikajú rôzne konšpiračné teórie, ktoré ohrozujú demokratickú spoločnosť ako takú. AVM je skupina ľudí, ktorí sú všeobecne proti očkovaniu, pričom toto hnutie zakladá na dezinformáciách o očkovaní a rôznych konšpiračných teóriách, ktoré nie sú vedecky podložené. Toto hnutie má aj napriek absurdnej ideológii veľa členov, ktorí odmietajú očkovanie a svoju ideológiu aktuálne propagujú napríklad u mladých ľudí na sociálnych sieťach s cieľom presvedčiť ich neočkovať sa proti ochoreniu COVID-19.

2.4.1. Dezinformácie o medicíne

V júni roku 2020 bola vydaná publikácia [9] zameraná na skúmanie toho, ako počet rastúcich falošných správ na sociálnych sieťach ohrozuje zdravie ľudí resp. formuje verejnú mienku o zdraví v nesprávnom smere. Autori tejto publikácie identifikovali a analyzovali 1225 článkov obsahujúcich dezinformácie o ochorení COVID-19. Keďže bola publikácia vydaná v prvom polroku 2020 bola táto téma veľmi aktuálna a nová, jej aktuálnosť ale pretrváva aj do dnešného dňa. Príspevky spojené s týmto ochorením sú publikované každodenne a vo veľkých počtoch. Tento veľký počet pribúdajúcich príspevkov je veľmi prirodzený a pochopiteľný, keďže drvivá väčšina svetovej populácie je vystavená informáciám o tomto ochorení alebo samotnému ochoreniu a dlhé mesiace karantény viedli k jedinému spôsobu socializácie a to cez sociálne siete, čo viedlo k zvýšeniu počtu ľudí, ktorí čítajú príspevky spojené s týmto ochorením.

Fakt, že pandémia COVIDu-19 pretrváva mnoho mesiacov a jej úplne odstránenie je komplikované a zdĺhavé priviedol mnohých používateľov internetu k hľadaniu „alternatívnych zdrojov liečby“, ktoré vo väčšine prípadov predstavujú riziko a nie sú absolútne založené na výskume alebo faktoch z oblasti medicíny. Príspevky, ktoré propagujú tieto dezinformácie sú mnohokrát rovnako závažné ako samotné ochorenie, proti ktorému sa snažia bojovať. Generálny riaditeľ Svetovej zdravotníckej organizácie (WHO) sa k tejto téme vyjadril s obavami, keďže „nebojujeme len proti pandémii, bojujeme aj proti infodémii“ (Dr. Tedros Adhanom Ghebreyesus, 2020, MSC2020), čiže inač povedané proti dezinformáciám devastačného charakteru.

Iba v samotnom mesiaci marca 2020 [10] bolo priemerne pridaných 46 000 Tweetov, ktoré boli fakticky nesprávne alebo naviazané na dezinformácie. Okrem bežných používateľov Twitteru, ktorí poväčšine nemajú vysokú sledovanosť, pridávali ale dezinformačné príspevky aj celebrity, verejne známe osoby alebo dokonca aj svetoví lídri. Ako príklady dezinformácií by sa dali uviesť napríklad príspevky s tvrdením, že pitie horúcej vody, alkoholu alebo alifatického alkoholu známeho ako metanol sú odporúčané pre zabránenie resp. liečenie ochorenia COVID-19. Takéto tvrdenia môžu a ako sa ukázalo aj majú vplyv na verejnú mienku a spôsobili zdravotné problémy viacerým ľuďom, ktorí tieto inštrukcie považovali za pravdivé a nasledovali ich.

2.4.2. Dezinformácie podnecujúce zločiny

Ďalej existovali príspevky, ktoré obviňovali ľudí konkrétnej etnicity alebo viery z rozširovania tohto vírusu. Konkrétne obviňovanie Moslimských komunít v Indii [11], s čim bol spojený hešteg „#CoronaJihad“ na Twitteri s cieľom obviňovať týchto ľudí z cieleného šírenia tohto vírusu. V USA to zase boli rasistické útoky „anti-ázijskeho“ charakteru. Tie viedli k vlnám príspevkov, ktoré rasisticky profilovali Američanov ázijského pôvodu [12] a rovnako ich obviňovali zo šírenia

ochorenia COVID-19 v USA. Rovnako boli spájaní so samotným zavinením vzniku tohto vírusu, keďže sa s veľkou pravdepodobnosťou predpokladá, že pacient „0“ bol ázijského pôvodu. Tieto príspevky, ktoré by sa okrem dezinformácií dali bezpochybne označiť aj ako šikana, sú priamym útokom a okrem iného výrazne vplývajú na psychické zdravie dotknutých skupín. Príspevky tohto charakteru na sociálnych sieťach naberali na popularite a následne viedli aj k zvýšeniu fyzických útokov na Američanov ázijského pôvodu. Niektoré z týchto útokov [14] boli tak brutálne, že mali za následok smrť. Tieto a mnohé podobné prípady dokazujú ako veľmi je spoločnosť ovplyvniteľná príspevkami na sociálnych sieťach a ako ľahko vie podľahnúť dezinformáciám, ktoré majú reálne katastrofálne dopady na spoločnosť ale aj na konkrétne skupiny a jedincov.

2.4.3. Manipulácia spoločnosti

Dezinformácie sa taktiež overili ako veľmi dobrý nastroj manipulácie v politickej oblasti. Spoločnosť Cambridge Analytica, ktorá pôsobila konkrétne na sociálnej sieti Facebook, následkom obvinení a súdnych procesov zanikla v roku 2018. Podľa výpovede ex zamestnanca, [4] spoločnosť okrem iného využívala dezinformácie ako nástroj pre propagáciu politických kandidátov vo voľbách po celom svete. Dezinformácie smerovali na konkrétnych vyprofilovaných jedincov pomocou dátových bodov, ktoré na nich postupne vytvárali. Ich cieľom boli politicky neutrálni používatelia Facebooku, keďže bolo oveľa jednoduchšie presvedčiť politicky neutrálneho používateľa ako používateľa, ktorý bol už zásadne zástancom opozičnej politickej strany. Následne sa snažili týchto neutrálnych používateľov ovplyvniť a presvedčiť ich, že politická strana, pre ktorú robili kampaň je správna voľba, zatiaľ čo všetky ostatné strany alebo kandidáti sú nečestní, majú nesprávne ideológie a ciele alebo škandalóznu minulosť. Takáto propaganda je veľmi nebezpečná nie len z dôvodu, že ovplyvňuje integritu volieb ale aj z dôvodu, že vytvára politickú a ideologickú polarizáciu u veľkého množstva ľudí. Tieto ale aj veľké množstvo ďalších faktov poukazuje na to, že dezinformácie sú veľmi veľkým problémom tejto doby a majú dopady, ktoré vedia ovplyvniť veľké množstvo oblastí nášho života.

2.5. Overovania informácii na webe

Následkom extrémneho vzrastu dezinformácií na webe vznikli stránky na overovanie faktov ako napríklad FactCheck.org, Factama.com alebo PolitiFact.com, ktoré sa snažia bojovať proti falošným správam tým, že pomáhajú overovať či sa jedná o pravdivú alebo nepravdivú informáciu. Nedostatkami takýchto fact-checking stránok [5] je ale to, že veľké množstvo z nich je zameraných hlavne na overovanie správ spojených s politikou, čo znamená, že pre všetky ostatné druhy dezinformácii je ich použitie veľmi obmedzené. To predstavuje dosť závažný problém, keďže dezinformácie rozhodne nekončia len pri politicky orientovaných témach.

Ešte väčším problémom je ale fakt, že tieto stránky sú založené a odkázané na manuálnu kontrolu správ. Táto manuálna kontrola je veľmi neefektívna, keďže zaberá veľmi veľa času a navyše je doslova nemožné týmto prístupom overovať informácie v reálnom čase. To predstavuje veľký problém, keďže obrovské množstvo informácii pribúda, je zdieľaných a prečítaných používateľmi každú sekundu a kým tieto stránky na overovanie faktov stihnú zareagovať veľké množstvo ľudí zatiaľ interaguje s dezinformačným príspevkom a je ním ovplyvnené.

Množstvo expertov ale aj technologických spoločností sa zhodlo na tom, že je nutné vytvoriť prístup, ktorý by dokázal rozpoznať dezinformácie. Je to ale veľmi komplikované, keďže informácie na internete majú dynamickú povahu čo znamená, že sa neustále menia. Dochádza k zmenám autorov dezinformácii, štýlu písania, témam, na ktoré sa zameriavajú atď. Napríklad počas obdobia volieb vznikajú falošné správy zamerané prevažne na propagandu v politickej sfére no ihneď po voľbách sa začnú orientovať na ďalšiu úplne inú tému, keďže voľby v tom čase už nie sú naďalej aktuálne. Tieto náhle zmeny predstavujú pre detekciu dezinformácií najväčšiu výzvu.

2.5.1. Nedôvera v overovanie faktov

Ďalším problémom tohto prístupu je samotná nedôvera vo „fact checking“ stránky. Tieto stránky poskytujú overené informácie pomocou vyhľadávania primárnych a sekundárnych zdrojov a taktiež poukazujú na dezinformačné články a zavadzajúce verejné vyhlásenia, ale dôvera resp. nedôvera v samotný fact checking je ďalším fenoménom online priestoru.

Publikácia z roku 2017 [13] s názvom „Dôvera a nedôvera v online služby pre overovanie faktov“ sa venovala práve otázke, či a ako sa názory a postoje ľudí menia v reakcii na skutočnosti, ktoré sú v rozpore s ich už existujúcimi názormi. Je relevantným predpokladom, že mnoho používateľov, ktorí sú roky angažovaní v skupinách propagujúcich napríklad konšpiračné teórie nezmení svoj názor ani vtedy, ak im bude preukázané, že ich názor je založený na dezinformáciách. Výskum v spomínanej publikácií niesol teda predpoklad, že služby poskytujúce overovanie faktov môžu byť potenciálne neúspešné pri znižovaní dezinformácii a to hlavne medzi ľuďmi, ktorí týmto dezinformáciám verili už doteraz. „Ľudia často ignorujú fakty, ktoré sú v rozpore s ich súčasným presvedčením, najmä v politike a kontroverzných spoločenských otázkach“. Experiment, ktorý bol v tomto výskume použitý nezhromažďoval spätnú väzbu pomocou dotazníka alebo prieskumu, namiesto dotazníka bol využitý analytický nástroj „Meltwater Buzz“. Tento nástroj umožnil zhromažďovať dáta priamo z online konverzácii, konkrétne z blogov, diskusných fór, online spravodajských stránok ale aj sociálnych sietí konkrétne z Twitteru a Facebooku. Tento pristúp bol zvolený z dôvodu, že [13] „konverzácie v sociálnych médiách predstavujú vysoko relevantný zdroj údajov, keďže pravdepodobne odrážajú surové, autentické vnímanie používateľov sociálnych médií“. Výsledkom

tohto experimentu bolo, že dve z troch fact-checking webstránok, o ktorých používatelia hovorili boli vnímané väčšinovo negatívne a teda jedna z týchto webstránok bola vnímaná pozitívnejšie. Pozitívne názory boli zamerané na užitočnosť týchto služieb, zatiaľ čo negatívne názory v príspevkoch sa týkali skôr dôveryhodnosti týchto služieb a nie ich samotnej užitočnosti. Negatívne hodnotenia komentujúcich vyjadrovali obavy z integrity týchto webových služieb a v niektorých prípadoch ich považovali za zaujaté, konkrétne zaujaté k politickej ľavici.

2.5.2. Zhodnotenie overovacích webov

Záverom je, že tieto služby nie sú ideálne pre dynamické prostredie online priestoru, keďže sú z pochopiteľných dôvodov pomalé a taktiež je ich vnímanie medzi niektorými používateľmi pomerne negatívne. Preto vzniká potreba pre iné prístupy, napríklad automatizované procesy resp. algoritmy, ktoré by dokázali odhaliť dezinformácie ešte predtým, než sa stihnú rozšíriť medzi väčšie množstvo používateľov a tieto evidentne dezinformačné príspevky zablokovať.

2.6. Sloboda slova v online priestore

Pri problematike dezinformácií vzniklo mnoho názorov na to, ako k dezinformáciám na webe pristupovať ale aj či je etické a morálne priamo zasahovať reguláciou resp. odstraňovaním takýchto príspevkov. Zachovanie autentickej a nekontrolovanej komunikácie je demokraticky pôsobiaci prístup ale zároveň je aj veľmi nebezpečný, keďže môže napomáhať rozkladu samotnej demokracie. Sloboda slova je veľmi podstatné ľudské právo ale taktiež by pravdepodobne malo mať isté hranice. Úplne neregulované sociálne siete prespievajú k chaosu ako bolo doteraz v opísaných prípadoch preukázané.

2.6.1. Autentická komunikácia

V akademickom časopise s názvom „European Journal of Communication“ bol publikovaný v roku 2020 článok [15], ktorý sa zaoberal práve myšlienkou, či je správne regulovať sociálne siete priamym zasahovaním do príspevkov. Facebook v roku 2017 vydal vysvetlenie, v ktorom sa hovorilo o tom, ako svoje algoritmy upravili tak, aby sa na nástenke s novinkami ľuďom zobrazovala „autentická komunikácia“. Algoritmy predvídali a hodnotili v reálnom čase, kedy môžu byť pre používateľa príspevky relevantnejšie, čo mal byť istý spôsob boja proti falošným správam bez priameho zasahovania do príspevkov. Vyjadrenie spoločnosti Facebook znelo nasledovne „Jednou z našich hodnôt pre „News Feed“ je autentická komunikácia. Počuli sme od našej komunity, že autentické správy sú tie, ktoré s nimi najviac rezonujú, teda tie, ktoré ľudia považujú za skutočné a nezavádzajúce“. Facebook teda upravil svoje algoritmy podľa predpokladu, že autentické správy sú úzko naviazané na reálne, faktické správy. To vedie k dôsledku, že ak tento predpoklad nie je stopercentný a medzi algoritmom vyhodnotené autentické správy sa dostane aj dezinformačný

príspevok (napríklad pre jeho popularitu a veľkú angažovanosť s príspevkom) zobrazí sa mnohým používateľom dezinformácia bez toho aby bola akokoľvek regulovaná. Napríklad ak veľké množstvo používateľov zdieľa a interaguje s príspevkom, ktorý je dezinformačnej povahy a používatelia reálne veria informáciám, ktoré príspevok zobrazuje nebudú ho nahlasovať a teda algoritmus ho potenciálne vyhodnotí ako autentický čo bude mať za následok ešte väčšiu angažovanosť s príspevkom.

2.6.2. Varovania pred dezinformáciami

Facebook sa taktiež rozhodol odstrániť výrazné varovania pri potenciálne dezinformačných príspevkoch známe ako „Disputed Flags“. Produktová manažérka pre spoločnosť Facebook publikovala vyjadrenie [16], v ktorom je tento krok odôvodnený tým, že akademický výskum z oblasti opravy dezinformácii naznačuje, že použitie výrazného symbolu ako napríklad červenej varovnej vlajky pri dezinformačných príspevkoch môže ešte viac posilniť používateľov vo viere v ich relevantnosť. Teda varovné symboly by mohli mať presne opačný efekt ako je ich zámer. Ak používatelia, ktorí majú isté presvedčenia uvidia príspevok potvrdzujúci ich presvedčenia ale zároveň je označený ako nepravdivý, bude to mať na nich práve spomínaný opačný efekt. Budú teda s veľkou pravdepodobnosťou ešte viac veriť v svoje predmetné presvedčenia a varovný symbol pri príspevku budú vnímať skôr ako „sprisahanie“ proti ich presvedčeniam alebo spôsob utláčania ich presvedčení. Iným problémom pri tomto označovaní príspevkov bolo, že mnoho používateľov vnímalo príspevok, ktorý nemal pri sebe varovný symbol automaticky ako pravdivý a faktický čo je nesprávny a nebezpečný predpoklad.

2.6.3. Regulácia príspevkov

Je teda zrejmé, že každý prístup ku boju proti dezinformáciám na sociálnych sieťach má svoje nedostatky a spôsob akým ku tomuto boju pristupovať je na rozhodnutí každej platformy. Pokus Facebooku zachovať takmer nekontrolovanú platformu mal veľké komplikácie a v posledných mesiacoch bola platforma nútená k ráznejším krokom ako napríklad vytvorenie „COVID-19 infocentra“, ktoré poskytuje overené a aktuálne informácie o pandémii, zatiaľ čo dezinformačné alebo nepresné príspevky resp. profily, ktoré tieto príspevky šíria sú čoraz častejšie suspendované či už dočasne alebo permanentne. Platforma Twitter bola taktiež nútená k ráznejším krokom, najviac kontroverzne vnímaným krokom je pravdepodobne permanentné odstránenie účtu bývalého prezidenta USA Donalda Trumpa [17], keďže jeho príspevky Twitter označil za „podnecujúce k násiliu a spochybňujúce demokratické voľby“, pre jeho vyjadrenia o volebnom podvode v posledných prezidentských voľbách. V niektorých prípadoch je teda pomerne ťažké určiť hranicu medzi slobodou slova a šírením falošných správ. Evidentné ale je, že hranica musí existovať, keďže jej absencia vedie k chaosu.

3. Analýza stavu problematiky

V tejto časti práce sú definované dôležité pojmy a popísané rôzne aktuálne využívané prístupy k detekcií antisociálneho správania na webe, teda využívané metódy, modely, ich výhody ale aj nedostatky v detekcií.

3.1. Dátové prúdy

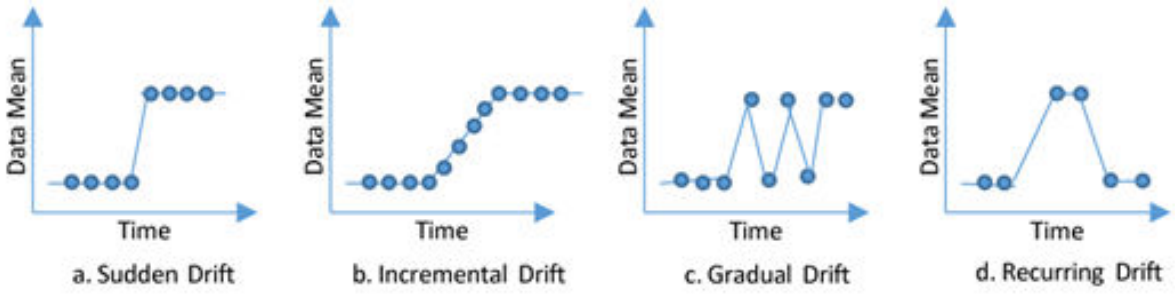
Všeobecne sa pojem dátový prúd stal veľmi rozšírený a je spomínaný v rôznych oblastiach. Všetky dáta prenášané cez internet sa teoreticky dajú označiť ako prúd dát. Pojem dátový prúd (ang. data stream) v oblasti dolovania dát a strojového učenia je všeobecne definovateľný ako nepretržitá a zároveň meniaca sa sekvencia dát, ktorá vstupuje do systému, kde sa má uložiť alebo nejakým spôsobom ďalej spracovať. Dátové prúdy môžu byť potenciálne nekonečné, [36] keďže dáta vstupujúce do streamu môžu neustále pribúdať čím sa stream ďalej predlžuje. Oproti tradičným spôsobom dolovania dát sa streamy líšia práve faktom, že sú potenciálne nekonečné a ich veľkosť je potenciálne neobmedzená. Limitáciou je dostupná veľkosť pamäte, pri ktorej nie je možné donekonečna ukladať pribúdajúce elementy streamu, preto streamy po spracovaní elementu resp. príspevku nezachovávajú daný element a vymažú ho alebo zachovajú iba sumarizáciu elementu. Streamy v oblasti objavovania znalostí resp. dolovania dát zahŕňa najčastejšie techniky klasifikácie, zhlukovania a asociačného dolovania. V poslednej dobe začala vznikať resp. narastať iniciatíva využívať dátové streamy pre úlohy falošných správ alebo iného antisociálneho obsahu podobného typu. Táto iniciatíva je narastajúca z dôvodu, že je čím ďalej tým viac výrazné, že štúdie s cieľom klasifikácie falošných správ pomocou statických metód nie sú naďalej prínosné a chýba im aspekt dynamického prostredia, keďže v online prostredí sú všetky príspevky rýchlo sa meniace a postupne pribúdajúce. V tejto diplomovej práci sú využívané streamy pre úlohu klasifikácie, konkrétne klasifikácie antisociálneho správania. Keďže k dispozícií sú statické datasety obsahujúce označkované príspevky spojené s antisociálnym správaním, je nutné ich pretvoriť na dynamické a nasimulovať tak postupné pribúdanie príspevkov. To je docielené práve pomocou streamov, ktoré umožňujú načítať celý dataset, pričom je nutné streamu určiť čo sú vlastnosti a čo ciele. Vlastnosti sú v tomto prípade samotné texty príspevkov a jediným cieľom je cieľový atribút, teda označenie či ide alebo nejde o antisociálne správanie.

