Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Korelace a kauzalita sentimentu extrahovaného z textu a tepové frekvence

Plzeň 2019 Milan Kuda

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 16. května 2019

Milan Kuda

Poděkování

Chtěl bych poděkovat pánům Ing. Romanu Moučkovi Ph.D. a Ing. Jaromíru Salamonovi za odborné vedení práce a cenné rady, které mi pomohly tuto práci zkompletovat.

Abstract

This master's thesis deals with the issue of extracting sentiment from texts and its comparison with sentiment predicated from the heart rate of the observed subject. The benefit of this thesis is that we could determine the subject's mood from the heart rate and adjust the behavior of computer systems accordingly when they are interacting with a human. To find a possible correlation between the predicated sentiment and the extracted sentiment, various texts from the Twitter social network and the measurements of heart rate from the Fitbit Charge HR wristband were used. The fusion of both data types and the usage of machine learning methods are also explained.

Abstrakt

Tato diplomová práce se zabývá problematikou extrakce sentimentu z textů a jeho porovnáním se sentimentem predikovaným na základě tepové frekvence pozorovaného subjektu. Přínos této práce tkví v tom, že bychom mohli určit náladu subjektu z tepové frekvence a podle toho upravit chování počítačových systémů při interakci s člověkem. Pro nalezení možné korelace mezi predikovaným sentimentem a extrahovaným sentimentem slouží texty ze sociální sítě Twitter a tepová frekvence změřená fitness náramkem Fitbit Charge HR. Vysvětlena je také fúze obou typů dat a použití metod strojového učení.

Obsah

1 Úvod

1.1 Motivace

Každý člověk během dne prožívá určité zážitky, které ovlivňují jeho chování a také tělesné funkce. Bylo by zajímavé zjistit, zda by bylo možné získat data o tepové frekvenci a k nim související emoce. Zjištění tepové frekvence je v této době jednoduché, umí to skoro každý chytrý telefon s fotoaparátem či nositelná chytrá elektronika. Jak ovšem zjistit, co člověk za celý den prožíval? Jako ideální forma se jeví si psát každou hodinu jakýsi digitální deník, kam by si člověk zapisoval své prožitky.

A k čemu by to vlastně bylo dobré? Pomocí tepové frekvence by mohlo jít odhalit, jakou má člověk náladu a například mu zpříjemnit prožití dne změnou prostředí. Dále by tato spojitost mohla být nápomocná ve zjištění, zda člověk netrpí mentálním onemocněním či se u něj nevyskytuje anomálie a neměl by vyhledat lékařskou pomoc.

1.2 Popis problému

V této práci jsou použity dva typy vstupních dat. Prvním typem je změřená srdeční frekvence pomocí fitness náramku Fitbit Charge HR. Pro tato data je potřeba navrhnout správný systém pro jejich filtraci, zpracování a uložení tak, aby se s nimi dalo dále pracovat při porovnání se sentimentem extrahovaným z dat.

Druhým typem dat je sentiment zaznamenaný pomocí sociální sítě Twitter, kde docházelo k přidávání příspěvků s popisem nálady člověka. Zpracování těchto dat bude provedeno pomocí jednoho z nástrojů pro extrakci sentimentu.

Oba typy dat již byly pořízeny během pilotního a kvaziexperimentu, jejichž popis bude uveden ve třetí kapitole.

Následně budou oba typy dat přiřazeny k sobě. Vzhledem k rozvoji strojového učení, neuronových sítí a umělé inteligence se využije těchto metod pro následné rozpoznání sentimentu z tepové frekvence.

1.3 Struktura práce

Diplomová práce je rozdělena do několika kapitol, kde každá kapitola představuje další krok k pochopení a vyřešení problému.

První kapitolou je Úvod do existujícího výzkumu (2), kde jsou popsány již nalezené a vyzkoušené metody, které se zabývaly podobnou tématikou. Z těchto materiálů bude následně čerpáno v dalších kapitolách.

Druhá kapitola Popis experimentu (3) a je zde popsán průběh experimentu a sběr dat.

Následuje kapitola Získání a zpracování sentimentu (4), kde je rozebrána extrakce sentimentu pomocí existujících nástrojů a jeho zhodnocení. Dále je zde srovnání nástrojů podle vhodnosti jejich použití v této práci.

Čtvrtá kapitola Tepové frekvence a její získání (5). Zde se zaobírám zpracováním časových řad a správnou reprezentací dat pro další použití.

Příprava dat (6) je název další kapitoly, ve které budou vysvětleny možnosti spojení sentimentu a tepové frekvence. Hlavní důraz je kladen na analytické a statistické metody.

V další kapitole Strojové učení (7) jsou rozebrány metody strojového učení a tato kapitola uzavře teoretickou část diplomové práce.

Následující kapitoly se zabývají praktickou částí. Další kapitolou je Analýza, návrh a implementace (8), kde je rozebrána základní struktura programu a poté jeho jednotlivé části podrobněji.

Předposlední kapitola Výsledky (9) se zabývá popisem získaných výsledků a popisem úspěšnosti použitých metod.

V kapitole Závěr (10) se čtenář seznámí se shrnutím dosažených výsledků, jejich možným uplatněním v praxi a dalším rozvojem této práce.

2 Úvod do existujícího výzkumu

Tato část diplomové práce klade důraz na nalezení a sepsání prací, které se zabývaly podobným tématem. Následně jsou zde popsány úvody do problematik srdečního rytmu, emocí a sentimentu, které budou v práci následně rozvedeny.

2.1 Existující práce

Tato diplomová práce vychází ze čtyř prací. Práce byly vybrány vzhledem k podobnosti témat, kde se autoři snažili odvodit spojitost mezi tělesnými funkcemi nebo případně detekovat stres z tělesných funkcí. Práce ECG-derived Blood Pressure Classification using Complexity Analysis-based Machine Learning [52] a Link between Sentiment and Human Activity Represented by Footsteps Experiment Exploiting IoT Devices and Social Networks [50] se zabývají metodami strojového učení a jejich použitím na data získaná pomocí měření tělesných funkcí. Další prací je Stress Detection Using Low Cost Hearth Rate Sensors [49], ve které je ukázáno, že i levná nositelná zařízení na měření tepové frekvence mohou identifikovat stres. Poslední prací je Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors, která se zabývá identifikací stresu pomocí měření tělesných funkcí u řidičů.