3.2. Konceptové drifty a ich detekcia

Pri technikách využívajúcich dátové prúdy sú pravdepodobne najväčšou prekážkou zmeny v dátach nastávajúce v závislosti od času. Tieto zmeny v dátach častokrát vedú pri neadaptívnych prediktívnych modeloch k zhoršujúcej sa efektivite. V oblasti strojového učenia sa tento jav nastávajúci pri niektorých zmenách v dátach nazýva konceptový drift. Pri mnohých známych

prediktívnych modeloch je postup založený na učení modelu z historických dát [37] a následne použitie modelu pre predikciu z nových doposiaľ neznámych dát. Tento prístup sa dá označiť ako statický, teda existuje predpoklad, že mapovania, ktoré sa model naučil z historických dát budú platné a efektívne aj pri nových dátach. Tento predpoklad môže byť správny pri niektorých problémoch, no rozhodne nie pri všetkých. Pri problematike klasifikácie textových príspevkov, hlavne príspevkov pochádzajúcich zo sociálnych sietí, je veľmi pravdepodobné a pomocou niekoľkých štúdií aj dokázané, že v takýchto dátach dochádza k zmenám, ktoré potenciálne vedú k vzniku konceptového driftu. Tieto zmeny tým pádom oponujú predpokladu statických prístupov a po nastaní zmien sú neadaptívne modely odkázané na stratu v efektivite a stávajú sa následkom konceptového driftu zastaralé. Konceptový drift predstavuje prekážku nie len z dôvodu jeho samotného nastania, ale aj pre jeho rôzne formy resp. typy, ktoré nie je možné predikovať. Zadefinované sú štyri typy konceptového driftu [23] podľa závažnosti driftu a rýchlosti jeho nastatia teda rýchlosti zmien. Drift s opakovaním sa pritom dá považovať za systematický, čo umožňuje jeho jednoznačnú identifikáciu resp. prípravu na jeho nastanie v budúcnosti.

- Náhly drift, v ktorom sa objavuje nový koncept v krátkom časovom období
- Inkrementálny drift, v ktorom sa objavuje starý koncept, ktorý sa postupom času inkrementálne vyvinie v nový koncept
- Postupný drift, v ktorom sa objavuje nový koncept, ktorý sa postupom času zmení na starý koncept
- Drift s opakovaním, v ktorom sa po istom čase môže zase zopakovať starý koncept



Obrázok 1 Typy konceptového driftu. Prevzatý z publikácie:“ Detecting Different Types of Concept Drifts with Ensemble Framework“ [23]

Konceptový drift predstavuje prekážku [37] napríklad pri časových radoch alebo profilovaní zákazníkov, keďže v správaní sa zákazníkov pri nakupovaní môže dôjsť k rôznym zmenám v závislosti od ekonomickej situácie, pracovnej situácie a rôznych ďalších nepredvídateľných faktorov, no najčastejšie sa vyskytuje pri dátových prúdoch teda streamoch, keďže v nich postupom času dochádza k zmenám, ktoré môžu konceptový drift vyvolať. Fakt, že je potenciálny výskyt

konceptového driftu v dátových prúdoch známy a dá sa predpokladať, že v nich nastane, vyvoláva potrebu včas ho odhaliť.

Odhalenie týchto driftov je možné pomocou detektorov driftu, ktoré sú zároveň využívané aj v experimentoch tejto diplomovej práce. Medzi využívané detektory driftu patria nasledujúce algoritmy [38]:

- ADWIN (ADaptive WINdowing) je adaptívny algoritmus so schopnosťou zisťovať zmeny a udržiavať aktuálne štatistiky o dátovom prúde. ADWIN umožňuje, aby aj algoritmy, ktoré nie sú prispôsobené na driftujúce dáta, boli voči tomuto javu odolné. Uchováva štatistiky z posúvajúceho sa okna, ktoré má premenlivú veľkosť a zároveň pri tom deteguje konceptové drifty. Ak je pomocou tohto detektora detegovaný drift, všetky dáta pred bodom kedy bol detegovaný sú zahodené a nepoužívajú sa ďalej, čo prakticky zabezpečuje, že sa model využívajúci ADWIN začne po detekcií driftu prispôsobovať zmene.
- EDDM (Early Drift Detection Method) je detektor zameraný hlavne na detekciu postupného driftu, teda driftu v ktorom sa objavuje nový koncept, ktorý sa postupom času zmení naspäť na starý koncept. Táto metóda je založená na sledovaní štatistík a funguje tak, že sleduje priemerné vzdialenosti medzi dvoma chybami.
- HDDM_A (Hoeffding Drift Detection Method Average-test) detektor driftu s testom kĺzavého priemeru, založený na Hoeffdingovej nerovnosti, pričom využíva pre odhadovanie samotný priemer. Na vstupe príma stream reálnych hodnôt a na výstupe vracia odhadovaný status streamu, teda či je stabilný, vo varovnej zóne alebo driftujúci.
- HDDM_W (Hoeffding Drift Detection Method Weighted average-test) je detektor driftu založený na McDiarmidovej nerovnosti, pričom pre odhadovanie využíva štatistiky z exponenciálneho váženého kĺzavého priemeru (EWMA). Na vstupe príma stream reálnych hodnôt a na výstupe vracia odhadovaný status streamu, teda či je stabilný, vo varovnej zóne alebo driftujúci.
- KSWIN (Kolmogorov-Smirnov Windowing) je detektor driftu založený na štatistickom teste Kolmogorov-Smirnov (KS). Používa posúvajúce sa okno, ktoré je na rozdiel od ADWINu fixnej veľkosti počas celého streamu, teda okno nemení svoju veľkosť na základe štatistík zo streamu.
- Page-Hinkley detektor funguje na základe Page-Hinkleyho testu teda vypočítavania pozorovaných hodnôt a ich priemeru, pričom drift bude detegovaný, ak je pozorovaný priemer v určitom okamihu väčší ako prahová hodnota lambda.

3.3. Adaptívne modely

Ako bolo popísane, konceptové drifty spôsobujú pri statických modeloch postupom času problémy vyúsťujúce do zníženej efektivity. Pri modeloch, ktoré slúžia na klasifikáciu antisociálneho správania je doležíte aby tieto modely boli pripravené na konceptové drifty a teda aby sa vedeli adaptovať. Existujú rôzne adaptívne modely, ktoré využívajú detektor driftu pre zistenie kedy a v ktorom bode sa majú preučiť a tým pádom si zachovať efektivitu aj napriek zmenám v dátach. Medzi takéto modely, ktoré sú využívané aj v experimentoch tejto diplomovej práce patria napríklad:

- Hoeffding Adaptive Tree je adaptívna verzia algortimu Hoeffding Tree, pričom ide o inkrementálny algoritmus rozhodovacieho stromu, ktorý je schopný sa učiť z veľkých dátových prúdov. Tento algoritmus využíva skutočnosť, že aj malá vzorka môže častokrát stačiť na výber optimálneho atribútu delenia. Tento predpoklad je matematicky dokázateľný pomocou Hoeffdingovej väzby, ktorá vyčísluje počet pozorovaní (teda v našom prípade počet príspevkov) potrebných na odhadnutie štatistiky v rámci istej presnosti (v našom prípade kvalita atribútu). Adaptívna verzia využíva pre detekciu driftov výlučne ADWIN metódu, pričom zároveň monitoruje efektivitu jednotlivých vetiev a nahradí ich novými vetvami, keď sa ich presnosť starých vetiev zníži a zároveň ak sú nové vetvy presnejšie.

- KNNADWIN je klasifikátor, ktorý je vylepšením bežného klasifikátora KNN. Tento vylepšený algoritmus je odolný voči konceptovému driftu a využíva ADWIN detektor driftu pre rozhodovanie toho, ktoré záznamy si má ponechať a ktoré má zabudnúť. Týmto krokom taktiež reguluje veľkosť okna záznamov. Hlavným rozdielom je, že má okno s premenlivou veľkosťou namiesto okna s pevnou veľkosťou a tiež aktualizuje ADWIN po každom inkrementálnom prispôsobení sa modelu.

- LPPNSE je klasifikátor pre inkrementálne učenie z nestatických, číže z dynamických prostredí známych ako Non-stationary environments (NSE), kde sa údaje v priebehu času menia. Tento klasifikátor sa učí z dávok údajov, pričom v týchto údajoch môže dochádzať ku konštantnému alebo nepravidelnému počtu driftov na čo je tento klasifikátor pripravený. Tento klasifikátor obsahuje taktiež nastavenia stavané na zvýšenie odolnosti voči cyklickému driftu. Taktiež je možné ho kombinovať s inými modelmi teda aj s ADWINom, čo ešte viac vylepšuje efektivitu tohto klasifikátora.

- SRP metóda kombinuje bagging a random subspaces, čo dokopy vytvára „random patches“ podľa čoho je aj táto metóda nazvaná. Klasifikácia pri tejto metóde je prispôsobená

vyvíjajúcim sa dátovým prúdom. SRP metóda zároveň umožňuje využívať rôzne detektory driftu, čo bolo pri experimentoch využívané na porovnávanie efektivity detektorov.

3.4. Prístupy k detekcii dezinformácii

S preukázaním očividných nedostatkov prístupu fact-checking stránok v boji proti dezinformáciám a príkladoch s reálnymi dopadmi dezinformácií na spoločnosť je jasná potreba regulácie dezinformácii na webe. V tejto komplexnej úlohe rastie neustále snaha vyvinúť prístup, ktorý by čo najefektívnejšie a najlepšie identifikoval dezinformácie na webe bez potreby manuálnej kontroly príspevkov. Do úvahy teda pripadá strojové učenie resp. modely, algoritmy a prístupy strojového učenia, ktoré by dokázali tieto príspevky identifikovať. Okrem iného v experimente [18], v ktorom sa ľudia bez špeciálneho školenia alebo tréningu snažili rozlíšiť pravdivé informácie od nepravdivých bolo výsledkom, že iba 54% príspevkov bolo identifikovaných ľuďmi správne, teda metódy strojového učenia majú v tomto rozlišovaní značnú výhodu oproti ľuďom.

3.4.1. Hlboké učenie

Pre účely detekcie dezinformácií ale aj iného antisociálneho správania v textovej podobe je možné používať prístupy hlbokého učenia. Prístupy používajúce na detekciu antisociálneho správania z textov hlboké učenie [34] pracujú výlučne z textom teda obsahom správ, pričom predspracovanie týchto textov nie je výrazné. Pri predspracovaní sa používa oddeľovanie viet, teda oddelenie každej vety od nasledujúcich viet s účelom vytvoriť jednotlivé záznamy, ak je tento krok nutný. Môže sa taktiež použiť odstránenie stop slov. Pri detekcií falošných správ alebo napríklad detekcií nenávistných príspevkov je dôležité zachovať kontext a pôvodný úmysel príspevkov, teda je dôležitá pôvodná podoba textu.

Pre detekciu antisociálneho správania na webe v oblasti hlbokého učenia sa používajú najčastejšie modely založené na neurónových sieťach ako Convolutional Neural Networks (CNN), Recurrent neural networks (RNN), Gated recurrent unit (GRU) alebo Long short-term memory (LSTM). CNN modely sú v tejto problematike výhodné, keďže sú rýchle a výsledkami priemerne najlepšie [34], napríklad oproti RNN modelom sú CNN modely až 5 krát rýchlejšie.

Pri niektorých štúdiách [19] sú využívané modely hlbokého učenia pre detekciu falošných správ v kombinácií s pred-trénovaním modelu na korpuse skúmanej témy. Tento prístup môže vykazovať veľmi dobré výsledky, jeho nedostatkom je ale nutnosť pred-trénovania na korpuse konkrétnej témy, teda na zmeny tém resp. adaptáciu na tieto zmeny nie je daný model stavaný. Pri takomto prístupe je teda reálne nasadenie modelu nepravdepodobné, keďže je zameraný iba na jednu tému, ktorú bol pomocou korpusu pripravený klasifikovať.

3.4.2. Nedostatky klasických prístupov

Mnoho z doteraz navrhnutých prístupov uvažuje nad detekciou falošných správ ako nad tradičnou úlohou dátovej analýzy, pri ktorej sa vôbec neberie ohľad na to, že tieto dáta nie sú statického ale dynamického typu, [6] keďže sú to dátové prúdy. Príspevky na sociálnych sieťach sa neustále menia či už témou, o ktorej hovoria alebo štýlom akým sú písané. Ďalším nedostatkom pri týchto klasických prístupoch je, že neberú do úvahy vplyv týchto zmien tém a štýlu príspevkov, pričom práve tieto zmeny môžu viesť k vzniku konceptového driftu. Konceptový drift znamená, že vlastnosti cieľového atribútu, ktoré sa model snaží predpovedať sa časom zmenia nepredvídaným spôsobom a teda čo platilo pred istým časom nemusí vôbec platiť v súčasnosti. To má za následok, že tradičné modely, ktoré sú zamerané na statické dáta síce na začiatku môžu vykazovať dobré niekedy až nadpriemerne dobré výsledky v detekcii dezinformácii, ale pri zmene tém o ktorých sa hovorí, autorov príspevkov, štýlov písania a rôznych iných zmenách sa ich presnosť detekcie výrazne znižuje, keďže sa na tieto zmeny nevedia adaptovať.

V roku 2020 vznikol experiment [6], v ktorom sa ako v jednom z prvých experimentov brala do úvahy dynamická povaha dát a dátové prúdenie. V experimente využívali Apache Kafka systém a takzvaný „Complex Event Processing engine“. Hlavným predmetom tejto štúdie bolo analyzovať prístupy používané na extrakciu vlastností pre potrebu klasifikácie dátových streamov. Pre klasifikačné modely v experimente bol vytvorený dataset, ktorý pozostáva z 26 000 záznamov pričom 50% bolo dezinformačného charakteru a ostatných 50% boli faktické správy. Tento dataset bol potom prispôsobený na streamovanie, aby sa docielila simulácia reálneho toku správ. Následne boli použité tri prístupy spracovania a to Streaming Ensemble Algorithm (SEA) ,Online bagging (OB), Single model (SM), pričom SEA a OB sú zložené prístupy a SM je jednomodelový prístup. Algoritmy ktoré v experimente použili na klasifikáciu boli: Gaussian Naive bayes, Multi-layer Perceptron, Hoeffding Tree s použitím Naive Bayes Addaptive prediction mechanism

Najpresnejší výsledok bola presnosť 0.8, ktorú dosiahol Multi-layer Perceptron v kombinácii s PCA. Na rozdiel od experimentu popísaného v bode 3.4.1 pracoval tento experiment s tokom resp. prúdom dát čo simuluje reálne okolnosti na sociálnych sieťach. Tento prístup je dôležité preskúmavať ďalej, keďže príspevky sa neustále menia a pribúdajú. Aj keď sa môže zdať, že tento prístup je nepresnejší, keďže najvyššia dosiahnutá presnosť je nižšia ako pri predošlom experimente, pravdou je, že je rovnako dôležitý a dosiahnutá presnosť je pri zohľadnený dátového prúdu dobrá.

3.4.3. Doménovo-adaptívny prístup

Ako bolo spomenuté online prostredie je veľmi dynamické a rýchlo sa meniace. Postupom času dochádza ku rôznym zmenám v príspevkoch, komentároch, autoroch príspevkov atď. Preto je rozumné uvažovať v tejto problematike nad metódami a prístupmi, ktoré by sa vedeli prispôsobovať na rôzne zmeny a zároveň vykazovať dobrú efektivitu detekcie falošných správ. Jeden z najnovšie popísaných adaptívnych prístupov je „Reinforced Adaptive Learning Fake News Detection“ (REAL-FND), tento prístup bol popísaný v nedávnej publikácii [20], konkrétne z Februára roku 2022.

Hoci mnohé modely hlbokého učenia boli navrhnuté na odhaľovanie falošných správ a niektoré z týchto modelov ukázali dobré výsledky, väčšina z nich je použiteľná len na jednu doménu teda na doménu, na ktorú boli modely trénované. Adaptívny model REAL-FND využíva zaujímavý a pomerne jedinečný prístup, kde pracuje s generalizovanými teda zovšeobecnenými doménovo nezávislými vlastnosťami na rozlíšenie falošných správ od reálnych. V iných metódach sa „reinforcement learning“ (RL) používa na modifikovanie parametrov modelov.

V REAL-FND sa to ale líši tým, že sa nemodifikujú pomocou RL parametre modelu, ale modifikuje sa reprezentácia, ktorá je modelom naučená a to tak, že doménovo špecifické vlastnosti sa nepoužijú, zatiaľ čo doménovo nemenné vlastnosti sa zachovajú. To v podstate znamená, že vlastnosti, ktoré by model naučili príliš špecifické pravidla sú ignorované a model sa učí iba z vlastností, ktoré platia všeobecne. Z klasifikátora, ktorý by bol efektívny iba pri jednej téme falošných správ sa teda potenciálne stane univerzálny resp. v istom zmysle adaptívny model pre detekciu falošných správ. Pri využití tohto prístupu sa dokázalo, že v porovnaní s doteraz používanými modelmi sa REAL-FND vie lepšie adaptovať na nové domény teda je adaptívnejší a pri použití modelu v rámci jednej konkrétnej domény vykazuje podľa metrík vysoký výkon teda je efektívny aj v takomto prípade.