V práci [52] se autoři snaží o klasifikaci krevního tlaku z elektrokardiogramu (EKG) do sedmi kategorií (nízký tlak, normální, normální vysoký, mírně závažný, středně závažný, závažný a izolovaná systolická hypertenze). V práci byl navrhnut systém pro změření EKG a hodnoty krevního tlaku, zpracování EKG, jeho analýzu a následně došlo k aplikaci různých metod strojového učení. Z článku vychází nejlépe metody J48 [45] a Bagging [45]. V práci byla také použita metoda ensembling (tj. kombinace výstupů různých metod, aby bylo dosaženo lepších výsledků) [40]. Autorům se novým přístupem podařilo zjistit vztah mezi EKG a krevním tlakem.

Další práce [50] se zabývala vztahem mezi náladou člověka vyjádřenou sentimentem extrahovaným z textu a jeho aktivitou reprezentovanou počtem kroků. Sentiment byl extrahován z textů psaných na sociální síť Twitter a aktivitu reprezentoval počet ušlých kroků během dne. Vyhodnocení sentimentu bylo prováděno jak jinými lidmi, tak pomocí strojového učení a jako

nejlepší metoda se ukázala Random Forest [45].

V práci [49] probíhalo sledování tělesných funkcí pomocí senzoru měření délky doby mezi depolarizacemi srdečních komor (R-R) [17] a následně byl navržen algoritmus pro zjištění stresu. V této práci bylo dosaženo přesnosti zjištění stresu 74,60 %, a je tedy možné použít levné senzory pro detekci srdeční frekvence pro zjištění stresu.

V poslední práci [29] byla zkoumána možnost predikce stresu pomocí EKG a senzoru, který měřil galvanickou změnu kůže. Data byla převzata ze studie provedené Healey a Picard [19], která se zabývala měřením fyziologických změn u řidičů. Ve studii byla provedena extrakce HRV signálu z EKG a stres byl rozdělen do dvou kategorií (stres nebo klidový stav). Jako nejlepší klasifikátor byl vyhodnocen lineární SVM, kde F-míra dosahovala 72,68 % při použití HRV, 97,87 % při použití EKG a 98,36 % při EKG spolu s galvanickou změnou.

2.2 Srdeční rytmus

Srdeční rytmus se dá vyjádřit jako střídání systoly (stažení srdeční svaloviny) a diastoly (uvolnění srdeční svaloviny), ke kterému dochází pravidelně [62]. Srdeční rytmus je řízen pomocí převodního systému srdečního, který se skládá z pěti částí: sinusový uzel, síňokomorvý uzel, Hisův svazek, Tawarova raménka a Purkňova vlákna. Srdeční rytmus je za běžných okolností určován sinusovým uzlem nacházejícím se v pravé síni [26].

Činnost srdce je ovlivňována nervovým centrem, které je umístěno v prodloužené míše a Varolově mostu. Činnost je ovlivňována pomocí sympatických a parasympatických vláken. Sympatikus zvyšuje srdeční frekvenci a připravuje tělo na případný stres. Parasympatikus působí opačně, snižuje srdeční frekvenci a uvádí organismus do klidu [26].

Pro měření srdečního rytmu se využívají srdeční frekvence a tepová frekvence. Srdeční frekvence je měřená přímo na srdci pomocí přístrojů jako je například EKG. Tepová frekvence se měří pohmatem či přístroji například na tepně zápěstí a stanovuje se jako počet tepových vln za minutu [26].

Normální hodnota tepové frekvence se u dospělého člověka pohybuje okolo 70 tepů za minutu. Tepová frekvence je proměnlivá, závisí na několika faktorech, například na fyzické námaze. Reaguje i na psychické podněty, změnu prostředí, únavu ale i na příjem kofeinu a jiných podpůrných látek [26]. Tepové frekvence je též závislá na věku, u malých dětí se pohybuje okolo 120 tepů za minutu a se stářím klesá i pod 60 tepů za minutu [8]. Pokud se tepová frekvence pohybuje nad 100 tepů za minutu, jedná se o tachykardii,

a jestliže se pohybuje pod 60 tepy za minutu, mluví se o bradykardii [55].

2.3 Emoce

Emoce je považovány za nejjednodušší zážitek, který slouží k uspokojení či neuspokojení hlavních potřeb organismu. Emoce slouží k vyhodnocení podnětů vycházejících z nitra organismu nebo vnějších podnětů. Taktéž umožňují rychlou reakci na měnící se podmínky a adaptaci na ně [53].

Největší částí nervového systému člověka je velký mozek, jehož součástí je limbický systém, který je zodpovědný za chování jedince, jež vede k přežití jeho a rodu. Chování k okolí lze rozdělit do dvou částí, za prvé apetitivní (přibližovací) chování, kdy jedinec vyhledává to, co má pro něj pozitivní biologickou hodnotu. Druhé chování se nazývá averzivní (únikové). V tomto případě se jedinec snaží vyhnout negativním událostem. Dalším důležitým oddílem mozku je mezimozek, při jehož dráždění vzniká kromě pocitů žízně a sytosti také agrese nebo strach [26].

Stresem se rozumí stav organismu, který vzniká působením stresoru. V tomto stavu jsou mobilizované obranné a nápravné systémy důležité pro přežití zátěžové situace. Stres se dělí na eustres (příjemný) a distres (nepříjemný). Příkladem eustresu může být výhra v loterii či splnění přání. Naopak distres může být způsoben například chorobou [26].

Stresorem se rozumí podněty, které vyvolávají určitý sled nervové, hormonální nebo celkové reakce organismu. Stresory jsou děleny do následujících kategorií:

- *fyzikální stresory* - chlad, hluk, vibrace
- *chemické stresory* - hlad, žízeň, jedy
- *bolest*
- *komplexní stresory* - nové prostředí, fyzická námaha
- *individuální stresory* - duševní vypětí, nedostatek spánku, ztráta milované osoby
- *stresory z nemoci* - hospitalizace, prostředí nemocnice
- *stresory skupinové* - školní, rodinné či sousedské prostředí
- *sociální stresory* - nezaměstnanost, pracovní nároky, odchod do důchodu

Stresory ovlivňuji funkci srdečního svalu prostřednictvím sympatiku a srdeční sval odpovídá zvýšením tepové frekvence. Toto zvýšení tepové frekvence je podobné situaci při fyzické zátěži. Nadměrným působením stresoru či při jeho dlouhodobém vlivu dochází k poruchám srdečního rytmu [26].

Stres tedy můžeme odečíst z fyziologických změn u člověka, nejčastějšími pozorovatelnými změnami jsou [49]:

- *Galvanická kožní reakce* - změny v elektrické vodivosti kůže; během stresu dochází k pocení a dochází ke snížení odporu kůže [51]
- *Elektromyogram* - měření elektrické aktivity svalů. Kvůli stresu dochází ke změnám ve stahování svalů a stres může být touto změnou identifikován [33]
- *Teplota kůže* - ve stresových situacích dochází také ke zvýšení teploty kůže [65]
- *Elektrická aktivita srdce* - nejčastěji používané metody pro zjištění stresu. Stres můžeme určit jednak z EKG, tepové frekvence či proměnlivosti tepové frekvence [58]

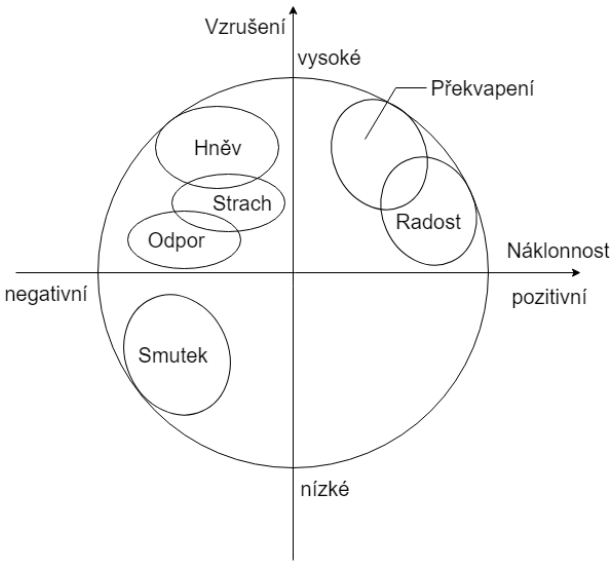
2.4 Sentiment

Emoce neovlivňují člověka pouze po fyzické stránce, ale projevují se i na jeho psychice. Ovlivňují tedy i to, jak se daný člověk vyjadřuje. Sentiment lze definovat jako autorův postoj, názor či vyjádření emocí nad určitým tématem. Rozdíl mezi sentimentem a názorem je ten, že názor pokaždé reflektuje přístup k určité problematice. Na druhou stranu sentiment zahrnuje i pocity či emoce, které nemusí vždy být spojené s daným tématem [61].

Detekcí sentimentu z psaného textu se zabývá obor Zpracování přirozeného jazyka (NLP - Natural Language Processing). Existují dva způsoby jak analyzovat sentimentu textu, a to kategorický model a dimenzionální model. V kategorickém modelu je sentiment vyhodnocován podle počtu výskytů elementů vyjadřující určitou emoci. Tento model je náchylný k chybám, protože je snadné použít špatnou klasifikací emocí. Z tohoto důvodu se začala používat metoda, kterou zavedl Paul Ekman [46]. On a jeho kolega Wallace Friesen navrhli rozdělení emocí do šesti základních skupin na základě pokusu s rozpoznáváním emocí na fotografiích [44]. Skupiny emocí jsou následující: hněv, smutek, radost, odpor, strach a překvapení.

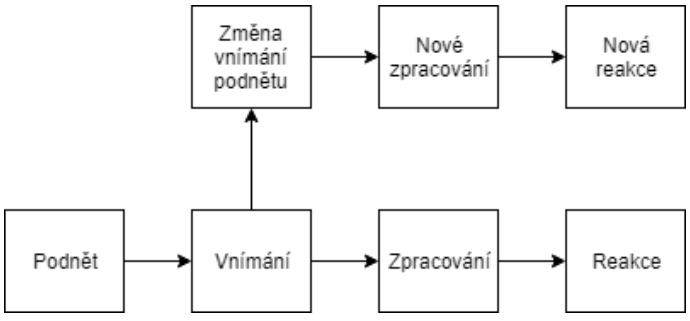
Dimenzionální model (obrázek 2.1) emocí byl vytvořen roku 1980 Jamesem Russelem. Model je definován jako dvoudimenzionální kruhový (circumplex) a je založen na předpokladu, že určité emoce si jsou bližší než

ostatní a všechny emoce mohou být vyjádřeny kombinací dvou základních dimenzí. Jednotlivé dimenze jsou tvořeny z náklonnosti a vzrušení [48]. V tomto modelu tedy není emoce jednotlivých elementů kategorizována, ale je vyjádřena polohou v prostoru. Nevýhodou může být, že takto vyjádřená emoce nemusí být čitelná lidmi [46].



Obrázek 2.1: Šest základních emocí v dvoudimenzionálním prostoru [24]

V anglických textech se můžeme setkat s pojmem opinion mining, který znamená totéž co analýza sentimentu. Pokud se budeme snažit "vydolovat názor" z příspěvků na Twitteru, je nutné počítat s tím, že emoce subjektu zde nejsou vyjádřené ihned a přímo. O tom, jaká bude reakce na stres nerozhoduje nejen stresor, ale i jeho zpracování, které může subjekt vědomě ovlivnit (obrázek 2.2). Člověk má tedy volbu, zda stresu podlehne nebo svojí reakci zmírní, a tím změní vnímání stresoru [43].



Obrázek 2.2: Změna zpracování stresoru

3 Popis experimentu

Pro tuto práci bude nutné získat data, která budou následně zpracována. Proto tato kapitola popisuje druhy experimentů a následně obsahuje popis provedených experimentů pro sběr dat obsahujících tepovou frekvenci a sentiment.

3.1 Druhy experimentů

Experiment neboli vědecký pokus je objektivní pozorování jevů, které se vyskytují v kontrolovaných podmínkách. Pomocí experimentu můžeme ověřit či vyvrátit platnost předpokladu a naleznout možné souvislosti.