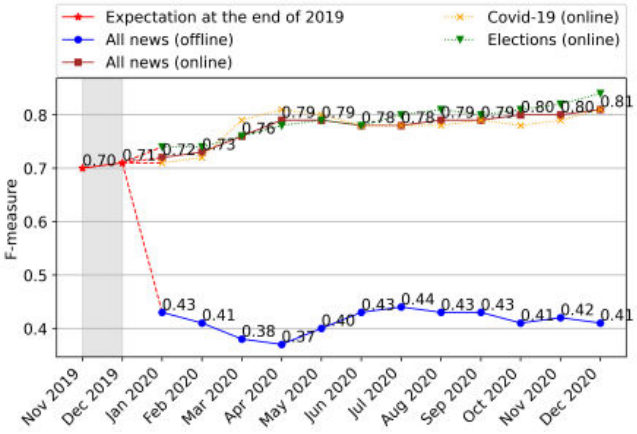
3.4.4. Detekcia nepodložených správ z dátových prúdov

Mnoho príspevkov na sociálnych sieťach by sa dalo označiť za fámy, teda príspevky, ktoré nesú informáciu, ktorej obsah nie je zatiaľ známy ani ako reálna správa ale ani ako falošná správa. Šírenie takýchto správ, ktoré majú tendenciu neskôr získať status falošnej správy sa taktiež dá pokladať za antisociálne správanie na webe, keďže to je v istom zmysle pod-úroveň samotných falošných správ. Na detekciu takýchto nepodložených správ bol navrhnutý model s názvom „CanarDeep“ [21], ktorý deteguje príspevky tohto typu v reálnom čase teda simuluje online dátové prúdy . Pre model boli použité dva rôzne typy vstupov a to vstupy kontextového typu a vstupy používateľského typu. Medzi vstupy používateľského typu patria napríklad vek používateľa ktorý uverejnil príspevok,

overený/neoverený status používateľa, celkový počet Tweetov používateľa. Medzi kontextové vstupy patria napríklad čiarky v príspevku, počet slov, symbol otáznika alebo výkričníka atď. Tieto vstupy boli použité spolu s klasifikátormi, (konkrétne HAN klasifikátor pre kontextové vstupy a MLP pre používateľské vstupy) výstupné predikcie týchto dvoch klasifikátorov boli skombinované pre klasifikáciu samotných Tweetov. Vyhodnotenie tohto prístupu vykazuje lepšie výsledky oproti súčasne zaužívaným metódam v tejto problematike , konkrétne pri metrike F1-skóre je CanarDeep až o 4.45% lepší ako ostatné metódy. Tento prístup zároveň dokazuje, že pri detekcii fám z Tweetov pri použití dátových prúdov je potenciálne prínosné použiť okrem textu aj iné atribúty resp. ich kombinácie, ktoré môžu napomôcť zlepšeniu klasifikátora v detekcii fám.

3.5. Vplyv konceptového driftu na klasifikáciu falošných správ

Väčšina štúdii v oblasti detekcie falošných správ neuvažuje nad dynamickým vývojom príspevkov a dynamickou povahou správ. Pracujú s tradičnými statickými modelmi, pričom spoliehajú na to, že charakteristiky príspevkov sa postupom času nemenia. Je ale dôležité sa zamyslieť nad tým, či tieto štúdie svojimi prístupmi k detekcií falošných správ nezaostávajú za aktuálnym vývojom. V dôsledku veľkého množstva štúdií používajúcich statické metódy vznikla v roku 2021 nová štúdia [22], ktorá sa venovala práve hypotéze, či si tradičné neadaptívne metódy zachovajú presnosť predikcie falošných správ aj pri zmenách charakteristík resp. tém príspevkov. V tejto štúdii boli použité dáta z rokov 2019 a 2020 s cieľom porovnať výsledky neadaptívneho offline učenia a adaptívneho online učenia. V experimente bola pomocou offline prístupu získaná hodnota F-miery pre posledné dva mesiace roku 2019 s priemerným výsledkom 0.7, podľa tohto výsledku sa určil odhad na ďalšie mesiace, s predpokladom, že ak by nenastal ani jeden konceptový drift, tak by offline model vykazoval približne rovnako dobré výsledky ako doteraz, teda okolo hodnoty 0.7 v F-miere.



Obrázok 2 Porovnanie prístupov pri detekcii falošných správ. Prevzatý z publikácie: „How concept drift can impair the classification of fake news“ [22]

Podľa Obrázku 1 je viditeľne, že úspešnosť klasifikácie pomocou offline metódy výrazne klesla oproti odhadovanému výkonu modelu. V sivo vyznačenej časti grafu sa nachádza hodnota F-miery, ktorá je 0.7 a zároveň je odhadovanou mierou pre ďalšie mesiace. S nástupom roku 2020 teda v mesiaci Január je ale viditeľný veľký pokles v F-miere pomocou offline metódy, v najhoršom bode teda v mesiaci Apríl 2020 klesla hodnota F-miery na 0.37 čo je celkovo o 0.33 menej ako predpokladaná hodnota. Tieto poklesy sú odôvodnené tak, že sú s veľkou pravdepodobnosťou spôsobené konceptovým driftom, konkrétne spojeným s prírastkom nových tém v roku 2020 ako napríklad COVID-19, prípadne prezidentskými voľbami v USA. COVID-19 obohatil príspevky vo veľkom počte o nové pojmy spojené s medicínou, liečbou a liekmi a keďže neadaptívny model sa týmto zmenám v pojmoch a témach nevie prispôsobiť, vykazoval oveľa nižšie výsledky ako pred týmito zmenami. Z obrázku 1 je taktiež viditeľný fakt, že adaptívny prístup vykazoval oveľa lepšie výsledky klasifikácie falošných správ v najvyššom bode až 0.81 F-miery. Z tejto štúdie je možné odvodiť záver, že statické modely výrazne zaostávajú v klasifikácii falošných správ pri zmenách v príspevkoch spôsobených konceptovým driftom.

3.6. Vplyv konceptového driftu na klasifikáciu falošných hodnotení

Ako už bolo spomenuté falošné hodnotenia resp. recenzie sú podobným problémom ako falošné správy a taktiež spadajú do kategórie antisociálneho správania na webe. Tieto falošné hodnotenia môžu vážne poškodiť meno značky ale aj prilákať zákazníkov ku kúpe produktov pomocou zavádzania a aj napriek aktívnemu boju proti týmto praktikám pretrváva fenomén falošných hodnotení dodnes. K detekcii falošných hodnotení sa dá pristupovať rôznymi spôsobmi ako napríklad kontrolou podozrivých účtov, overovaním relevancie rapídne pribúdajúcich pozitívnych alebo negatívnych komentárov v krátkom časovom rozpätí alebo detekciou pomocou metód strojového učenia.

Modely detekcie falošných hodnotení pracujú podobne ako modely detekcie falošných správ s textovými dátami, preto je relevantné aj v tejto problematike uvažovať nad prípadnými následkami konceptového driftu a spôsobmi ako s nim bojovať ak k nemu dôjde. Väčšina súčasných prístupov [23] zameraných na detekciu falošných hodnotení neberú do úvahy dôležitosť zachovania chronologického poradia falošných hodnotení. To je ale pri tejto problematike dôležitý faktor, ktorý by nemal byt ignorovaný, keďže postupom času sa menia spôsoby akými používatelia obchádzajú spam filtre resp. akými spôsobmi obchádzajú samotné detektory falošných hodnotení.

Pri falošných hodnoteniach môže teda často krát dôjsť k zmene spôsobu akým sú písane napríklad aké špeciálne znaky alebo emotikony používajú pre obídenie detekčného systému. Podobne ako pri falošných správach dochádza aj pri falošných hodnoteniach ku zmenám témy, dalo by sa

predpokladať, že v prípade falošných hodnotení je zmien tém oveľa viac, keďže môžu hodnotiť služby alebo produkty, pričom v každej kategórii sa logicky komentuje o inom produkte alebo službe.

Ako príklad by sa dala uviesť spoločnosť Amazon s ich internetovým obchodom, ktorý distribuuje obrovské množstvo produktov rôzneho typu od rôznych nezávislých predajcov. Ak by model detekcie falošných hodnotení na takejto webstránke nebol adaptívny resp. rezistentný voči konceptovému driftu tak by nemalo zmysel ho používať, keďže by nebol efektívny v dynamickom prostredí takéhoto internetového obchodu.

V roku 2021 bola vydaná publikácia [23], ktorá sa venovala vplyvu konceptového driftu na detekciu falošných hodnotení. Výsledkom experimentu popísaného v tejto publikácii je, že výkonnosť všetkých použitých metód pre detekciu falošných hodnotení časom klesla v dôsledku zmien charakteristík falošných hodnotení a času, preto je odporúčané predikčné modely často aktualizovať. Autori experimentu uviedli, že „Experimentálne výsledky naznačujú, že existuje silná korelácia negatívneho typu medzi konceptovým driftom a výkonnosťou klasifikácie, keďže konceptový drift negatívne ovplyvňuje predikcie modelu“. Z tohto vyjadrenia ale aj celého spomenutého experimentu je možné vyvodiť záver, že konceptový drift predstavuje rovnako veľkú hrozbu pri detekcií falošných hodnotení ako aj pri detekcií falošných správ a dá sa predpokladať, že sa tento problém vyskytuje pri detekcií akéhokoľvek antisociálneho správania na webe ak je textového typu.

3.7. Vizualizácia falošných správ

Samotná detekcia falošných správ je komplexná úloha, do ktorej zasahuje mnoho faktorov. Okrem viac známych faktorov ovplyvňujúcich modely detekcie falošných správ ako napríklad témy, časové rozpätie príspevkov, štýl akým sa príspevky píšu atď. môžu existovať aj iné faktory, ktoré je výhodné vizualizovať. Vizualizácia je vhodná pre jednoduchšiu interpretáciu a celkové pochopenie prepojení medzi faktormi, ktoré ovplyvňujú modely detekcie a samotnú detekciu.

V tejto oblasti môže byť nápomocný nastroj s názvom „FakeNewsTracker“ [24], ktorého cieľom je pomôcť a uľahčiť študovanie falošných správ. Tento nástroj dokáže automaticky zhromažďovať falošné správy, pomocou ktorých poskytuje datasety pre účel skúmania falošných správ. Obsahuje taktiež nástroje pre detekciu falošných správ pomocou viacerých modelov strojového učenia. Nakoniec dokáže vizualizovať falošné správy a teda zobraziť charakteristiky falošných správ, ktoré je možné následne skúmať. Pre vizualizáciu sú používané rôzne vizualizačné techniky.

FakeNewsTracker používa pre zhromažďovanie správ weby pre overovanie správ. Pri týchto weboch je výhodou, že okrem samotnej kategorizácie príspevku (teda či je pravdivý alebo nie) poskytujú aj odôvodnenie prečo je podľa nich príspevok klasifikovaný tak ako je. To je pri exploračnom nastrojí ako tomto výhodné, keďže spätná väzba týchto overovacích webov poskytuje viac informácii a kontext. Okrem zhromažďovania samotných falošných správ sa zhromažďujú aj informácie o príspevkoch a používateľoch, ktorí o týchto falošných správach písali na platforme Twitter. Využíva sa pritom „pokročilé vyhľadávanie“ [25] na Twitteri, ktoré umožňuje a zjednodušuje rôzne kroky ako napríklad nájsť Tweety so špecifickými frázami, slovami, vetami, heštegmi alebo písane v špecifickom jazyku. Ďalej poskytuje toto vyhľadávanie možnosti nájsť príspevky konkrétneho účtu, všetky odpovede na príspevky konkrétneho účtu, označenia konkrétneho účtu ale aj geografické miesta, z ktorých sa Tweetovalo. Tieto všetky informácie, ktoré sa dajú zhromaždiť poskytujú nové možnosti analýzy falošných správ pomocou FakeNewsTracker nástroja, keďže v ňom možno vyhľadávať prepojenia samotných falošných správ a ľudí, miest, heštegov atď. používaných v spojení s týmito falošnými správami.

3.8. Detekcia nenávistných príspevkov

Ako bolo spomenuté k antisociálnemu správaniu na webe patria aj nenávistné resp. ofenzívne príspevky známe ako „hate speech“. Tieto príspevky sa vyskytujú prevažne v textovej forme podobne ako falošné správy a falošné hodnotenia preto sú prístupy pre ich detekciu podobné. To či sa jedná alebo nejedná o príspevok nenávistného charakteru je mnohokrát nejasné. Existujú pritom aj argumenty proti regulácií takýchto príspevkov, keďže by regulácia potláčala slobodu slova resp. prvý dodatok ústavy ak sa jedná o USA. Najväčšie sociálne siete ako Twitter a Facebook však majú pri regulácii týchto príspevkov jasno a príspevky takéhoto typu regulujú alebo odstraňujú.

Anonymita na internete zabezpečuje ideálne podmienky pre používateľov, ktorých úmysel je pridávať nenávistný obsah a to vedie k narastajúcemu počtu príspevkov nenávistného typu. Rovnako ako u falošných správ je aj pri tejto problematike nutné používať automatizované prístupy detekcie týchto príspevkov. Európska Únia začala v roku 2016 na sociálne siete ako Twitter, Facebook alebo YouTube vyvíjať nátlak, keďže požadovali aby spomenuté sociálne siete preskúmavali väčšinu nahlásených príspevkov za nenávistné prejavy [31] s podmienkou, že majú byť tieto príspevky preskúmané do 24 hodín. Pri veľkom počte nahlasovaných príspevkov je manuálna kontrola každého príspevku v krátkom čase rovnako nemožná ako spomenutá manuálna kontrola falošných správ, preto sú nutné automatizované prístupy napríklad pomocou strojového učenia.

Niektoré štúdie používajúce metódy strojového učenia [32] pre detekciu nenávistných komentárov sú však problematické, keďže výsledky klasifikácie sú pre ľudí mnohokrát zle interpretovateľné. Tento fakt predstavuje problém, keďže v niektorých prípadoch by mohli používatelia, ktorí boli automaticky takýmto systémom zablokovaní požiadať o manuálne prešetrenie ich zablokovania a v takomto prípade by bolo nutné, aby pracovník sociálnej siete rozumel na základe čoho systém o zablokovaní používateľa rozhodol. Pri nenávistných príspevkoch je dôležitý kontext, keďže aj príspevky, ktoré na prvý pohľad pôsobia ako nenávistné nemusia porušovať pravidlá sociálnej siete. Tieto výnimky môžu platiť napríklad pri satirických príspevkoch alebo príspevkoch v rámci chránenej komunity čo predstavuje pre detekciu nenávistných komentárov ďalšiu výzvu. Niektoré prístupy detekcie nenávistných príspevkov využívajú slovníky, ktoré môžu metódam pomáhať v efektivite, no tieto slovníky musia byť neustále udržiavané a obohacované o nové pojmy spojené s touto problematikou, čo predstavuje problém. Existujú ale aj prístupy, ktoré nepoužívajú slovníkový prístup, ktorého údržba je pomerne náročná, ale využívajú prístupy, ktoré nie sú na slovníkoch ani zdrojoch tretích strán vôbec závislé. Jedným z takýchto prístupov je napríklad „Multi-view Support Vector Machine“, [33] ktorý využíva rôzne pohľady na príspevky, čím získava rôzne aspekty nenávistných príspevkov, čo vedie k efektívnejšej klasifikácii.

4. Návrh a implementácia riešenia zvolenej problematiky

Motiváciou k výskumu, ktorý je popísaný v tejto diplomovej práci je fakt, že v súčasnosti neexistuje veľké množstvo štúdií a experimentov, ktoré by sa venovali problému detekcie falošných správ z textových dát pomocou dátových prúdov teda pomocou dynamických metód. Absencia týchto štúdií znamená, že je táto problematika zatiaľ pomerne nová a nepreskúmaná, keďže väčšina štúdií v prostredí detekcie falošných správ z textov pristupuje k detekcií pomocou statických metód bez použitia dátových prúdov. Z tohto dôvodu sa praktická časť tejto diplomovej práce zameriava na problematiku detekcie falošných správ práve pomocou dynamických metód a to pomocou dátových prúdov, ktoré slúžia na vytvorenie simulácie pribúdania textových príspevkov rovnako ako na sociálnych sieťach.

Jedným z cieľov praktickej časti tejto diplomovej práce je v jednotlivých experimentoch odsledovať a overiť či zmena témy príspevkov vedie k vzniku konceptových driftov, ktoré následne znižujú úspešnosť klasifikácie falošných správ. Ďalším cieľom je následne overiť aké výsledky budú v daných experimentoch dosahovať adaptívne modely disponujúce detektorom driftov, pričom úspešná detekcia driftu zabezpečuje preučenie sa modelu s cieľom adaptovať ho na nové okolnosti. Následne porovnať výsledky rôznych adaptívnych modelov navzájom, ale aj oproti modelom bez adaptívnych nastavení teda bez detektorov driftu a overiť, či sú adaptívne modely na dátových prúdoch efektívnejšie ako tie neadaptívne. Dynamický prístup používaný v tejto diplomovej práci ale aj samotné ciele diplomovej práce sú odlišné od väčšiny štúdií v tejto problematike a teda výsledky tejto práce by mali byť prínosné v oblasti detekcií falošných správ pomocou dátových prúdov s použitím adaptívnych modelov.

4.1. Popis praktickej časti diplomovej práce

Praktická časť tejto diplomovej práce pozostáva z experimentov, ktorých cieľom je otestovať výkonnosť jednotlivých modelov s použitím dátových prúdov a sledovať rozdiely medzi adaptívnymi a neadaptívnymi prístupmi klasifikácie falošných správ. Pre praktickú časť práce bol používaný programovací jazyk Python. Dátové prúdy sú použité ako simulácie reálnych okolností v online priestore, teda simulujú dynamickosť sociálnych sietí s neustále pribúdajúcimi príspevkami od používateľov. V takomto dynamickom prostredí a teda aj v dátových prúdoch je bežným javom, že dochádza ku konceptovému driftu. Pri príspevkoch môže dochádzať k zmenám témy, štýlu akým sú napísané, používateľom, ktorý ich uverejnil atď. Pri experimentoch boli použité rôzne adaptívne modely ale aj ich neadaptívne verzie pre porovnanie ich efektívnosti v klasifikácií antisociálneho správania z dátových prúdov.

V experimentoch boli použité rôzne datasety s rôznym obsahom, jeden dataset je zameraný na ofenzívne príspevky na webe, čo je samozrejme taktiež súčasť antisociálneho správania, ostatné datasety sú zamerané na falošné správy a dezinformácie z webu, najčastejšie z platformy Twitter. V jednotlivých experimentoch boli tieto datasety používané aj samostatne, čo znamená že k nim neboli pridávané žiadne ďalšie príspevky z iných datasetov. V ďalších experimentoch došlo ale aj ku kombinovaniu datasetov pre zvýšenie počtu príspevkov v dátovom prúde a spojenie rôznych tém príspevkov, čo vyvoláva viacero zmien, ktoré môžu viesť k vzniku konceptových driftov, na ktoré sa modely musia prispôsobiť. Pre zachovanie autentickosti a všetkých charakteristík príspevkov neboli pri predspracovaní odstraňované ani menené žiadne časti príspevkov resp. textu, tento prístup plynie z motivácie otestovať ako efektívne budú jednotlivé modely klasifikovať príspevky so zachovaním štýlu akým boli napísané resp. čo má príspevok vyjadrovať.

Pre vyhodnocovanie modelov je používaná metóda s názvom „prequential evaluation“ [35] pričom slovo prequential je zložené zo slov prediktívny a sekvenčný. Táto metóda vyhodnocovania bola vytvorená konkrétne pre dátové prúdy, pričom má každý záznam streamu dva účely . Každý záznam zo streamu je najprv použitý pre otestovanie modelu, čo znamená, že je vytváraná predikcia a následne je ten istý záznam použitý pre trénovanie modelu. Pomocou tejto metódy je zabezpečené, že sú jednotlivé modely testované stále na záznamoch, s ktorými sa zatiaľ nestretli.

Pri všetkých modeloch používajúcich KSWIN detektor driftu založený na metóde okien bolo používané nastavenie s veľkosťou posúvajúceho sa okna na 100. Toto nastavenie veľkosti je pri KSWIN fixné, teda nikdy nemení svoju veľkosť. Zároveň bola pri všetkých modeloch využívajúcich KSWIN nastavená hodnota α na 0.005, keďže je doporučené aby táto hodnota bola veľmi malá teda nie väčšia ako 0.01, keďže je veľmi citlivá a zvýšenie by mohlo negatívne ovplyvniť detekciu driftov. Pri modeloch využívajúcich ADWIN detektor sa veľkosť okna nenastavuje dopredu, keďže o jeho veľkosti rozhoduje samotný algoritmus na základe analýzy štatistík. Pri všetkých modeloch s ADWIN detektorom bol parameter δ, nastavený na malú hodnotu, konkrétne 0.002 z rovnakých dôvodov ako pri KSWIN detektore.

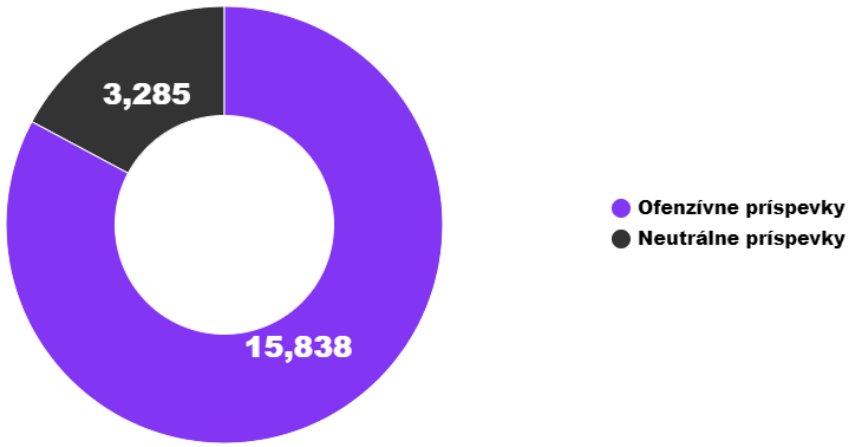
4.2. Detekcia ofenzívnych príspevkov

Prvým experimentom bolo otestovanie jednotlivých modelov s účelom klasifikácie ofenzívnych príspevkov. Pre tento experiment boli využité voľne dostupné dáta [26] z roku 2019, ktoré obsahujú Tweety a Retweety nenávistného a ofenzívneho charakteru. Pôvodné dáta obsahovali 6 atribútov:

- tweet – text daného príspevku
- count – počet užívateľov, ktorí klasifikovali obsah tweetov do tried podľa ich charakteru
- hate_speech – počet užívateľov, ktorí klasifikovali daný tweet ako nenávistný
- offensive_language – počet užívateľov, ktorí klasifikovali daný tweet ako ofenzívny
- neither – počet užívateľov, ktorí klasifikovali daný tweet ako neutrálny
- class – zaradenie tweetu do príslušnej skupiny podľa hodnotení užívateľov

Pre účel experimentu bol zachovaný atribút tweet a cieľový atribút.

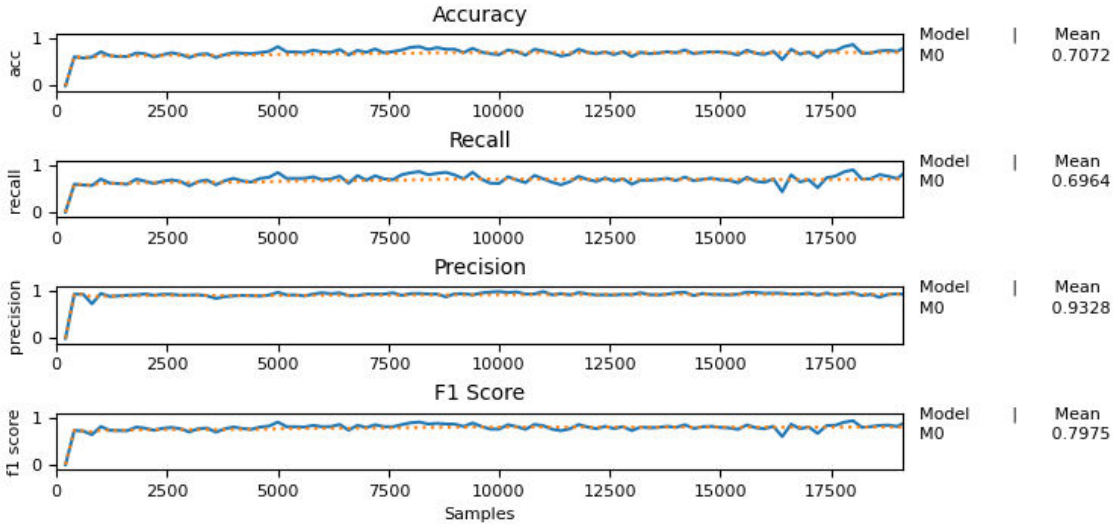
Ako bolo spomenuté, modely z balíka scikit-multiflow sú určené na binárnu klasifikáciu, z tohto dôvodu boli v atribúty „hate_speech“ a „offensive_language“ zlúčené a sú unifikované do triedy - offensive. Prázdne hodnoty a duplikáty boli z datasetu odstránené. Po tomto kroku obsahoval dataset 19 123 riadkov resp. príspevkov a 2 atribúty teda tweet a cieľový atribút. Podľa ADWIN detektoru bolo identifikovaných 7 driftov v bodoch: 767, 1279, 1343, 1407, 17791, 17855, 18367, okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.



Obrázok 3 Rozloženie predikovaného atribútu

4.2.1. Základný model pre porovnávanie

Ako základný („default“) model, ktorý slúžil pre porovnávanie výsledkov s ostatnými modelmi bol zvolený Naive Bayes klasifikátor so základnými nastaveniami. Nastavenia tohto modelu sú v balíku scikit-multiflow nemenné. Naive Bayes je klasifikačný algoritmus známy svojou jednoduchosťou.



Obrázok 4 Priebeh modelu Naive Bayes v úlohe klasifikácie ofenzívnych príspevkov

Základný model Naive Bayes klasifikátor dosahoval pri klasifikácií ofenzívnych príspevkov úspešnosť klasifikácie 70,72% a hodnota F1-miery bola 79.75%, výsledky sú spriemerované na celom streame.

4.2.2. Hoeffding Tree

Následne boli pre tento dataset použité Hoeffding Tree klasifikátory. Pre experiment boli použité tri verzie tohto algoritmu a to verzia, ktorá používa pre predikciu Naive Bayes bez ADWIN detektora zmien, následne bola použitá aj adaptívna verzia používajúca pre predikciu mechanizmus „Naive Bayes Adaptive“ v kombinácii s ADWIN detektorom zmien a keďže je predikovaný atribút v tomto datasete nevybalancovaný bola použitá aj verzia Majority Class, ktorá je pre takéto prípady dobre použiteľná. ADWIN je adaptívny algoritmus so schopnosťou zisťovať zmeny a udržiavať aktuálne štatistiky o dátovom prúde. ADWIN umožňuje, aby aj algoritmy, ktoré nie sú prispôsobené na driftujúce dáta, boli voči tomuto javu odolné.

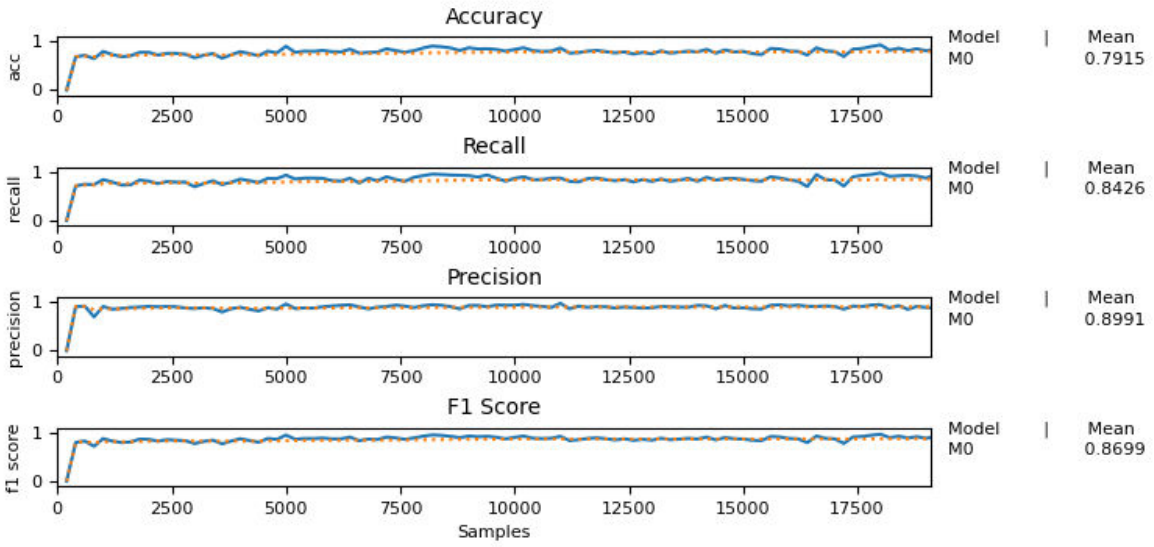
Tabuľka 1 Porovnanie výsledkov Hoeffding Tree modelov v úlohe klasifikácie ofenzívnych príspevkov

Metrika	**Hoeffding Tree + Majority Class**	**Hoeffding Tree + Naive Bayes**	**Hoeffding Tree + ADWIN + Naive Bayes Adaptive**
Úspešnosť klasifikácie	0.8277	0.6900	0.8331
Návratnosť	1.0	0.6972	0.9874
Presnosť	0.8277	0.9067	0.8393
F-miera	0.9057	0.7883	0.9073

Verzia používajúca Hoeffding Tree v kombinácii s Naive Bayes bez adaptívnych nastavení a bez ADWIN detektora bola výsledkami najmenej efektívna, v presnosti klasifikácie a F-miere výrazne zaostávala oproti ostatným dvom použitým verziám. Verzia používajúca Naive Bayes Adaptive v kombinácii s ADWIN detektorom bola výsledkami najlepšia spomedzi troch použitých verzii aj keď tesne za ňou bola verzia používajúca Majority Class, ktorá svoje dobré výsledky dosiahla vďaka schopnosti adaptovať sa na nevybalancovanú predikovanú triedu pri datasetoch s väčším počtom záznamov.

4.2.3. Batch Incremental

Tento klasifikátor je možné kombinovať s rôznymi modelmi. Zozbiera sa okno príkladov, ktoré sa potom používajú na trénovanie nového modelu, každý model sa potom pridáva do súboru modelov. Existuje hranica maximálneho počtu týchto modelov, ktorá ak je dosiahnutá, tak sa najstarší model vymaže a pridá sa najnovší model. Tento prístup bol použitý pre otestovanie ako resp. či vôbec zvýši efektívnosť základného modelu, teda Naive Bayes klasifikátor, ak je kombinovaný s inkrementálnym prístupom dávkovania.



Obrázok 5 Priebeh modelu Batch Incremental v kombinácii s Naive Bayes v úlohe klasifikácie ofenzívnych príspevkov

Ako možno vidieť z grafu na obrázku 6, základný model Naive Bayes v kombinácii s inkrementálnym učením z dátových prúdov pomocou dávkovania je výrazne efektívnejší ako základný model bez inkrementálneho prístupu. Úspešnosť klasifikácie sa oproti základnému modelu zvýšila o 8% a F-miera o 7%. Z týchto výsledkov možno sledovať, že aj neadaptívny model sa v kombinácií s inkrementálnou metódou môže stať efektívnejším no za výsledkami kompletne adaptívnych metód napriek tomu zaostáva.

4.2.4. Streaming Random Patches

SRP klasifikátor na rozdiel od väčšiny iných dostupných metód ponúka možnosť zmeniť drift detektor používaný v modely na zisťovanie zmien. SRP teda poskytuje možnosť otestovať rôzne detektory v kombinácii s adaptívnym modelom. Ako základný model pre porovnávanie bol zvolený SRP v kombinácii s Naive Bayes pričom nastavenia modelu boli zmenené tak aby boli adaptívne nastavenia a detektory driftu pri tejto verzii vypnuté. Následne pre všetky ostatné kombinácie SRP boli použité adaptívne modely, konkrétne Hoeffding Adaptive Tree v kombinácii s jednotlivými detektormi konceptového driftu. Cieľom tohto kroku je porovnať rozdiely vo výsledkoch modelov s použitím rôznych detektorov driftu, na datasete ofenzívnych príspevkov.

Tabuľka 2 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie ofenzívnych príspevkov

Metrika	Default	ADWIN	KSWIN	EDDM	PageHinkley	HDDMA	HDDMW
Úspešnosť klasifikácie	0.7058	0.8353	0.8356	0.8242	0.8368	0.8363	0.8329
Návratnosť	0.6949	0.9669	0.9642	0.9766	0.96	0.9641	0.9660
Presnosť	0.9324	0.8536	0.8555	0.8378	0.8592	0.8563	0.8519
F-miera	0.7963	0.9067	0.9066	0.9019	0.9069	0.9070	0.9054

Z výsledkov je viditeľný fakt, že všetky modely bez ohľadu na to aký detektor používali mali dobré výsledky, ktoré boli výrazne lepšie ako výsledky základného neadaptívneho modelu bez detektora driftu. Všetky výsledky boli veľmi podobné a rozdiely boli v podstate zanedbateľné, preto by sa dalo skonštatovať, že pri detekcii ofenzívnych príspevkov v tomto datasete s použitím Streaming Random Patches klasifikátora v kombinácii s adaptívnym modelom Hoeffding Adaptive Tree nezáleží na tom aký detektor je používaný.

4.2.5. Learn PPNSE

Pre experiment boli použité tri verzie tohto klasifikátora s cieľom porovnať ich výkonnosť v klasifikácii ofenzívnych príspevkov. V prvej verzií bola použitá kombinácia s Decision tree neadaptívnym klasifikátorom. V druhej verzii bol PPNSE kombinovaný s Hoeffding Adaptive Tree, teda adaptívnym klasifikátorom, ktorý disponuje ADWIN detektorom a v poslednej bol PPNSE kombinovaný s KNNADWIN adaptívnym klasifikátorom s ADWIN detektorom driftu.

Tabuľka 3 Porovnanie výsledkov LPPNSE modelov v úlohe klasifikácie ofenzívnych príspevkov

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.7536	0.8027	0.8381
Návratnosť	0.8351	0.9435	0.9669
Presnosť	0.8628	0.8384	0.8561
F-miera	0.8487	0.8879	0.9082

Pri PPNSE klasifikátori dosahovala najlepšie výsledky kombinácia s adaptívnym modelom KNNADWIN, ktorý má zabudovaný ADWIN drift detektor. Oproti Neadaptívnej verzii bola adaptívna lepšie v presnosti klasifikácie o 7% a v F-miere o 6%. Celkovo tak v tomto experimente mali najlepšie výsledky adaptívne modely v kombinácií s detektormi driftu.

4.2.6. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie a zároveň aj podľa F-miery dosahoval model LPPNSE v kombinácií s adaptívnym modelom KNNADWIN.

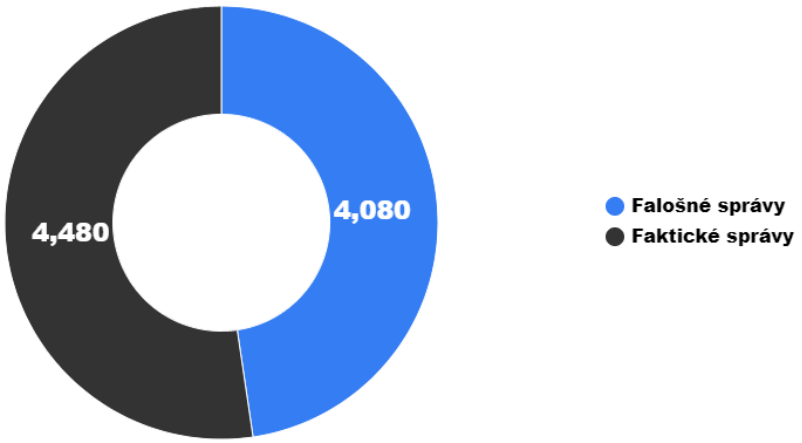
Tabuľka 4 Porovnanie výsledkov všetkých modelov pri detekcií ofenzívnych príspevkov

Metriky	NB DEF	BI. + NB	HT + NB	HAT + NBA	HT + MC	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.7072	0.7915	0.6900	0.8331	0.8277	0.7058	0.8353	0.8356	0.8242	0.8368	0.8363	0.8329	0.7536	0.8027	**0.8381**
Návratnosť	0.6964	0.8426	0.6972	0.9874	1.0	0.6949	0.9669	0.9642	0.9766	0.96	0.9641	0.9660	0.8351	0.9435	**0.9669**
Presnosť	0.9328	0.8991	0.9067	0.8393	0.8277	0.9324	0.8536	0.8555	0.8378	0.8592	0.8563	0.8519	0.8628	0.8384	**0.8561**
F-miera	0.7975	0.8699	0.7883	0.9073	0.9057	0.7963	0.9067	0.9066	0.9019	0.9069	0.9070	0.9054	0.8487	0.8879	**0.9082**

4.3. Detekcia falošných správ v oblasti COVID19

V ďalšom experimente bol používaný dataset [27] obsahujúci Twitter príspevky spojené s ochorením COVID-19. Ako v predošlom tak aj v tomto experimente zatiaľ nedošlo ku kombinovaniu s inými datasetmi resp. témami, čiže boli používané iba príspevky spojené s COVID-19. Z tohto prístupu je možné predpokladať, že konceptových driftov resp. detekcií bude pomerne málo, keďže sú príspevky zamerané na prakticky rovnakú tému, používajú mnoho rovnakých slov napríklad výrazov z medicíny, názvov vakcín atď. V tomto experimente je teda predpokladom, že rozdiely v efektívnosti klasifikácie medzi adaptívnymi a neadaptívnymi metódami bude menší ako pri predošlom experimente.

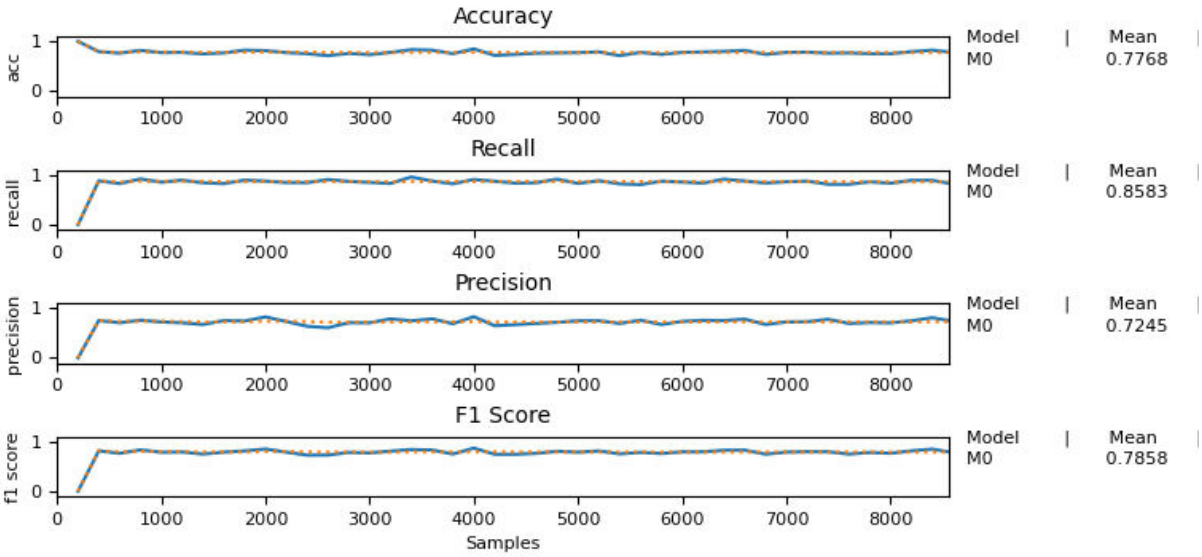
Dataset obsahuje 8560 záznamov a 2 atribúty, konkrétne text samotných Tweetov a cieľový atribút. Podľa ADWIN detektoru boli identifikované 2 drifty v bodoch: 4095 a 4863, okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.