3.1.1 Pilotní experiment

Pilotní experiment je předběžná studie malého rozsahu, která má za úkol zjistit proveditelnost, časové a finanční nároky, nepříznivé případy a zlepšení vůči studovanému předpokladu před zahájením výzkumného projektu. Pilotní experiment má tedy zajistit, aby v návrhu řízeného experimentu nechyběly důležité poznatky například pro sběr dat, a tím se zamezilo opakování řízeného experimentu. [22]

3.1.2 Kvaziexperiment

Kvaziexperiment je empirická studie používaná k odhadnutí dopadu předpokladu na sledovanou skupinu, kde nelze náhodně přiřadit sledovaným objektům experimentální podmínky. Kvaziexperiment tedy sdílí podobnosti s řízeným experimentem, ale liší se právě v absenci náhodného přiřazení či kontroly. Z těchto důvodů v tomto typu experimentů chybí kontrolní skupina. Místo toho kvaziexperiment typicky dovoluje výzkumníkům kontrolovat přiřazení účastníků, ale musí být využito jiného kritéria než náhodného přiřazení. Kvaziexperiment slouží hlavně k nalezení vztahů mezi proměnnými a má nižší míru validity než řízený experiment. [11]

3.1.3 Řízený experiment

Řízený experiment se zabývá zjišťováním vztahů mezi příčinou a následkem. Vychází z porovnávání mezi experimentální skupinou a kontrolní skupinou,

které by se měly lišit pouze v hledaném účinku. Na rozdíl od kvaziexperimentu se zde používá náhodné přiřazení buď do experimentální, či kontrolní skupiny. Náhodné rozdělení do skupin zajišťuje, že případné rozdíly ve skupinách budou dílem náhody. [2]

3.2 Provedení experimentů

V této diplomové práci budou použita data nasbíraná během pilotního a kvaziexperimentu. V obou experimentech docházelo ke sběru tepové frekvence pomocí nositelné elektroniky a sentimentu, který je extrahován z textu psaných příspěvků na sociální síť Twitter.

3.2.1 Pilotní experiment

Pilotní experiment měl jediného účastníka a délka experimentu byla 2x 50 dní. Účastníkem byl 35-letý muž s kontrolovaným vysokým tlakem.

Měření tepové frekvence

Pro měření byla použita dvě různá zařízení, jmenovitě Fitbit Charger HR a Basis Peak. Měření probíhalo nepřetržitě, pouze s přestávkami pro nabíjení. Tepová frekvence je u zařízení Fitbit Charger HR měřena náhodně každých 5 až 15 sekund po měření trvá dalších 5 sekund. U Basis Peak je měření prováděno 32x za minutu a zařízení provádí agregaci dat 1x za minutu.

Sentiment

Sentiment byl vyjadřován v anglickém jazyce pomocí příspěvků na sociální síť Twitter, tudíž délka příspěvku nesměla přesáhnout 140 znaků (toto omezení bylo dáno samotnou sociální sítí). Příspěvky obsahovaly vyjádření autorových pocitů pomocí textu implicitně obsahující sentiment. A také obsahovaly explicitní označení #p pro pozitivní a #n pro negativní sentiment. Text příspěvku bude sloužit pro extrakci sentimentu a označení je pouze pro kontrolu.

K psaní příspěvků docházelo každých 45 minut s maximální hodnotou 21 příspěvků denně o víkendu (od 9:00 do půlnoci) a 23 příspěvků pro pracovní dny (od 7:30 do půlnoci). Minimálně však bylo vyžadováno 20 příspěvků denně.

3.2.2 Kvaziexperiment

Po provedení pilotního experimentu byl připraven kvaziexperiment, který byl rozšířen na sedm účastníků (4 ženy, 3 muži, věk 19 až 25 let s průměrnou hodnotou 20,25 let a směrodatnou odchylkou 1,98 roku) a byla zkrácena doba experimentu. Ta nyní činila celkově 14 dní, ale pouze v 10 dnech bylo prováděno měření, viz obrázek (obrázek 3.1). Výběr účastníků pro experiment byl proveden mezi studenty studující stejný obor a rodný jazyk všech účastníků byla čeština. Během experimentu byly zaznamenány následující údaje: pohlaví, věk, váha a výška účastníka, název Twitter a Fitbit účtů, sériové číslo náramku a datum startu experimentu.

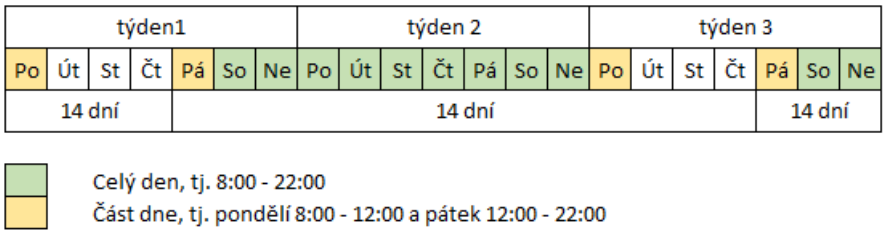
Měření tepové frekvence

Měření tepové frekvence probíhalo opět pomocí nositelné elektroniky, pouze pomocí Fitbit Charger HR. Každému zařízení byl přidělen technický účet (bodyinnumbers01-04). Sběr dat probíhal nepřetržitě, minimálně však od 8:00 do 22:00.

Sentiment

Sentiment v kvaziexperimentu byl vyjadřován v rodném jazyce pomocí sociální sítě Twitter. Délka příspěvku byla nově omezena 280 znaky (omezení sociální sítě). Pro každý náramek byl vytvořen účet na sociální síti (bodyinnumbers01-04) a příspěvky obsahovaly vyjádření autorových pocitů pomocí textu implicitně obsahující sentiment. Příspěvky opět obsahovaly explicitní označení #p pro pozitivní a #n pro negativní sentiment.

K psaní příspěvků docházelo každých 60 minut mezi 8:00 a 22:00 a bylo vyžadováno přesně 15 příspěvků za den.



Obrázek 3.1: Časový rámec kvaziexperimentu

4 Získání a zpracování sentimentu

Předchozí kapitola byla věnována sběru dat, v této je popsáno co je to sentiment, jak je možné jej analyzovat, problémy analýzy a jaké přístupy a nástroje lze použít.

4.1 Analýza sentimentu

Jak již bylo napsáno v druhé kapitole, sentiment se definuje jako autorův postoj, názor nebo vyjádření emocí nad určitým tématem (2.4). Bylo zde popsáno i základní rozdělení emocí a popsán dimenzionální model (2.3).

Zjištění sentimentu pomocí metod zpracování přirozeného jazyka ovšem není bezchybné a má určitá úskalí. Hlavním z nich je struktura jazyka, kde mohou být jednotlivá vyjádření nejasná či špatně formulována. Existuje pět základních problémů při analýze sentimentu [27].

- *rozlišení smyslu* - Stejná slova mohou mít různý význam podle kontextu, např. přídavné jméno malý může být vnímáno jak pozitivně (malá spotřeba benzínu) tak negativně (malý zavazadlový prostor)
- *porovnávání* - S porovnáním se často setkáváme zejména v recenzích "výdrž baterie mobilu Y je lepší než mobilu X". Vyskytuje se zde pozitivní slovo lepší, ale může být složité určit, jaká je klíčová informace tohoto porovnání.
- *negace* - Negace může úplně změnit smysl věty. Máme-li dvě věty: "je velká šance, že se telefon rozbije" a "je velká šance, že se telefon nerozbije", je na první pohled zřejmé, že negace úplně změnila smysl věty.
- *intenzita* - V jazyku můžeme narazit i na sílu názorů (mírný nebo silný) a následně může být problematické určit, jak má být výsledek klasifikován, pokud používáme více kategorií.
- *sarkasmus* - Posledním problémem je zjištění sarkasmu. Ten může špatně či úplně nečitelný, ale na druhou stranu může být nápomocen při podrobnější analýze sentimentu.

Pro potřeby této práce bude nutné programově určit sentiment z psaného textu a jeho zařazení do jedné z kategorií. Proto bude nutné se podívat na možnosti extrakce sentimentu z textu a vybrat vhodnou variantu pro tuto práci. Velmi časté je rozdělení do pěti skupin (velmi negativní, negativní, neutrální, pozitivní, velmi pozitivní). Občas se ovšem pro zjednodušení a lepší shodu se slovníky tato kategorizace zjednodušuje pouze na pozitivní a negativní. [30].

Analýzu sentimentu můžeme rozdělit podle velikosti zkoumaného textu a rozlišujeme tři základní úrovně [27, 28].

- *Analýza celku* - Analýza celku nebo dokumentu je nejednodušší forma klasifikace, kde se celek bere jako základní jednotka informace. Tato forma klasifikace není vhodná pro dokumenty, které obsahují názory na různá témata, jako jsou blogy či fóra.

- *Analýza věty* - Větná analýza je považována za nejpodrobnější analýzu dokumentu. Každá věta je brána jako samostatná jednotka se svým názorem. Pro každou větu je tedy zjištěn její sentiment. Analýza na úrovni vět má dva základní procesy, a to klasifikaci objektivity a klasifikaci sentimentu.

  V klasifikaci objektivity nás zajímá názor autora. Věty mohou být buď objektivní, nebo subjektivní. V objektivní větě se nalézají fakta, ale nevyjadřuje názor. V subjektivních větách lze nalézt názor autora. Objektivní věty (věty konstatující fakta) by neměly hrát žádnou roli v polaritě článku a mohou být vyloučeny.

  Druhou částí je klasifikace sentimentu, kde je sentiment věty urče podle slov, která vyjadřují buď pozitivní nebo negativní zabarvení věty.

- *Analýza aspektu* - Jak analýza dokumentu, tak analýza vět nezjišťují, co se ve skutečnosti lidem líbí a co ne. Místo zkoumání větných konstrukcí se analýza aspektu zaměřuje přímo na názor. Tato analýza je postavena na myšlence, že názor se skládá ze sentimentu (pozitivní nebo negativní) a cíle názoru. Například věta "Kvalita hovorů s iPhone je dobrá, ale výdrž baterie je krátká" obsahuje dva aspekty, a to kvalitu hovoru a výdrž baterie. Cílem jsou tedy kvalita hovoru, jejíž sentiment je kladný a výdrž baterie, která má záporný sentiment. [20]

  Analýza aspektu tedy produkuje strukturovaný souhrn jednotlivých názorů a takto převádí nestrukturovaný text na strukturovaná data pro další analýzy.

4.2 Metody a techniky pro extrakci sentimentu

Typy zdrojů vhodných pro extrakci sentimentu se rozdělují na čtyři základní: mikroblogy, diskuzní fóra, recenze produktů a vědecké citace. V těchto typech je možné nalézt sentiment autora. A. Yousif s kolegy provedl analýzu metod pro zjištění sentimentu vztahujících se k posouzení vědeckých citací. Mezi základní patřilo strojové učení (66,7 %), analýza pomocí slovníku (13,3 %), deep learning (13,3 %) a analýza pomocí klíčových slov (6,7 %), procentuální zastoupení těchto metod bylo zjištěno v rámci průzkumu. [66]

4.2.1 Strojové učení

Strojové učení se soustředí na vytváření modelů, které mají schopnost učit se z poskytnutých dat. Klasifikace sentimentu je prováděna ve dvou krocích. V prvním je trénován model pomocí trénovacích dat, kde je již sentiment klasifikován. V druhém kroku jsou klasifikována data, která nebyla použita pro trénink. [41]

Mezi používané metody patří například Naive Bayes, Support Vector Machine a Rozhodovací strom.

Naïve Bayes (7.3.2) je jednoduchý pravděpodobnostní model pro klasifikaci, který se opírá o předpoklad nezávislosti na příznacích (features) za účelem klasifikace vstupních dat. Support Vector Machine (7.3.1) je nepravděpodobnostní binární lineární klasifikátor. Jeho cílem je nalézt rovinu, která příznaky rozdělí tak, aby trénovací data náležela odlišným poloprostorům.

Poslední metodou je Decision Tree (Rozhodovací strom) (7.3.3), jenž rozděluje trénovací data na menší části za účelem identifikace vzorů, které mohou být použity pro klasifikaci.