Obrázok 6 Rozloženie predikovaného atribútu

4.3.1. Základný model pre porovnávanie

Ako základný model bol opäť použitý Naive Bayes klasifikátor. Tentokrát má základný model lepšie výsledky v porovnaní s predošlým experimentom a to hlavne pri miere presnosti klasifikácie.



Obrázok 7 Priebeh modelu Naive Bayes v úlohe klasifikácie falošných správ COVID19

Základný model Naive Bayes klasifikátor dosahoval pri klasifikácii príspevkov falošných správ o COVIDe-19 úspešnosť klasifikácie 77,68% a hodnota F-miery bola 78.58%, výsledky sú spriemerované na celom streame

4.3.2. Hoeffding Tree

Podobne ako pri predošlom experimente aj v tomto boli použité Hoeffding Tree klasifikátory pre porovnanie so základným modelom. Tentokrát už nebolo nutné zahrnúť aj prístup s Majority Class, keďže predikovaný atribút je počtom vyvážený.

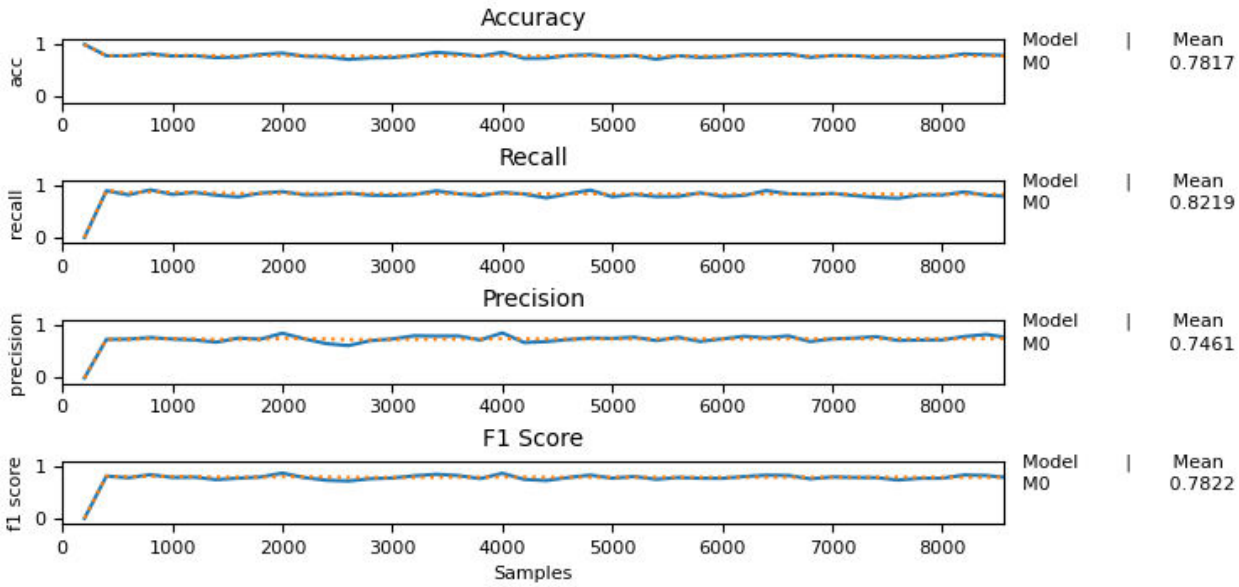
Tabuľka 5 Porovnanie výsledkov Hoeffding Tree modelov v úlohe klasifikácie falošných správ COVID19

Metrika	**Hoeffding Tree + Naive Bayes**	**Hoeffding Tree + ADWIN + Naive Bayes Adaptive**
Úspešnosť klasifikácie	0.7667	0.7947
Návratnosť	0.8293	0.8046
Presnosť	0.7226	0.7741
F-miera	0.7723	0.7890

Z výsledkov a porovnania týchto modelov je viditeľné, že Hoeffding Tree s použitím Naive Bayes bol menej efektívny v porovnaní s adaptívnou verziou v kombinácii s ADWIN detektorom. Výsledky adaptívnej verzie tohto modelu mali pomerne blízko k výsledkom základného neadaptívneho modelu, čo potvrdzuje predpoklad, že v rámci tohto datasetu nedochádza k dostatočne výrazným zmenám v príspevkoch na to, aby boli adaptívne modely výrazným spôsobom lepšie oproti neadaptívnym.

4.3.3. Batch Incremental

Pri inkrementálnom prístupe dávkovania spojenom so základným modelom Naive Bayes sú v tomto experimente taktiež badateľné oveľa menšie zmeny k lepšiemu ako pri predošlom experimente, takže aj pri tomto modely sa splnil predpoklad určený na začiatku.



Obrázok 8 Priebeh modelu Batch Incremental v kombinácii s Naive Bayes v úlohe klasifikácie falošných správ o COVIDe19

4.3.4. Streaming Random Patches

Rovnako ako v predošlom experimente aj v tomto boli otestované modely s použitím rôznych detektorov konceptového driftu. Cieľom porovnania je zistiť či bude základný model bez detektora dosahovať v tomto experimente približne rovnaké výsledky ako modely s detektormi a zistiť, ktorý detektor je najefektívnejší.

Tabuľka 6 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie falošných správ o COVIDe19

Metrika	Default	ADWIN	KSWIN	EDDM	PageHinkley	HDDMA	HDDMW
Úspešnosť klasifikácie	0.7780	0.8034	0.8057	0.8038	0.8145	0.8041	0.7996
Návratnosť	0.8568	0.8115	0.8269	0.8197	0.8321	0.8261	0.8187
Presnosť	0.7268	0.7841	0.7794	0.7802	0.7904	0.7774	0.7744
F-miera	0.7865	0.7976	0.8024	0.7995	0.8107	0.8010	0.7959

Ako bolo predpokladané, základný model bez detektoru driftu a adaptívnych nastavení bol výsledkami pomerne blízko k ostatným modelom, ktoré sú adaptívne a používajú drift detektor. Tento jav je spôsobený malým množstvom zmien, čo je pre tento dataset prirodzené, keďže obsahuje iba príspevky o COVIDe-19 .

Najefektívnejším modelom bol model, ktorý používal detektor driftu s názvom Page Hinkley, pričom tento model dosahoval úspešnosť klasifikácie lepšiu o 3% a F-mieru lepšiu o 2% oproti základnému modelu.

4.3.5. Learn PPNSE

Pre experiment boli opäť použité tri verzie PPNSE klasifikátora s cieľom porovnať ich výkonnosť v klasifikácií. V prvej verzií bola použitá kombinácia s Decision Tree neadaptívnym klasifikátorom v druhej verzii bol PPNSE kombinovaný s Hoeffding Adaptive Tree, teda adaptívnym klasifikátorom využívajúcim ADWIN a v tretej verzii s KNN-ADWIN.

Tabuľka 7 Porovnanie výsledkov LPPNSE modelov v úlohe klasifikácie falošných správ o COVIDe19

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.7157	0.7862	0.7492
Návratnosť	0.6970	0.8177	0.9501
Presnosť	0.7039	0.7546	0.6662
F-miera	0.7004	0.7850	0.7832

Pri PPNSE klasifikátori dosahovala najlepšie výsledky kombinácia s adaptívnym modelom Hoeffding Adaptive Tree, ktorý má zabudovaný ADWIN drift detektor. Najväčší rozdiel bol medzi adaptívnou verziou a neadaptívnou verziou používajúcou Decision Tree, kde bola úspešnosť klasifikácie u adaptívneho modelu presnejšia o 7% a F-miera o 9%.

4.3.6. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie a zároveň aj podľa F-miery dosahoval model SRP v kombinácií s Page Hinkley detektorom driftu.

Tabuľka 8 Porovnanie výsledkov všetkých modelov pri detekcií falošných správ o COVIDe19

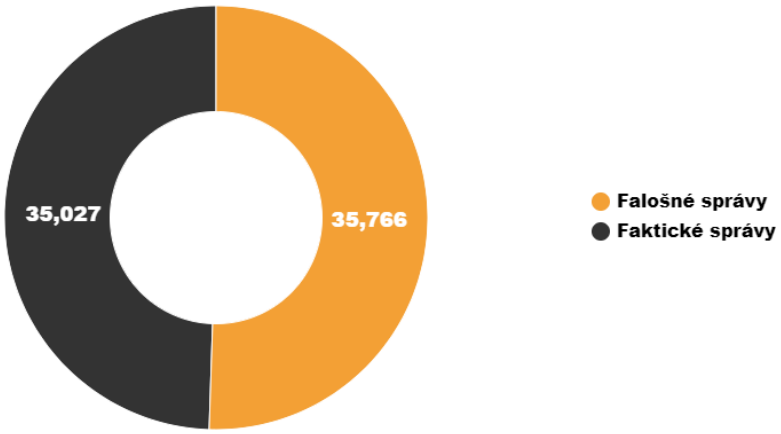
Metriky	NB DEF	BI. + NB	HT + NB	HAT + NBA	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.7768	0.7817	0.7667	0.7947	0.7780	0.8034	0.8057	0.8038	**0.8145**	0.8041	0.7996	0.7157	0.7862	0.7492
Návratnosť	0.8583	0.8219	0.8293	0.8046	0.8568	0.8115	0.8269	0.8197	**0.8321**	0.8261	0.8187	0.6970	0.8177	0.9501
Presnosť	0.7245	07461	0.7226	0.7741	0.7268	0.7841	0.7794	0.7802	**0.7904**	0.7774	0.7744	0.7039	0.7546	0.6662
F-miera	0.7858	0.7822	0.7723	0.7890	0.7865	0.7976	0.8024	0.7995	**0.8107**	0.8010	0.7959	0.7004	0.7850	0.7832

4.4. Detekcia dezinformačných príspevkov

Pre ďalší experiment boli zvolené dáta, [28] ktoré obsahujú faktické a dezinformačné články. Jednotlivé záznamy sú tvrdenia resp. vyjadrenia obsiahnuté v článkoch publikovaných na internete a zdieľaných cez sociálne siete.

Počet záznamov v datasete je 70 793 čiže ide o pomerne obsiahly dataset, pričom príspevky sú politického charakteru, popisujú však rôzne témy ako napríklad prezidentské voľby, migračná kríza, náboženstvo atď. Dáta obsahovali aj dátum a vďaka tomu ich bolo možné zoradiť podľa časovej postupnosti a tým pádom v istom zmysle nasimulovať reálne podmienky pribúdajúcich príspevkov v dátovom prúde so zmenami v témach aj keď majú rovnakú nad kategóriu – politické príspevky. Podľa ADWIN detektoru bolo identifikovaných 12 driftov v bodoch:

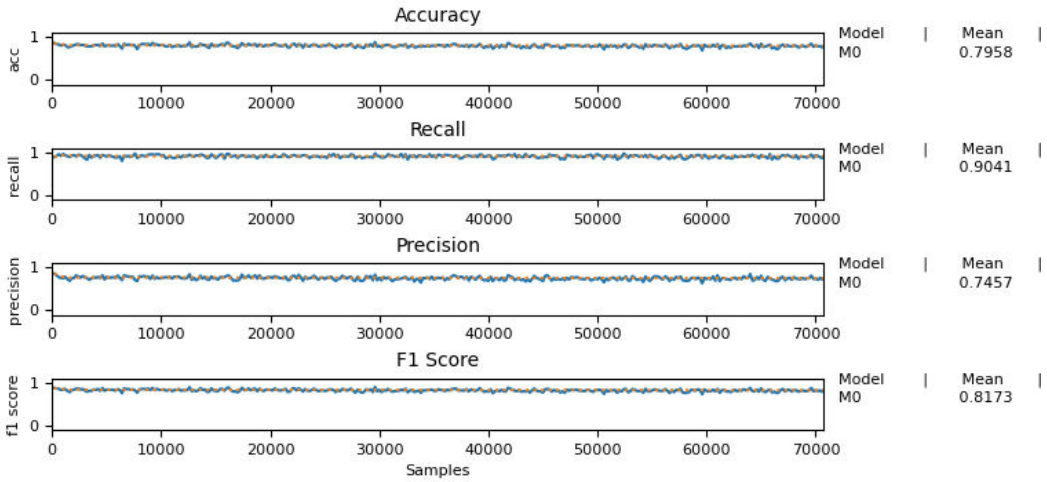
3839, 4863, 5247, 7007, 15871, 16895, 20991, 30815, 58527, 58879, 60831, 69247 okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.



Obrázok 9 Rozloženie predikovaného atribútu

4.4.1. Základný model pre porovnávanie

Ako základný model, ktorý slúžil pre porovnávanie výsledkov s ostatnými modelmi bol opäť zvolený Naive Bayes Klasifikátor. Základný model Naive Bayes klasifikátor dosahoval pri klasifikácií dezinformačných príspevkov úspešnosť klasifikácie 79,58% a hodnota F-miery bola 81.73%, výsledky sú spriemerované na celom streame. Pri zohľadnený, že sa nejedná o adaptívny model sú tieto výsledky pomerne dobré.



Obrázok 10 Priebeh modelu Naive Bayes v úlohe klasifikácie dezinformačných príspevkov

4.4.2. Hoeffding Tree

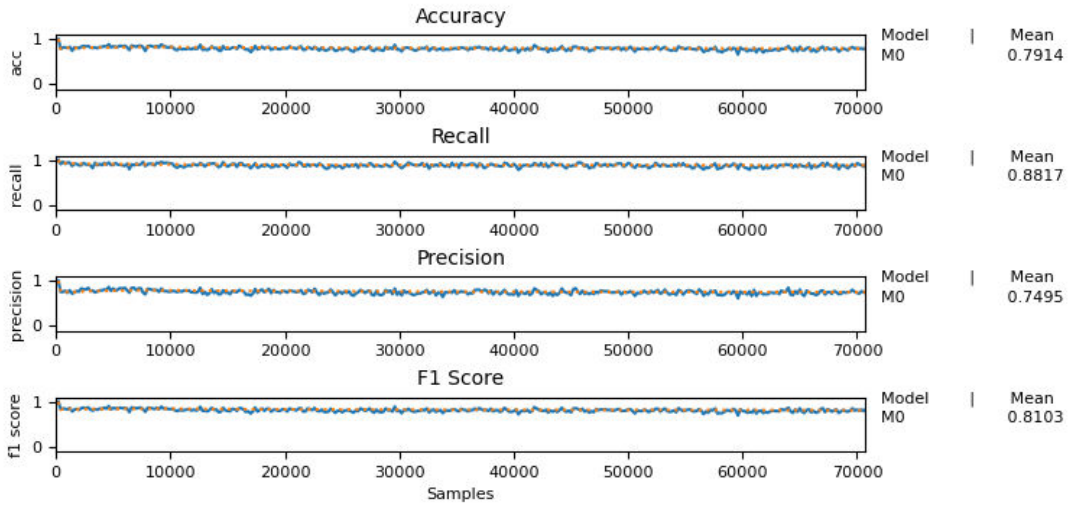
Modely pre porovnanie so základným modelom boli Hoeffding Tree modely. Nie je nutné zahrnúť aj prístup s Majority Class, keďže predikovaný atribút je vyvážený.

Tabuľka 9 Porovnanie výsledkov Hoeffding Tree modelov v úlohe klasifikácie dezinformačných príspevkov

Metrika	**Hoeffding Tree + Naive Bayes**	**Hoeffding Tree + ADWIN + Naive Bayes Adaptive**
Úspešnosť klasifikácie	0.8193	0.8240
Návratnosť	0.8070	0.8305
Presnosť	0.8305	0.8228
F-miera	0.8186	0.8266

Vo výsledku bol najlepším modelom adaptívny Hoeffding Adaptive Tree s ADWIN detektorom a Naive Bayes Adaptive prístupom. Výsledky boli hodnotami v jednotlivých metrikách pri týchto modeloch pomerne blízko.

4.4.3. Batch Incremental



Obrázok 11 Priebeh modelu Batch Incremental v kombinácii s Naive Bayes v úlohe klasifikácie dezinformačných príspevkov

Pri použití tohto modelu využívajúceho metódy inkrementálneho učenia neboli výsledky základného modelu spojeného s týmto prístupom výrazne zlepšené. Inkrementálny prístup dávkovania teda nebol v tomto prípade efektívny a nezlepšil základný model Naive Bayes.

4.4.4. Streaming Random Patches

Ďalším krokom bolo použitie SRP modelu v kombinácií s Hoeffding Adaptive Tree s rôznymi detektormi driftu s cieľom porovnať ich výkonnosť a zistiť, ktorý z nich dosahuje najlepšie výsledky. Najlepšie výsledky dosahovala kombinácie SRP modelu s detektorom driftu Page Hinkley, pričom bol od základného modelu v presnosti klasifikácie lepší o 6% a v F-miere o 4%. Ostatné kombinácie SRP modelu s detektormi mali taktiež veľmi dobré výsledky. Vo výsledku sa dá zhodnotiť, že všetky adaptívne modely mali dobré výsledky klasifikácie s viditeľnými rozdielmi oproti základnému neadaptívnemu SRP modelu bez detektora driftu.

Tabuľka 10 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie dezinformačných príspevkov

Metrika	Default	ADWIN	KSWIN	EDDM	PageHinkley	HDDM_A	HDDM_W
Úspešnosť klasifikácie	0.7977	0.8330	0.8551	0.8298	0.8609	0.8529	0.8569
Návratnosť	0.9047	0.8396	0.8541	0.8437	0.8607	0.8496	0.8571
Presnosť	0.7478	0.8315	0.8583	0.8237	0.8634	0.8578	0.8592
F-miera	0.8188	0.8355	0.8562	0.8336	0.8621	0.8537	0.8581

4.4.5. Learn PPNSE

Opäť boli použité tri verzie PPNSE klasifikátora pričom základná verzia bola neadaptívna bez detektora driftu a ostatné dve verzie boli adaptívne využívajúce ADWIN detektor driftu.

Tabuľka 11 Porovnanie výsledkov LPPNSE modelov v úlohe klasifikácie dezinformačných príspevkov

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.6893	0.8570	0.8139
Návratnosť	0.6881	0.8957	0.7303
Presnosť	0.6941	0.8335	0.8809
F-miera	0.6911	0.8635	0.7986

Pri PPNSE klasifikátori opäť dosahovala najlepšie výsledky kombinácia s adaptívnym modelom Hoeffding Adaptive Tree, ktorý má zabudovaný ADWIN drift detektor. Najväčší rozdiel bol medzi adaptívnou verziou a neadaptívnou verziou používajúcou Decision Tree, kde bola úspešnosť klasifikácie u adaptívneho modelu Hoeffding Adaptive Tree presnejšia o 17% a F-miera o 17%.

4.4.6. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie mal model SRP v kombinacií s Page Hinkley detektorom driftu. Podľa F-miery bol najefektívnejší LPPNSE v kombinácií s Hoeffding Adaptive Tree.