4.2.2 Analýza pomocí slovníku

Analýza pomocí slovníku obsahuje již připravený seznam slov a jejich polaritu. Avšak díky složitosti jazyka má tato metoda omezení například s rozpoznáváním záporu. [66]

Jedním z možnách slovníků je SentiWordNet, který rozšiřuje slovník synonymické řady slovníku WordNet o hodnocení pozitivity, negativity a objektivity. [13]

4.2.3 Deep learning

Deep learning využívá ke klasifikaci sentimentu neuronových sítí (7.3.4). Neuronové sítě jsou tvořeny vrstvami a jednotlivé vrstvy se skládají z neuronů. Neurony jsou navzájem propojeny mezi vrstvami a každé propojení má svojí váhu. Pomocí nastavování vah se může síť naučit vykonávat různé úkoly podobně jako mozek.

Mezi neuronové sítě, které se využívají v extrakci sentimentu patří Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) a Long Short Term Memory network (LSTM). CNN byla původně navržena jako dopředná neuronová síť pro vizuální detekci. CNN je vhodná pro použití ve zpracování přirozeného jazyka díky přítomnosti konvolučních vrstev pro extrakci příznaků. Díky této vrstvě vzniká propojovací vzor mezi neurony přilehlých vrstev.

Rekurentní neuronové sítě (RNN) obsahují spojení, kde se mezi neurony tvoří řízený cyklus. Na rozdíl od dopředných sítí obsahuje tato síť "paměť" a během výpočtu výstupu je každý vstup ovlivněn předchozími vstupy. Sítě tohoto typu trpí problémem zvaným "vanishing gradient" kdy je změna gradientu příliš malá a vrchní vrstvy zůstanou beze změny. Long Short Term Memory (LSTM) network odstraňuje problém RNN pomocí speciálních uzávěrů k vnitřnímu stavu buňky, a tím zabraňují mizení gradientu. [23, 67]

4.2.4 Analýza pomocí klíčových slov

V analýze pomocí klíčových slov jde o hledání nejdůležitějších pojmů v dokumentu a jejich následnou klasifikaci. Podle klasifikace sentimentu klíčových slov je určen i sentiment dokumentu. [66]

4.3 Word embedding

Word embedding je jednou z populárních metod zpracování přirozeného jazyka. Přístup spočívá v převedení slov na vektory v mnohorozměrném prostoru tak, aby slova s podobným významem byla blízko sobě a slova nepodobná daleko od sebe. Další vlastností tohoto zpracovaní je "linguistic regularity" tím rozumíme, že část morfologických a sémantických vlastností se dá reprezentovat unikátním vektorem. Pokud bychom tedy k původnímu slovu přičetli zmíněný vektor, dostaneme se k blízkosti nového slova, které se ovšem od původního liší právě vlastností reprezentovanou vektorem. [39]

4.3.1 Word2Vec

Word to Vector (Word2Vec) používá neuronové sítě k vytvoření spojení mezi kontextem a slovem. K tomu používá buď Continuous Bag of words (CBOW) nebo skip-gram model. CBOW používá jako vstup kontext, který je reprezentován předchozími a následujícími slovy, počet vstupních slov záleží na délce definovaného okna. Následně je predikováno hledané slovo. Skip-gram se snaží obráceně predikovat kontext podle vstupního slova. Vektory jsou poté reprezentovány jako váhy v neuronové síti. [7, 38]

4.3.2 GloVe

Global Vectors for Word Represantation (GloVe) používá k reprezentaci kontextu matici, do které zaznamenává, jak často se slova vyskytují ve společném kontextovém okně. Tato matice dosahuje velkých rozměrů, a proto se používá její faktorizace k dosažení efektivnější reprezentace. [7, 42]

4.4 Dostupné nástroje

Dosud zde byly popsány pouze nástroje či postupy, které je potřeba implementovat. Ovšem existují i již hotové nástroje, které budou popsány nyní. Následně z nich budou vybrány a použity zvolené nástroje pro extrakci sentimentu.

4.4.1 CoreNLP

CoreNLP[1] je soubor nástrojů k jazykové analýze. Jedním z jeho možných využití je i extrakce sentimentu pomocí neuronové sítě. Nástroj pracuje tak, že věta je rozebrána do stromové struktury přes jednotlivá slovní spojení až k základním slovům, které jsou následně ohodnocena. Následně jsou vypočítána ohodnocení slovních spojení, až je nakonec ohodnocena celá věta.

Protože zjištění sentimentu není kvůli problémům popsaným výše jednoduchý úkol (4.1), využívá CoreNLP trénovaný model Sentiment Treebank. Tento model pro angličtinu obsahuje stromový model sentimentu pro 215 154 frází v 11 855 větách a dosahuje přesnosti až 85,4% při analýze jedné věty.

CoreNLP je vydáváno pod licencí GNU GPL v3 a pro svůj v běh vyžaduje Javu. Je s ním možné ovšem komunikovat přes příkazy či webové rozhraní a existuje také mnoho prostředků, jak využívat tento nástroj v jiných

[1] https://stanfordnlp.github.io/CoreNLP

programovacích jazycích. CoreNLP je možné používat k analýze angličtiny, arabštiny, čínštiny, francouzštiny, němčiny a španělštiny. [54]

4.4.2 Natural Language Toolkit

Natural Language Toolkit (NLTK)[2] je platforma pro práci s textem vyvíjená pro Python a obsahuje přes 50 zdrojů slov. Samotné NLTK je vyvíjeno pod licencí Apache 2.0 License, a je tedy volně k použití. Samotný NLTK obsahuje již třídu Sentiment Analyzer, s jejíž pomocí lze vytvořit klasifikátor pro zjištění sentimentu.

Ovšem obsahuje také modul VADER (Valence Aware Dictionary and Sentiment Reasoner), což je nástroj pro extrakci sentimentu založený na slovnících a pravidlech. Každé klíčové slovo tedy obsahuje hodnocení sentimentu a VADER z těchto hodnot určí konečný sentiment textu. Mimo pouhého hledání ve slovníku VADER také bere v potaz kontext, slova napsaná velkými písmeny (zdůraznění), vykřičníky, stupňování přídavných jmen či věty s ale. [31]

S NLTK se dají použít i další třídy, jednou z nich je například Sentiment Classifier[3]. Tato třída využívá slovník wordnet a statistiku výskytu slov z NLTK, aby mohla určit sentiment slov, která mají více významů. Text je klasifikován pouze jako pozitivní či negativní, a liší se tím od VADERu, který umí klasifikovat text i jako neutrální.