Tabuľka 12 Porovnanie výsledkov všetkých modelov pri detekcií dezinformačných príspevkov

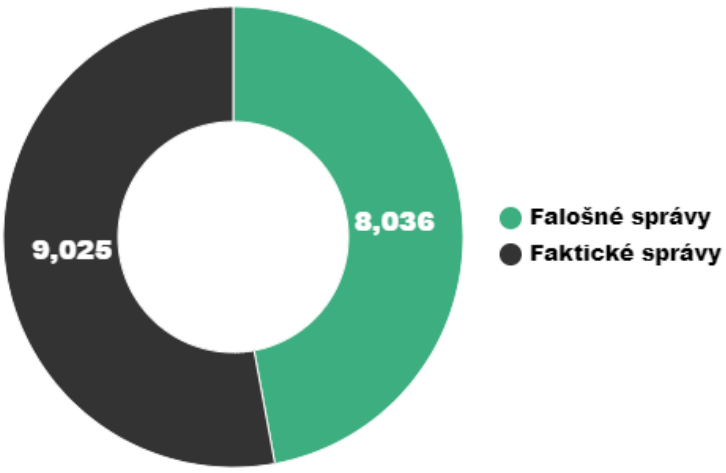
Metriky	NB DEF	BI. + NB	HT + NB	HAT + NBA	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.7958	0.7914	0.8193	0.8240	0.7977	0.8330	0.8551	0.8298	**0.8609**	0.8529	0.8569	0.6893	**0.8570**	0.8139
Návratnosť	0.9041	0.8817	0.8070	0.8305	0.9047	0.8396	0.8541	0.8437	**0.8607**	0.8496	0.8571	0.6881	**0.8957**	0.7303
Presnosť	0.7457	07495	0.8305	0.8228	0.7478	0.8315	0.8583	0.8237	**0.8634**	0.8578	0.8592	0.6941	**0.8335**	0.8809
F-miera	0.8173	0.8103	0.8186	0.8266	0.8188	0.8355	0.8562	0.8336	**0.8621**	0.8537	0.8581	0.6911	**0.8635**	0.7986

4.5. Detekcia falošných správ z kombinovaného dvoj-témového datasetu

Pre ďalší experiment boli skombinované dva datasety pre vytvorenie prechodu z jednej témy, o ktorej sa hovorí online na inú tému. V rámci prvého kroku v tomto experimente boli vyselektované príspevky s politickou tematikou, ktoré boli zamerané konkrétne na voľby resp. obsahovali príspevky spomínajúce voľby.

Do úvahy bolo brané časové hľadisko príspevkov teda podľa dátumu bolo vyselektovaných posledných resp. časovo najnovších 8502 príspevkov spomínajúcich voľby. Ako ďalší dataset bol vybraný covid dataset, ktorý obsahoval 8559 príspevkov teda pre zachovanie rovnomernosti medzi datasetmi nebol upravovaný počet záznamov a patrili pod neho rôzne pod-témy, teda príspevky nie

len spomínajúce COVID-19 samotný ale napríklad aj vakcináciu a rôzne nepodložené liečby resp. liečba vírusu spojená s konvertovaním na istú vieru atď. Oba datasety teda potenciálne obsahujú rôzne pod-témy ale hlavná veľká zmena témy z prechodu volebných príspevkov na prechod covidových príspevkov nastáva po 8502. príspevku. Cieľom je otestovať ako dobre sa adaptívne modely oproti neadaptívnym prispôsobia na takúto veľkú zmenu a ako veľmi bude táto zmena pri klasifikácii viditeľná.

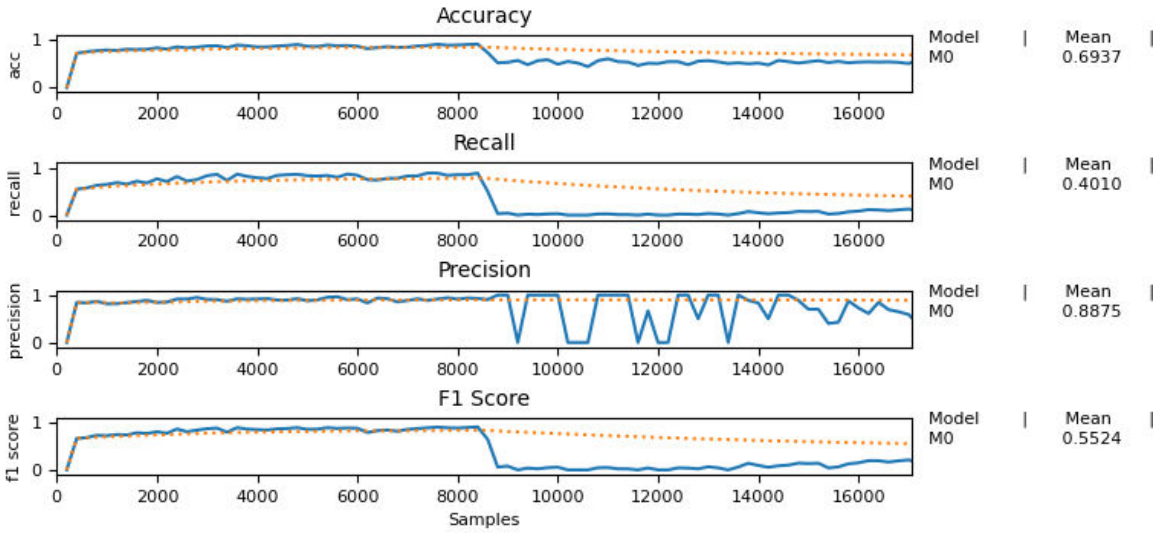


Obrázok 12 Rozloženie predikovaného atribútu

4.5.1. Základný model pre porovnávanie

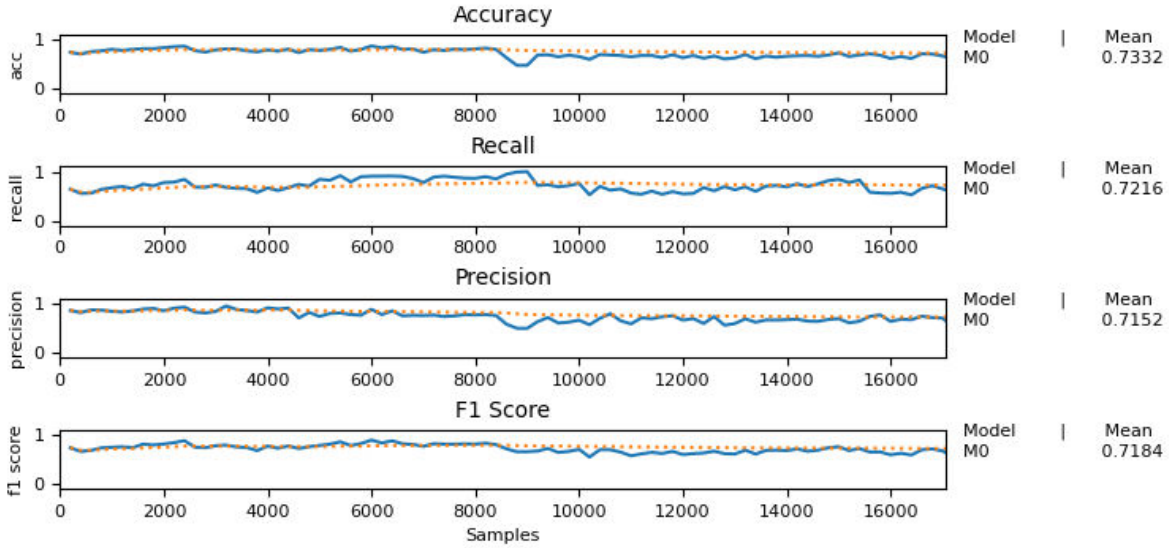
Ako základný model, ktorý slúžil pre porovnávanie výsledkov s ostatnými modelmi bol zvolený Naive Bayes Klasifikátor. Základný model Naive Bayes klasifikátor dosahoval pri klasifikácií príspevkov kombinovaného datasetu úspešnosť klasifikácie 69,37% a hodnota F1-skóre bola 55.24%, výsledky sú spriemerované na celom streame.

Základný neadaptívny model bol teda veľmi neefektívny pri klasifikácii dezinformácií kombinovaného datasetu. Taktiež je podľa vykresleného grafu v obrázku 15 veľmi dobre viditeľné, že krátko po zmene témy z príspevkov o voľbách na príspevky o covide nastal veľký problém v klasifikácii. Po 8500. príspevku je viditeľný veľký pokles v presnosti klasifikácie a F-miere, čo je prirodzené, keďže tento model nedisponuje vlastnosťami, ktoré by sa dokázali na túto zmenu prispôsobiť. Podľa ADWIN detektoru bolo identifikovaných 10 driftov v bodoch: 3807, 4639, 6047, 8863, 8895, 8927, 9215, 9727, 10751, 11103 okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.



Obrázok 13 Priebeh modelu Naive Bayes v úlohe klasifikácie falošných správ pri dvoch témach

Pre porovnanie so základným modelom bol zvolený adaptívny model Hoeffding Adaptive Tree, ktorý disponuje ADWIN detektorom driftu s použitím Naive Bayes Adaptive (NBA) metódy. Pri použití adaptívneho modelu je viditeľné z grafu na obrázku 16, že tento prístup sa vedel oveľa lepšie vysporiadať so zmenou témy nastávajúcou po 8502. príspevku a zachoval si približne rovnakú úspešnosť klasifikácie a F-mieru počas celého prúdu dát.



Obrázok 14 Priebeh modelu Hoeffding Adaptive Tree s použitím Naive Bayes Adaptive v úlohe klasifikácie falošných správ pri dvoch témach

4.5.2. Detekcia zmien v kombinovanom datasete

Pomocou ADWIN detektora boli zistené v dátach drifty, ktorých bolo 10. Každý zistený drift dáva modelom, ktoré disponujú detektorom driftu spätnú väzbu na to aby sa adaptovali. Okrem samotných detektorov driftov modely disponujú aj varovnými metódami, ktoré ich varujú pred potenciálnym nastátím driftu ešte pred tým než nastane pomocou rôznych indikátorov a menších zmien v dátach, preto nie sú samotné počty zistených driftov nikdy veľmi veľké ale aj napriek tomu sú modely viditeľne efektívnejšie ak disponujú týmito detektormi.

Z desiatich zistených driftov boli tri pred zmenou témy na COVID-19, to znamená že sa vyskytli v rámci prvej témy. Ako príklady driftu je možné uviesť nasledujúce detekcie. Prvý zistený drift nastal pri príspevku, kedy sa prvý krát používali slová ako „Yushan“, čo je názov najvyššieho vrchu v Taiwane, taktiež sa v datasete prvý krát objavilo meno „Fengshan“, okrem spomenutia volieb sa v tomto príspevku riešilo potenciálne napadnutie Taiwanu Čínou a práve tieto nové pojmy resp. téma príspevku ako celok viedli k detekcii driftu.

V ďalšom detegovanom príspevku išlo o kontroverziu spojenú s voľbou prísediaceho sudcu najvyššieho sudu v USA, v tomto príspevku bola prvý krát a zároveň opakovane (konkrétne 6 krát) použitá fráza „Position unclear“, ktorá bola používaná pre vyjadrenie rôznych politikov k tejto téme volieb. Toto opakovanie vopred nevídanej frázy spojené s rôznymi menami pravdepodobne vyvolalo drift detegovaný v tomto príspevku.

Po príspevku 8502 končili príspevky s témou spomínajúcou voľby a začali príspevky s témou spomínajúcou COVID-19. V krátkej dobe po tom (v rámci 400 príspevkov po zmene) došlo k trom ďalším detekciám driftu, tento fakt je viditeľný aj podľa grafov, ktoré pri adaptívnych modeloch s detektorom zobrazujú krátku stratu presnosti klasifikácie a následne jej obnovenie na doterajšie hodnoty (hodnoty ako pred driftom) je teda nutné poznamenať, že drift nemusí a často krát pravdepodobne ani nie je zaznamenaný ihneď po zmene témy ale až po niekoľkých desiatkach resp. stovkách príspevkov nasledujúcich po zmene témy, v tomto prípade sa adaptívne modely dokázali adaptovať a vrátiť úspešnosť klasifikácie na dostatočne dobré hodnoty približne do 500 až 1000 príspevkov po náhlej a výraznej zmene z témy na tému .

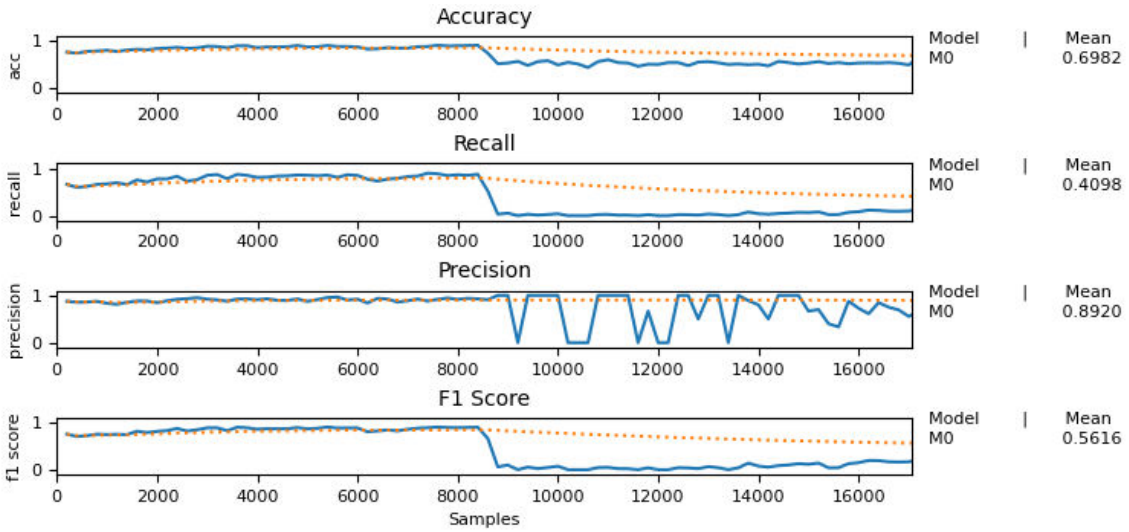
4.5.3. Streaming Random Patches

Pre kombinovaný dataset bol použitý opäť SRP klasifikátor pre otestovanie každého detektora driftu a vytvorenie ich porovnania.

Tabuľka 13 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie falošných správ pri dvoch témach

Metrika	Default	ADWIN	KSWIN	EDDM	PageHinkley	HDDM_A	HDDM_W
Úspešnosť klasifikácie	0.6982	0.7695	0.7670	0.7239	0.7654	0.7663	0.7565
Návratnosť	0.4098	0.7383	0.7398	0.6931	0.7317	0.7312	0.7206
Presnosť	0.8920	0.7648	0.7599	0.7134	0.7615	0.7634	0.7526
F-miera	0.5616	0.7514	0.7497	0.7031	0.7463	0.7470	0.7362

Najlepšie hodnoty dosahoval SRP s ADWIN detektorom driftu, zatiaľ čo najhoršie výsledky podľa presnosti klasifikácie a F-miery dosahovala „Early drift detection method“ skratkou EDDM. Základný model bez akéhokoľvek detektora alebo adaptívnych nastavení viditeľný na grafe v obrázku 18 bol po zmene témy v podstate nepoužiteľný.



Obrázok 15 Priebeh základného SRP modelu bez detektora driftu a adaptivnych nastavení

4.5.4. Learn PPNSE

Taktiež bol použitý PPNSE klasifikátor s troma variantami teda jednou neadaptívnou a dvomi adaptívnymi metódami.

Tabuľka 14 Porovnanie výsledkov LPPNSE modelov v úlohe klasifikácie falošných správ pri dvoch témach

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.6739	0.7713	0.7611
Návratnosť	0.6569	0.7382	0.7012
Presnosť	0.6532	0.7676	0.7711
F-miera	0.6550	0.7526	0.7345

Najlepšie hodnoty dosahoval adaptívny model Hoeffding Adaptive Tree ktorý v sebe obsahuje ADWIN detektor driftu. Tento prístup v kombinácii s PPNSE klasifikátorom dosahoval zároveň najlepšie výsledky zo všetkých modelov použitých pri tomto datasete.

4.5.5. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie a zároveň aj podľa F-miery dosahoval model LPPNSE v kombinácií s adaptívnym modelom Hoeffding Adaptive Tree.

Tabuľka 15 Porovnanie výsledkov všetkých modelov pri detekcií falošných správ pri dvoch témach

Metriky	NB DEF	HAT + NBA	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.6937	0.7332	0.6982	0.7695	0.7670	0.7239	0.7654	0.7663	0.7565	0.6739	**0.7713**	0.7611
Návratnosť	0.4010	0.7216	0.4098	0.7383	0.7398	0.6931	0.7317	0.7312	0.7206	0.6569	**0.7382**	0.7012
Presnosť	0.8875	0.7152	0.8920	0.7648	0.7599	0.7134	0.7615	0.7634	0.7526	0.6532	**0.7676**	0.7711
F-miera	0.5524	0.7184	0.5616	0.7514	0.7497	0.7031	0.7463	0.7470	0.7362	0.6550	**0.7526**	0.7345

4.6. Detekcia falošných správ zo zmiešaného viac-témového datasetu

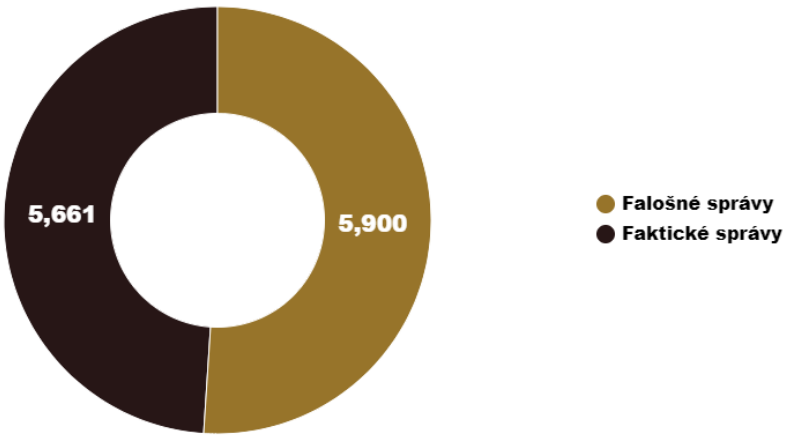
Pre tento experiment boli použité tri datasety, z toho boli dva zamerané na COVID19 a jeden bol zameraný na falošné správy s rôznou tematikou, hlavne politicky orientovanou, ktorý neobsahoval príspevky o COVIDe-19. Pre otestovanie ako sa modely budú správať pri viacerých zmenách témy, nie len pri jednej veľkej zmene ako pri predošlom experimente, boli pre tento experiment vyselektované konkrétne príspevky podľa zadaných kritérií. Následne boli tieto príspevky zoradené podľa témy aby bolo overiteľné či došlo k zmenám resp. ku zhoršeniu presnosti klasifikácie pravé v týchto bodoch.

Dokopy bolo vyselektovaných osem tém, ktoré obsahovali resp. neobsahovali zadané slová. Z covid datasetu boli vyselektované určité príspevky a z nich bola následne vytvorená téma. Konkrétne prvá téma z tohto datasetu obsahovala iba príspevky, ktoré obsahujú slovo vakcína. Ďalšia téma obsahovala iba príspevky, ktoré obsahujú slovo nákaza resp. infekcia (infection). Ďalšia téma obsahovala iba príspevky spomínajúce Čínu v spojení s COVID-19 a v poslednej téme boli príspevky, ktoré hovorili o smrti a úmrtiach spojených s COVID-19. Samozrejme sa týmto selektovaním niektoré príspevky dostali do viacerých kategórii preto boli príspevky s duplicitnými textami odstránené a ponechané iba v jednej z kategórii. Z politicky zameraného datasetu bola taktiež vyberaná selekcia príspevkov podľa kritérií a následne z nich boli tvorene samostatné kategórie resp. témy.