Další používanou knihovnou je TextBlob[4], který pracuje na základě pravidel. Jeho výhodou je velmi jednoduché rozhraní. TextBlob umí určovat polaritu textu (zda je text negativní či pozitivní) a také jeho objektivitu.

4.4.3 Ostatní nástroje

- *FastText*[5] - je open-source knihovna napsaná v C++ vyvíjená společností Facebook, s neoficiálními wrappery je použitelná i v jazyce Python
- *PyText*[6] - Deep learning framework postavený na PyTorch. Jeho zaměření je hlavně na jednoduchou implementaci zpracování textu

[2]https://www.nltk.org/
[3]https://pythonhosted.org/sentiment_classifier/
[4]https://textblob.readthedocs.io/en/dev/
[5]https://fasttext.cc
[6]https://github.com/facebookresearch/pytext

- *AllenNLP*[7] - Deep learning framework postavený na PyTorch zaměřený na vytváření prototypů
- *FLAIR*[8] - NLP knihovna postavená na PyTorch zabývající se rozpoznáváním pojmenovaných entit, označováním části řeči, rozlišováním smyslu a klasifikací

Často se zde vyskytuje knihovna PyTorch[9], jedná se o open-source knihovnu pro strojové učení, která se používá v Pythonu.

[7] https://allennlp.org/
[8] https://github.com/zalandoresearch/flair
[9] https://pytorch.org/

5 Tepové frekvence a její získání

V druhé kapitole bylo popsáno, že srdeční rytmus je periodické střídání stažení a uvolnění srdečního svalu (2.2). V této kapitole budou uvedeny metody měření tepové frekvence a tepová pásma.

5.1 Tepová frekvence

Abychom lépe pochopili tepovou frekvenci je nutné se seznámit se základními pojmy. Tím prvním je klidová tepová frekvence, která by měla odrážet zdraví a fyzický trénink člověka. Druhým pojmem je maximální tepová frekvence, které může člověk při zátěži dosáhnout bez ohrožení na zdraví.

5.1.1 Klidová tepová frekvence

Klidová tepová frekvence se měří, pokud je člověk vzhůru, v běžné teplotě a nebyl vystaven žádné námaze či stimulaci, jako je například stres nebo překvapení. Normální klidová tepová frekvence je 60 až 90 úderů za minutu. Nad 90 úderů za minutu roste možnost srdečních chorob. Tepová frekvence je také závislá na věku a s věkem klesá, kdy u dětí dosahuje až 120 úderů a ve stáří může klesnout i pod 60 úderů za minutu (tabulka 5.1) [26].

- *Tachykardie* - pokud je klidová frekvence vyšší než 100 úderů za minutu. Může se vyskytovat při emocionálních stavech jako je stres nebo úzkost a nebo v těhotenství. Ale také může být důsledkem sepse, horečky, kardiomyopatie a dalších onemocnění.
- *Bradykardie* - klidová frekvence je nižší než 60 úderů na minutu. Často se může vyskytovat u trénovaných sportovců, ale zde se nejde o příznak nemoci. U nesportovců může jít o důsledek infarktu myokardu či nitrolebeční poranění [55].

5.1.2 Maximální tepová frekvence

Maximální tepová frekvence se zjišťuje většinou pomocí srdečního zátěžového testu, kdy je subjektu měřeno EKG a zároveň je vystaven fyzické zátěži např.

Věk	Tepová frekvence v úderech za min.
Novorozenci (0 - 3 měsíce)	99-149
Kojenci (3 – 6 měsíce)	89–119
Kojenci (6 – 12 měsíce)	79-119
Děti (1 – 10 roků)	69–129
Děti přes 10 let, dospělí, senioři	59–99
Trénovaní atleti	39–59

Tabulka 5.1: Tepová frekvence a věk [25]

běhání na běžeckém pásu. Intenzita běhu se postupně zvyšuje, dokud nejsou detekovány změny srdeční funkce. Test trvá obvykle od 10 do 20 minut [16].

Hodnotu maximální tepové frekvence (5.1) můžeme určit také pomocí výpočtu. Nejčastěji používaný vzorec formulovaný v roce 1970 doktory Williamem Haskellem a Samuelem Foxem (5.1) [14, 47].

$$HR_{max} = 220 - věk \tag{5.1}$$

Tento vzorec byl vytvořen na základě 11 různých vědeckých prací. Vzorec je jednoduchý na zapamatování a velmi se rozšířil díky použití v přístrojích na sledování srdeční tepové frekvence. Je nutné zmínit, že tento vzorec ovšem není úplně přesný a chyba se může pohybovat v 7 až 11 úderech za minutu. Podle srdeční frekvence rozlišujeme následující tepová pásma [12]:

- *Nízká aktivita* - 50 až 59 % - rozcvička, zóna pro přípravu před výkonem
- *Fitness* - 60 - 69 % - fitness, zvýšené spalování tuků
- *Aerobní* - 70 - 79 % - vytrvalostní trénink, zlepšení stavu kardiovaskulárního a respiračního systému
- *Anaerobní* - 80 - 89 % - zátěžový trénink, zvyšuje se objem kyslíku, který je organismus schopen přijmout
- *VO2 Max* - 90 - 100 % - maximální zátěž, tělo je schopno spálit nejvyšší počet kalorií, zvyšuje se riziko zranění a prodlužuje doba na regeneraci

Často se také můžeme setkat s tabulkou (obrázek 5.1), kde je porovnání různých zátěžových zón podle věku a tepové frekvence.

Zátěžové zóny										
		Věk								
		20	25	30	35	40	45	50	55	65	70
Tepová frekvence	100%	200	195	190	185	180	175	170	165	155	150
		VO2 Max - maximální zátěž									
	90%	180	176	171	167	162	158	153	149	140	135
		Anaerobní - zátěžový									
	80%	160	156	152	148	144	140	136	132	124	120
		Aerobní - vytrvalostní									
	70%	140	137	133	130	126	123	119	116	109	105
		Fitness - spalování tuků									
	60%	120	117	114	111	108	105	102	99	93	90
		Nízká aktivita - rozcvička									
	50%	100	98	95	93	90	88	85	83	78	75

Obrázek 5.1: Zátěžové zóny [12]

5.2 Měření tepové frekvence

V této části budou popsány hlavní metody pro měření tepové frekvence, včetně fotopletysmografu, který používá Fitbit Charger HR[1] (obrázek 5.2) a byl využit v této práci.