Ako prvá bola vytvorená téma spomínajúca rasizmus ale zároveň bola určená podmienka aby príspevky v tejto téme neobsahovali ostatné kľúčové slova z iných tém, ktoré sa budú ďalej používať teda napríklad aby sa v týchto príspevkoch o rasizme nehovorilo zároveň o armáde, keďže armáda je ďalšia téma, ktorá bola vytváraná pre experiment. V ďalšej téme, teda téme spomínajúcej armádu boli okrem tejto špecifikácie určené aj podmienky aby v príspevkoch bola spomenutá konkrétne vojna a Amerika pre zníženie počtu príspevkov a zároveň boli určené rovnaké podmienky ako pri predošlej téme teda aby príspevky o armáde nespomínali rasizmus a ďalšie slová používane pre ostatné témy. Posledné dve témy boli tvorené z príspevkov spomínajúcich zdravotníctvo a nakoniec vieru, taktiež boli určené podmienky aby nespomínali predošlé kľúčové slová teda rasizmus a armádu. Z týchto kritérií, ktoré určovali čo príspevky musia a nesmú obsahovať bolo vytvorených dokopy osem tém a bol z nich vytvorený novy dataset, pričom témy nasledovali v nasledujúcom poradí: rasizmus, vakcinácia, zdravotníctvo, nákaza (spojené s témou COVID), viera, Čína (spojené s témou COVID), US armáda a vojny, smrť a úmrtia (spojené s témou COVID). Témy boli cielene zoradené tak, aby sa striedali príspevky z politického datasetu s témami z COVID datasetov. Následne bol tento dataset použitý pri modeloch a boli skúmané výsledky a priebeh klasifikácie príspevkov. Podľa ADWIN detektoru bolo identifikovaných 11 driftov v bodoch: 4095,

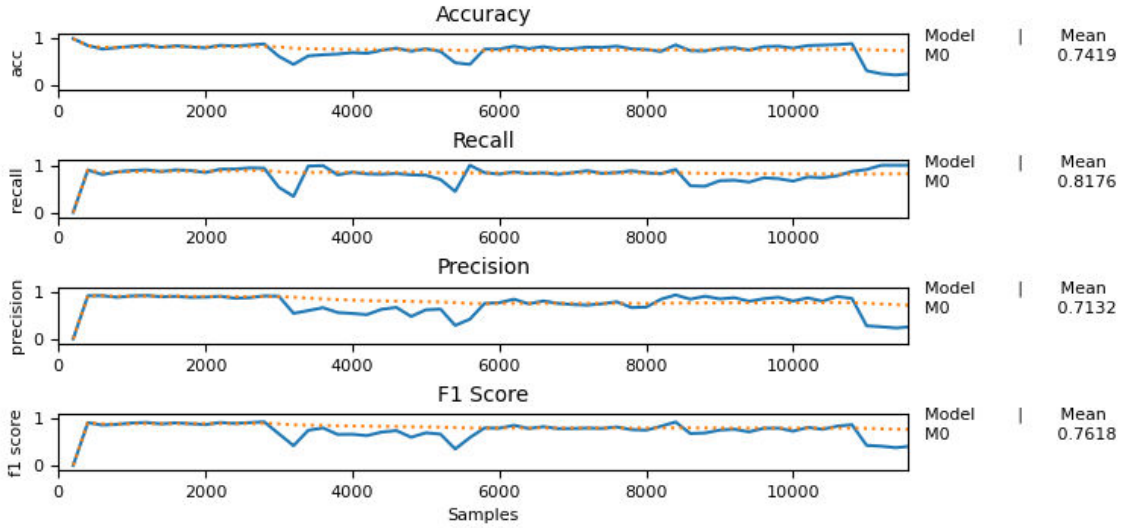
4607, 4671, 5119, 5759, 5951, 6399, 7487, 9055, 9983, 10143 okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.

Zmeny tém nastali v bodoch: 2895 z témy rasizmu na tému vakcinácie, 3599 z témy vakcinácie na tému zdravotníctva, v 5150 z témy zdravotníctva na nákazu spojnú s COVID-19, v 5619 z témy nákazy na tému viera, 7984 z témy viery na tému Čína, V 8350 z témy Čína na tému US armáda a vojna, v 10816 z témy US armády na tému úmrtí spojených s COVID-19



Obrázok 16 Rozloženie predikovaného atribútu

4.6.1. Základný model pre porovnávanie



Obrázok 17 Priebeh modelu Naive Bayes v úlohe klasifikácie falošných správ pri viacerých témach

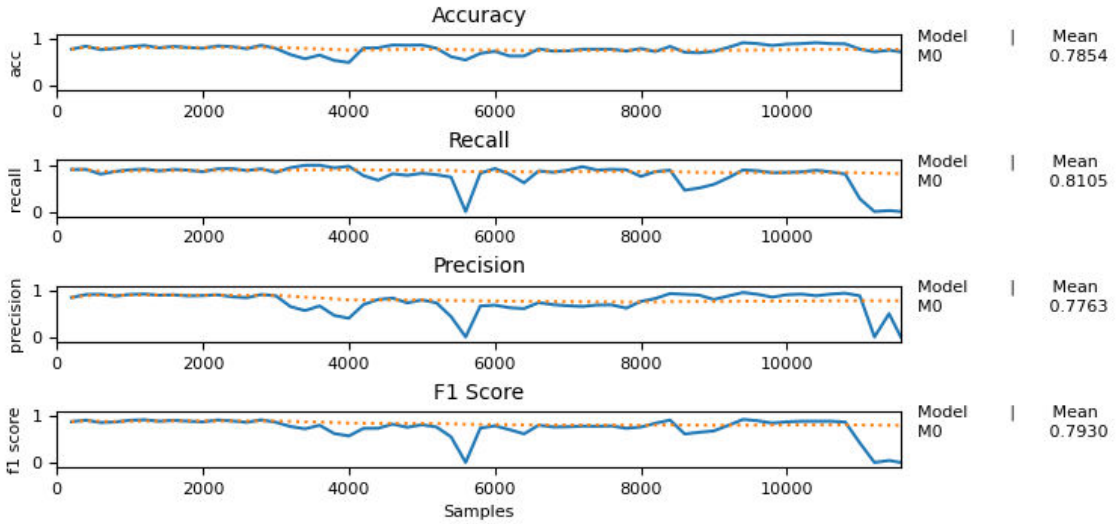
Z priebehu grafu základného modelu Naive Bayes na obrázku 20 je viditeľne, že prvé problémy vznikli pri zmene témy vakcinácie na tému zdravotníctvo, ďalšia znížená úspešnosť klasifikácie je viditeľná približne pri príspevku 5600. teda pri zmene z témy nákazy na vieru. Najväčší úpadok v presnosti a F-miere je viditeľný po prechode z témy US armády na tému úmrtí spojených s COVID-19. Ostatné prechody medzi témami nemali ani pri tomto neadaptívnom modely veľký vplyv na jeho presnosť teda neboli dostatočne výrazne alebo nebol počet príspevkov z danej témy dostatočne veľký na to aby zmeny vykázali viditeľný pokles na grafe. Úspešnosť klasifikácie základného modelu je 74% a jeho F-miera je 76%, výsledky sú spriemerované na celom streame.

4.6.2. Hoeffding Adaptive Tree

Pre porovnanie priebehu a výsledkov základného modelu bol použitý model Hoeffding Adaptive Tree, ktorý používa ADWIN detektor zmien. Pri adaptívnom modely je viditeľný progres v efektívnosti keďže úspešnosť klasifikácie je 79% a F-miera taktiež 79%.

Zaujímavý jav je možné sledovať pri grafe modelu na obrázku 21, kde vidno veľmi podobný priebeh ako pri neadaptívnom modely čiže problémy nastali približne pri rovnakých bodoch. V porovnaní s predošlým experimentom je taktiež badateľné, že adaptívny model využívajúci ADWIN je síce efektívnejší ako neadaptívny ale zmena v metrikách nie je tak výrazná ako pri predošlom experimente. Príčinou tohto javu môže byť pomalšia rýchlosť preučenia modelu resp. rýchlosť reakcie na drift, keďže je v použitom datasete menšie množstvo príspevkov v rámci jednej témy ale väčšie množstvo samotných tém, čiže zmien, ktoré môžu viesť k driftu.

Nastavenia pri tomto modely boli rovnaké ako pri modely z predošlého experimentu. Pri predošlom experimente v ktorom došlo iba k jednej veľmi veľkej zmene tém je veľmi dobre viditeľné ako neadaptívny model zaostáva za adaptívnym, v predošlom experimente ale platilo, že dataset obsahoval 8500 príspevkov v obidvoch témach a taktiež bolo viditeľné, že sa adaptívne modely úplne adaptujú až po 500-1000 príspevkoch po zmene témy. V tomto experimente dochádza ku zmene tém po oveľa menšom množstve príspevkov a teda modely sa tak rýchlo nestihnú úplne adaptovať resp. tieto zmeny tak rýchlo detegovať a preto je badateľný menší rozdiel v efektívnosti medzi adaptívnym a neadaptívnym modelom.



Obrázok 18 Priebeh modelu Hoeffding Adaptive Tree s použitím Naive Bayes Adaptive v úlohe klasifikácie falošných správ pri viacerých témach

4.6.3. Streaming Random Patches

Pre porovnanie jednotlivých detektorov driftu a porovnaní voči základnému modelu bez detektora driftu bol použitý opäť SRP klasifikátor.

Tabuľka 16 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie falošných správ pri viacerých témach

Metrika	Default	ADWIN	KSWIN	EDDM	PageHinkley	HDDM_A	HDDM_W
Úspešnosť klasifikácie	0.7429	0.8264	0.8369	0.8217	0.8304	0.8338	0.8341
Návratnosť	0.8196	0.8236	0.8432	0.8158	0.8535	0.8543	0.8457
Presnosť	0.7152	0.8324	0.8367	0.8298	0.8197	0.8244	0.8305
F-miera	0.7638	0.8280	0.8399	0.8228	0.8362	0.8391	0.8380

Pozerajúc sa na úspešnosť klasifikácie a F-mieru je v spojení s SRP klasifikátorom najefektívnejší KSWIN drift detektor. Oproti základnému SRP modelu bez detektora driftu a adaptívnych nastavení je efektívnejší o 10% v presnosti klasifikácie a o 8% v F-miere, výsledky sú spriemerované na celom streame.

4.6.4. Learn PPNSE Klasifikátor

Tabuľka 17 Porovnanie výsledkov LPPNSE modelov v úlohe klasifikácie falošných správ pri viacerých témach

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.7318	0.8227	0.8243
Návratnosť	0.7425	0.8220	0.8354
Presnosť	0.7306	0.8260	0.8200
F-miera	0.7365	0.8240	0.8276

Pri PPNSE boli rozdiely medzi neadaptívnym a adaptívnymi modelmi viac výrazne aj napriek tomu, že oba modely využívajú ADWIN detektor driftu. Pri presnosti klasifikácie aj F-miere boli adaptívne modely presnejšie o približne 9%.

4.6.5. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie a zároveň aj podľa F-miery dosahoval model SRP v kombinácií s KSWIN detektorom driftu.

Tabuľka 18 Porovnanie výsledkov všetkých modelov pri detekcií falošných správ viacerých témach

Metriky	NB DEF	HAT + NBA	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.7419	0.7854	0.7429	0.8264	**0.8369**	0.8217	0.8304	0.8338	0.8341	0.7318	0.8227	0.8243
Návratnosť	0.8176	0.8105	0.8196	0.8236	**0.8432**	0.8158	0.8535	0.8543	0.8457	0.7425	0.8220	0.8354
Presnosť	0.7132	0.7763	0.7152	0.8324	**0.8367**	0.8298	0.8197	0.8244	0.8305	0.7306	0.8260	0.8200
F-miera	0.7618	0.7930	0.7638	0.8280	**0.8399**	0.8228	0.8362	0.8391	0.8380	0.7365	0.8240	0.8276

4.7. Detekcia tém pomocou LDA

Pre posledný experiment bol zvolený prístup s názvom Latentná Dirichletova Alokácia (LDA). Tento prístup bol použitý na kombinovanom datasete pričom LDA slúžila na to, aby sa dala určiť téma príspevku bez toho, aby bolo nutné príspevky jednotlivo filtrovať ako pri predošlom experimente resp. čítať príspevky a hodnotiť o akú tému sa jedná. Pomocou LDA metódy a top kľúčových slov popisujúcich každý príspevok boli určené okruhy resp. názvy tém opisujúce jednotlivé príspevky. Témy v datasete boli:

1) Muslims /Muslim State
2) Covid
3) Police officer / Shooting
4) Funding/ Money
5) President Obama
6) Society
7) Trump/ Republicans

Týchto sedem tém reprezentujúcich príspevky datasetu bolo usporiadaných v poradí aké je uvedené vo vymenovaní. Téma číslo 6 teda „Society“ bola pritom najviac obšírna téma. Ako pri predošlých kombinovaných experimentoch tak aj pri tomto môže existovať v rámci každej témy viacero pod-tém. Cieľom tohto prístupu je otestovať či sa na datasete vytvorenom a usporiadanom na základe výstupu z LDA podarí dostatočne dobre zobraziť rozdiely medzi adaptívnymi a neadaptívnymi modelmi v spojení s datovaním prúdmi. Podľa ADWIN detektoru bolo identifikovaných 16 driftov v bodoch: 607, 639, 735, 799, 1727, 2879, 3007, 3039, 3295, 5311, 5407, 5695, 5823, 7775, 7871, 8383 okrem samotných detekcií využíva ADWIN aj informácie varovného charakteru, ktoré predikujú potenciálne nastanie driftu.

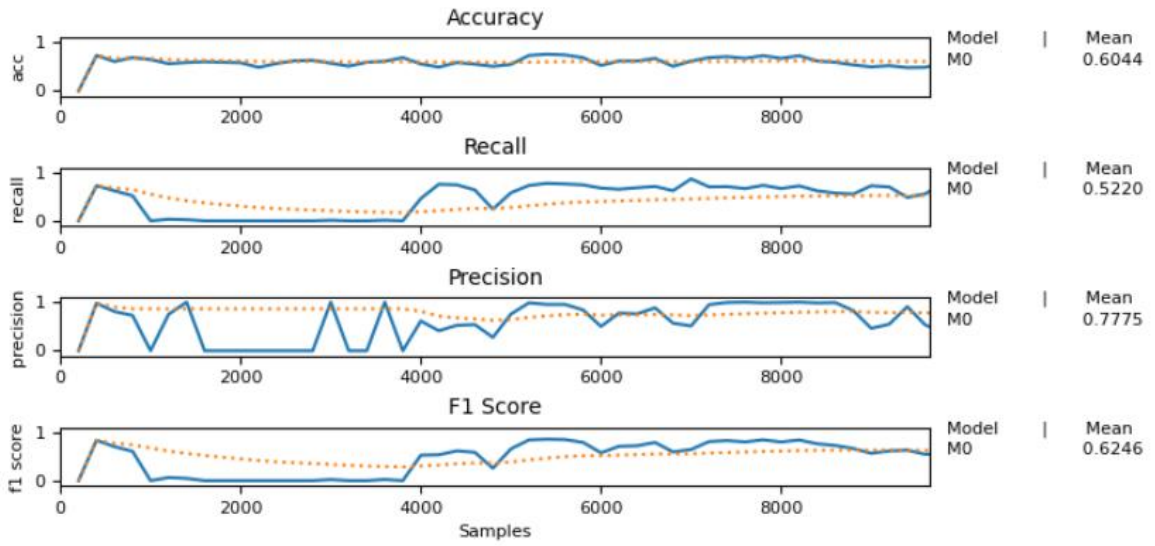
4.7.1. Základný model pre porovnávanie

Základný model Naive Bayes dosahoval úspešnosť klasifikácie 60% a F-mieru 62%, výsledky sú spriemerované na celom streame. Z grafu na obrázku 22 je viditeľné, že si základný model v podstate celú dobu udržiaval svoju úspešnosť klasifikácie približne na rovnakej hodnote teda okolo výsledných 60% pri F-miere je ale viditeľné, že pred dosiahnutím prvej tisícky príspevkov došlo k obrovskému úpadku F-miery takmer na nulu, pričom tento jav trval do približne štyri-tisíceho príspevku.

Podľa označení tém v datasete je možné zistiť, že prvá zmena témy nastala pri príspevku 773 a bola to konkrétne zmena z témy moslimov a moslimského štátu na tému COVID-19, tento prechod teda vysvetľuje pokles v F-miere viditeľný na grafe. Téma COVIDu-19 trvala až po príspevok 4883 pričom

v tomto bode sa téma menila na tému policajné zásahy. Približne v tomto bode na grafe je viditeľné taktiež zhoršenie v F-miere.

Ako posledná výraznejšie viditeľná zmena je cca v bode 8700 kedy nastáva zmena z témy spoločnosť na tému Trump / Republikáni, v F-miere je viditeľná klesajúca tendencia po tejto zmene. Nie je ale natoľko výrazná ako prvé dve spomenuté zmeny. Ostatné prechody z témy na tému nespôsobovali až tak výrazne poklesy ako tri spomenuté zmeny, to mohlo byť spôsobené prelínaním sa tém čo by v podstate znamenalo, že aj keď je daný príspevok v jednej konkrétnej téme, môže mať veľa spoločných znakov resp. málo odlišných znakov s príspevkami v inej téme. Ak je takýchto príspevkov v rámci jednej témy väčšie množstvo, zmeny z témy na tému nemusia v realite predstavovať badateľné zmeny a následkom toho ich ani model nebude považovať za dostatočné na to aby prešli z varovania na samotnú detekciu driftu.

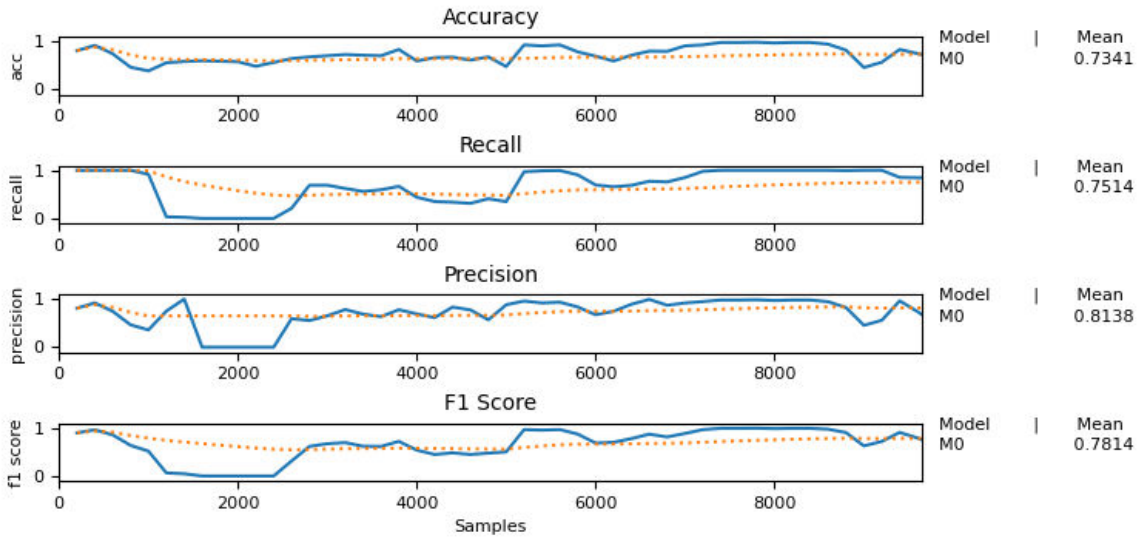


Obrázok 19 Priebeh modelu Naive Bayes v úlohe klasifikácie falošných správ (LDA)

Pre porovnanie so základným modelom bol použitý adaptívny model Hoeffding Adaptive Tree využívajúci NBA čo je vlastne adaptívna verzia základného modelu. Úspešnosť klasifikácie bola pri adaptívnom modely 73% a F-miera bola 78%, čo je oproti základnému modelu 13% zvýšenie v presnosti klasifikácie a 15% zvýšenie v F-miere.

Pri adaptívnom modely je viditeľný prvý pokles v F-miere v podobnom bode ako pri neadaptívnom modely, na rozdiel od neadaptívneho modelu trvá ale oveľa kratšie kým sa model v F-miere dostane opäť na dobré hodnoty, teda viditeľne sa na tému COVIDu-19 prispôsobuje, keďže dosahuje dobré výsledky už po dve tisícom príspevku. Viditeľný pokles na grafe pri zmene témy na COVID-19 opäť dokazuje, že adaptívnym modelom využívajúcim detektor driftu trvá niekedy aj do tisíc príspevkov kým sa na zmenu adaptujú, dôležité ale je, že k adaptácii dochádza. V prípade experimentov v tejto

diplomovej práci, ktoré používajú datasety obsahujúce občas len niekoľko tisíc záznamov je to výraznejší jav, ale pri reálnych okolnostiach v online prostredí kedy v dátových prúdoch pribúdajú milióny príspevkov je potenciálne adaptácia do tisíc príspevkov veľmi dobrá a nepredstavovala by ani len najmenší problém. To platí samozrejme za predpokladu, že by aj pri takom obrovskom množstve dát stále platila detekcia driftu do tisíc príspevkov od zmeny, čo nie je potvrdené.



Obrázok 20 Priebeh modelu Hoeffding Adaptive Tree s použitím Naive Bayes Adaptive v úlohe klasifikácie falošných správ (LDA)

4.7.2. Streaming Random Patches Klasifikátor

Najefektívnejšia bola kombinácia SRP s HDDM_A detektora driftu a najmenej efektívna bola kombinácia s EDDM detektorom driftu. Všetky kombinácie používajúce detekciu driftu boli ale výrazne lepšie ako základný model.

Tabuľka 19 Porovnanie výsledkov SRP v kombinácii s jednotlivými detektormi driftu v úlohe klasifikácie falošných správ (LDA)

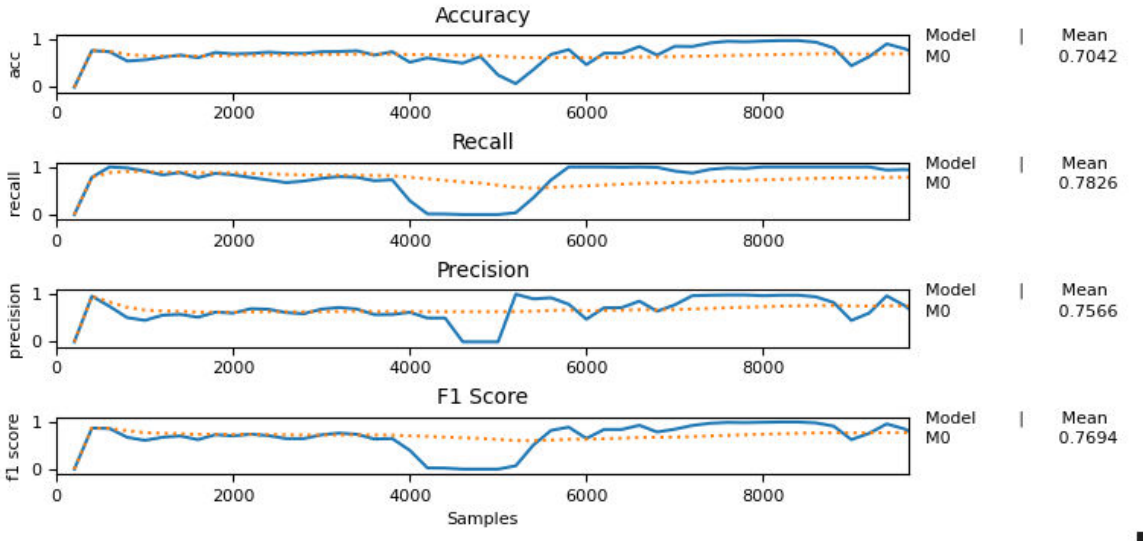
Metrika	Default	ADWIN	KSWIN	EDDM	Page Hinkley	HDDM_A	HDDM_W
Úspešnosť klasifikácie	0.6043	0.8070	0.8080	0.7853	0.8054	0.8116	0.8025
Návratnosť	0.5241	0.8637	0.8665	0.8482	0.8728	0.8604	0.8452
Presnosť	0.7778	0.8364	0.8359	0.8189	0.8286	0.8446	0.8429
F-miera	0.6262	0.8499	0.8510	0.8333	0.8502	0.8524	0.8441

4.7.3. Learn PPNSE Klasifikátor

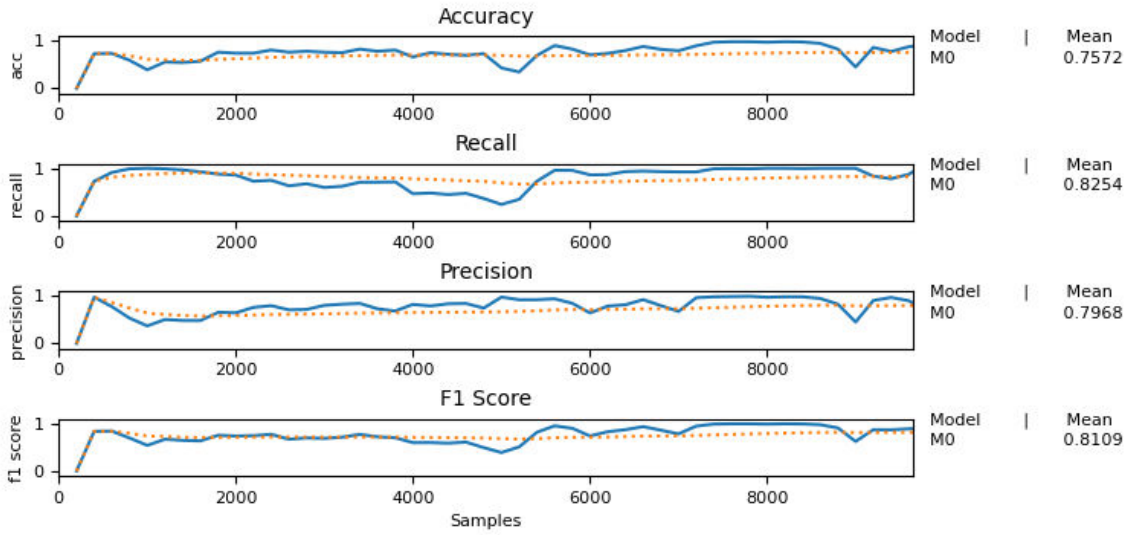
Pri poslednej použitej metóde PPNSE je zaujímavým výsledkom, že kombinácia PPNSE a Hoeffding Adaptive Tree dosahuje v presnosti klasifikácie a F-miere mierne horšie výsledky ako základný model používajúci Decision Tree, tento fakt sa vyskytuje iba v tomto experimente keďže v predošlých experimentoch platilo stále, že adaptívna verzia dosahovala lepšie výsledky.

Zaujímavým pozorovaním je taktiež, že Hoeffding Adaptive Tree bez kombinácie s PPNSE klasifikátorom dosahoval lepšie výsledky čiže by sa dalo tvrdiť, že PPNSE v tomto prípade zhoršil efektívnosť tohto modelu. Najefektívnejšia bola kombinácia s KNN-ADWIN, ktorá dosahovala najvyššiu úspešnosť klasifikácie aj F-mieru.

Metrika	**LearnPPNSEClassifier + Decision tree**	**LearnPPNSEClassifier + Hoeffding Adaptive Tree**	**LearnPPNSEClassifier + KNN-ADWIN**
Úspešnosť klasifikácie	0.7303	0.7042	0.7572
Návratnosť	0.7929	0.7826	0.8254
Presnosť	0.7824	0.7566	0.7968
F1-miera	0.7876	0.7694	0.8109



Obrázok 21 Priebeh LearnPPNSEClassifier s Hoeffding Adaptive Tree (LDA)



Obrázok 22 Priebeh LearnPPNSEClassifier s KNNADWIN (LDA)

Pri pohľade na grafy modelov Hoeffding Adaptive Tree v kombinácii s PPNSE a KKNADWIN s PPNSE je možne vidieť, že pri Hoeffding modely nastali problémy viditeľné v F-miere približne po štyri tisícom príspevku a následne sa miera začala zlepšovať približne medzi príspevkami 5500-5700 kedy došlo k zmene témy z policajných zásahov na financovanie. Tento problém u modelu používajúceho KNNADWIN nenastal a model sa vysporiadal so zmenami lepšie.

4.7.4. Porovnanie výsledkov

Z celkového porovnania výsledkov je viditeľné, že najlepšie výsledky podľa presnosti klasifikácie a zároveň aj podľa F-miery dosahoval model SRP v kombinácií s HDDMA detektorom driftu.

Metriky	NB DEF	HAT + NBA	SRP DEF	SRP + ADWIN	SRP + KSWIN	SRP + EDDM	SRP + PH	SRP + HDDMA	SRP + HDDMW	LPPNSE + DT	LPPNSE + HAT	LPPNSE + KNN ADWIN
Úspešnosť klasifikácie	0.6044	0.7341	0.6043	0.8070	0.8080	0.7853	0.8054	**0.8116**	0.8025	0.7303	0.7042	0.7572
Návratnosť	0.5220	0.7514	0.5241	0.8637	0.8665	0.8482	0.8728	**0.8604**	0.8452	0.7929	0.7826	0.8254
Presnosť	0.7775	0.8138	0.7778	0.8364	0.8359	0.8189	0.8286	**0.8446**	0.8429	0.7824	0.7566	0.7968
F-miera	0.6246	0.7814	0.6262	0.8499	0.8510	0.8333	0.8502	**0.8524**	0.8441	0.7876	0.7694	0.8109

Záver

V súčasnosti predstavujú falošné správy a dezinformácie jeden z najväčších problémov na sociálnych sieťach a celkovo v online priestore.

V tejto diplomovej práci je poskytnutý prehľad antisociálneho správania na webe a taktiež je poukázané na jeho závažnosť. Sú spomenuté rôzne druhy tohto správania a dopady aké má na ľudí či už ako na skupiny alebo jednotlivcov. Taktiež je podaný prehlaď aktuálne a historicky používaných metód a prístupov pre reguláciu a detekciu falošných správ na sociálnych sieťach.

V práci sú popísané experimenty, ktoré boli vykonané s cieľom porovnať efektívnosť klasifikácie falošných správ pomocou adaptívnych modelov v kombinácií s detektormi driftu z dátových streamov. Taktiež bolo cieľom otestovať hypotézu či budú neadaptívne modely oproti tým adaptívnym menej efektívne v klasifikácii falošných správ z dátových prúdov. Je možné zhodnotiť, že pri všetkých šiestich experimentoch sa adaptívne modely celkovo preukázali ako efektívnejšie. Najväčšie rozdiely možno sledovať pri experimente na kombinovanom datasete s dvomi témami, kde bol nedostatok neadaptívnych metód najviac viditeľný, keďže boli tieto metódy po zmene témy extrémne nepresné v klasifikácii falošných správ, zatiaľ čo adaptívne metódy dosahovali aj napriek zmene témy dobré, konštantné výsledky. Pri porovnaní detektorov driftu dosahovali všetky detektory vo väčšine prípadov podobné výsledky, čiže nie je možné jednoznačne určiť jeden najlepší prístup. Dôležité je ale spomenúť, že vo všetkých prípadoch dosahovali metódy disponujúce akýmkoľvek detektorom driftu a adaptívnym nastavením lepšie výsledky ako základné neadaptívne metódy, s ktorými boli porovnávané. Dá sa teda skonštatovať, že adaptívne prístupy sú prínosom v detekcii antisociálneho správania na webe. Taktiež je nutné poukázať na dôležitosť dátových prúdov využívaných v experimentoch, keďže umožňujú testovať modely dynamicky a nie staticky, čo bolo taktiež preukázané za veľmi dôležité. Využívaná knižnica scikit-multiflow slúžila pre experimenty dostatočne dobre, aj keď má svoje obmedzenia ako napríklad fakt, že väčšina modelov disponuje výlučne ADWIN detektorom driftu a nedá sa použiť iný detektor.

Celkovo je ale možné zhodnotiť, že problematika spojená s detekciou antisociálneho správania na webe je veľmi aktuálna a je nutné pokračovať v jej riešení prístupmi strojového učenia, keďže sa javí ako možnosť s najväčším potenciálom na úspech.

Zoznam použitej literatúry

[1]. P. Dizikes. [2018]. Study: On Twitter, false news travels faster than true stories. Massachusetts Institute of Technology News Office.

[2]. C. Shao, G.L. Ciampaglia, O. Varol, et al. [2018]. The spread of low-credibility content by social bots. Nature Communications 10, Article number 4787.

[3]. A. Bovet, H.A. Makse. [2019]. Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications 9, Article number 7.

[4]. BBC NEWS, Interview. [2018]. Cambridge Analytica planted fake news. **[Online].** Dostupné na internete: *https://www.bbc.com/news/av/world-43472347*

[5]. X. Zhang, A.A. Ghorbani. [2020]. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, Volume 57, Issue 2.

[6]. P. Ksieniewicz, P. Zyblewski, M. Choraś, R. Kozik, A. Giełczyk, M. Woźniak. [2020]. Fake News Detection from Data Streams. 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8.

[7]. P. Fichman, M. R. Sanfilippo. [2016]. Online Trolling and Its Perpetrators. Rowman & Littlefield Publishers 2016. ISBN 9781442238503.

[8]. L. Hautala. [2021]. Amazon's never-ending fake reviews problem, explained. **[Online]**. Dostupné na internete: *https://www.cnet.com/tech/services-and-software/features/amazons-never-ending-fake-reviews-problem-explained/*

[9]. S. B. Naeem, R. Bhatti, A. Khan. [2020]. Regular Feature: International Perspectives and Initiatives. Health Information & Libraries Journal, 38: 143-149.

[10]. E. Hollowood, A. Mostrous. [2020]. Fake news in the time of C-19 **[Online]**. Dostupné na internete: *https://www.tortoisemedia.com/2020/03/23/the-infodemic-fake-news-coronavirus/*

[11]. L. Frayer, [2020]. Blamed For Coronavirus Outbreak, Muslims In India Come Under Attack. WUIS - University of Illinois, Springfield.

[12]. H. L. Cheng, H. Y Kim, J. D. Reynolds, Y. Tsong, Y. J. Wong. 2021. COVID-19 anti-Asian racism: A tripartite model of collective psychosocial resilience. American Psychologist, 76(4), 627–642.

[13]. P. B. Brandtzaeg, A. Følstad. [2017]. Trust and Distrust in Online Fact-Checking Services Communications of the ACM. Volume 60, Issue 9, pp 65–71.

[14]. BBC NEWS, [2021]. Covid 'hate crimes' against Asian Americans on rise **[Online]**. Dostupné na internete: *https://www.bbc.com/news/world-us-canada-56218684*

[15]. J. Andersen, S.O. Søe. [2019]. Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. European Journal of Communication. Volume 35, Issue 2, pp 126-139.

[16]. T. Lyons. [2017]. Replacing Disputed Flags With Related Articles. **[Online]**. Dostupné na internete: *https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/*

[17]. Twitter Inc. [2021]. Permanent suspension of @realDonaldTrump **[Online]**. Dostupné na internete: *https://blog.twitter.com/en_us/topics/company/2020/suspension*

[18]. C. F. Bond, B. M. DePaulo. [2006]. Accuracy of Deception Judgments. Personality and Social Psychology Review. 10(3), pp. 214–234.

[19]. A. Wani, I. Joshi, S. Khandve, V. Wagh. [2021]. Evaluating Deep Learning Approaches for Covid19 Fake News Detection. Communications in Computer and Information Science, Volume 1402. ISBN 978-3-030-73696-5.

[20]. A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido, H. Liu. [2022]. Domain Adaptive Fake News Detection via Reinforcement Learning.

[21]. D.K. Jain, A. Kumar, A. Shrivastava. [2022]. CanarDeep: a hybrid deep neural model with mixed fusion for rumour detection in social data streams. Neural Computing and Applications.

[22]. R. M. Silva, T. A. Almeida. [2021]. How concept drift can impair the classification of fake news. Anais do IX Symposium on Knowledge Discovery, Mining and Learning, Rio de Janeiro. SBC, pp.121-128.

[23]. R. Mohawesh, S. Tran, R. Ollington, S. Xu. [2021]. Analysis of concept drift in fake reviews detection. Expert Systems with Applications Journal, Volume 169. ISSN 0957-4174.

[24]. K. Shu, D. Mahudeswaran, H. Liu. [2019]. FakeNewsTracker: a tool for fake news collection, detection, and visualization. Computational and Mathematical Organization Theory. 25. 10.1007/s10588-018-09280-3.

[25]. Twitter Inc. [n.d.]. How to use advanced search. **[Online]**. Dostupné na internete: *https://help.twitter.com/en/using-twitter/twitter-advanced-search*

[26]. T. Davidson. [2019]. Hate speech and offensive language. **[Dataset]**. Dostupné na internete: *hate-speech-and-offensive-language/data at master · t-davidson/hate-speech-and-offensive-language · GitHub*

[27]. E. Aghammadzada. [2021]. COVID19 Fake News Dataset NLP. **[Dataset]**. Dostupné na internete: *https://www.kaggle.com/datasets/elvinagammed/covid19-fake-newsdataset-nlp?select=Constraint_Val.csv*

[28]. P. K. Verma, P. Agrawal, R. Prodan. [2021]. WELFake dataset for fake news detection in text data **[Dataset].** Dostupné na internete: *https://zenodo.org/record/4561253*

[29]. J. Banks. [2010]. Regulating hate speech online. International Review of Law, Computers and Technology. Volume 24, Issue 3, pp. 233-239.

[30]. Twitter Inc. [n.d.] Hateful conduct policy **[Online].** Dostupné na internete: *https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy*

[31]. A. Hern. [2016]. Facebook, YouTube, Twitter and Microsoft sign EU hate speech code **[Online]**. *https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code*

[32]. P. Fortuna, S. Nunes. [2019]. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, Volume 51, Issue 4, Article Number 85, pp 1–30.

[33]. S. MacAvaney ,H. R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder. [2019]. Plos One, 14(8), Volume 14.

[34]. S. Girgis, E. Amer, M. Gadallah. [2018]. Deep Learning Algorithms for Detecting Fake News in Online Text. 13th International Conference on Computer Engineering and Systems (ICCES), pp. 93-97.

[35]. Scikit Multiflow Documentation. [n.d.]. Evaluate Prequential **[Online]**. Dostupné na internete:*https://scikitmultiflow.readthedocs.io/en/stable/api/generated/skmultiflow.evaluation.EvaluatePrequential.html*

[36]. S. Homayoun, M. Ahmadzadeh. [2016]. A review on data stream classification approaches. Journal of Advanced Computer Science & Technology. 5. 8. 10.14419/jacst.v5i1.5225.

[37]. J. Brownlee. [2017]. A Gentle Introduction to Concept Drift in Machine Learning **[Online].** Dostupné na internete: *https://machinelearningmastery.com/gentle-introduction-concept-drift-machine-learning/*

[38]. Scikit Multiflow Documentation. [n.d.]. API Reference. **[Online]**. Dostupné na internete: *https://scikit-multiflow.readthedocs.io/en/stable/api/api.html*

Prílohy

Príloha A: CD médium – diplomová práca v elektronickej podobe, prílohy v elektronickej podobe. Zdrojové kódy v ipynb súboroch (Jupyter Notebooky) a datasety použité pre experimenty.

Príloha B: Používateľská príručka

Príloha C: Systémová príručka