5.2.1 EKG

Elektrokardiografie neboli EKG slouží pro snímání změny elektrického potenciálu, který je způsoben srdeční aktivitou. Časový záznam je uložen v podobě elektrokardiografu pro vyhodnocení. EKG je neinvazivní metoda a pro sběr dat používá elektrody připojené buď na kůži, nebo například na stěnu jícnu. Elektrická aktivita srdce vybuzuje mechanickou aktivitu a pomocí EKG lze tedy zjišťovat i srdeční choroby.

Ke snímání EKG dochází například pomocí dvanáctisvodového systému, kdy se elektrody umisťují na pravé a levé zápěstí, levou nohu a hrudník. Z tohoto plyne, že tato metoda je vhodná pro klinické vyšetření a nemůže být použita v této práci.

[1] https://www.fitbit.com/be/chargehr

Obrázek 5.2: Fitbit charger HR

Z elektrokardiografu se dá následovně vyčíst tepová frekvence. Ta se určuje jako vzdálenost mezi depolarizacemi srdečních komor tj. R-R interval. [17]

5.2.2 Hrudní pás

Hrudní pás bývá součástí příslušenství u sporttesterů a skládá se z monitorovací a vysílací části. Monitorování tepové frekvence hrudním pásem se podobá snímání EKG a opět se jedná o neinvazivní metodu. Hrudní pás obsahuje dvě elektrody, které snímají změny elektrického potenciálu srdce. Zde se také počítá s tím, že sledovaný objekt se potí, a tím se zvyšuje vodivost spojení mezi kůží a elektrodami. Snímaná hodnota se zpracovává v monitorovací části, kde je signál vyčištěn od nežádoucích artefaktů a šumu a odečtena tepová frekvence. Následně je výsledek pomocí vysílací části odeslán například uživateli do jeho hodinek či mobilního telefonu [34].

Mezi výhody snímání tepové frekvence hrudním pásem patří přesnost, ale jeho nošení může být nepříjemné, a není tak vhodný pro sledování v delším časovém úseku. Další nevýhodou je nutnost mít u sebe zařízení pro příjem dat.

5.2.3 Pletysmografie

Další neinvazivní metodou je pletysmografie, která pro zjištění tepové frekvence používá objemové změny částí těla. Změna je způsobena pulzy srdce,

které následně vyvolává tlakové a objemové změny v krevním řečišti. Objemové změny se následně dají měřit různými metodami [3].

- *mechanický pletysmograf* - zaznamenává změny pomocí pneumatického nebo vodního systému. Tento přístroj má tvar náprstku, ve kterém je uzavřen článek prstu. Prostor náprstku je vyplněn vzduchem nebo vodou a přenáší objemové a tlakové změny k membráně, která je následně snímá.

- *impedanční pletysmograf* - ke sledování používá změny v elektrické impedanci, které jsou závislé na změnách objemu krve. Pro tento účel je využito průchodu malého střídavého proudu sledovanou částí těla. Při průchodu pulzu dojde ke zvýšení elektrické vodivosti. Při použití je nutné mít na paměti, že kůže nemá všude stejný odpor, a tím může docházet k chybám při měření.

- *elektrokapacitní pletysmograf* - tato metoda se liší zejména v tom, že nedochází ke kontaktu s tělem. Při měření je první elektroda vzdálena 1 až 2 mm od těla a druhou elektrodu představuje sledovaný orgán. Při průchodu pulzu dochází ke změnám vzdálenosti mezi elektrodami, která má za následek změnu elektrické kapacity vzdušného prostoru. Tato metoda má nevýhodu v tom, že pozorovaný objekt musí být v absolutním klidu.

- *fotoelektrický pletysmograf* - metoda využívající pro měření světlo. Zdroj světla svítí na tkáň a následně je snímáno množství odraženého světla. Množství zachyceného světla je závislé na objemu krve v tkáni. Podle toho, jak objem roste a klesá, je možné určit tepovou frekvenci. Kvalita výsledného měření závisí i na použité vlnové délce světelného zdroje. Nejčastěji se používá zelené či červené světlo a infračervené záření. Větší vlnová délka zajišťuje, že světlo proniká hlouběji do tkáně, ale výsledek může být následně ovlivněn pohyby v tkáni.

  Fotoelektrické pletysmografy dělíme na dva typy, a to reflexní a průsvitové. Reflexní pletysmograf obsahuje zdroj světla i snímač vedle sebe. Může být umístěn například na hodinkách. Průsvitový pletysmograf snímá průchod světla celou částí těla, a snímač je tedy na opačné straně než zdroj. Tato metoda je přesnější, ale není tolik vhodná pro nositelnou elektroniku.

5.2.4 Nepřesnosti měření

Jelikož v této práci je využito fotoelektrické pletysmografie, je nutné také zmínit, jak při tomto druhu sledování tepové frekvence může dojít k nepřesným měřením [10].

- *Špatné utažení senzoru* - při malém utažení senzoru dochází k pohybu měřícího přístroje, které může vést ke změnám ve výsledcích vlivem vnikání okolního světla na senzor. Přístroj by měl být pevně utažen, ale ne příliš, jinak dojde k zamezení cirkulace krve v končetině, a tím je taktéž ovlivněno měření.

- *Pohyb sledované osoby* - pohyb ovlivňuje tkáň, kde se tepová frekvence měří. Při pohybu se v tkáni mění objem krve, který znesnadňuje rozpoznání pulzů srdce. Správné umístění senzoru nad zápěstím by mělo tento problém eliminovat.

- *Nervózní chování* - pokud je osoba nervózní a například provádí stále stejnou činnost (např. klepání prstem či podupávání), způsobuje toto chování záchvěvy, které ovlivňují měření. Opět pomůže správné utažení a umístění senzoru, ale je také dobré nositele na tuto skutečnost upozornit, aby se vyhýbal činnosti ovlivňující měření.

- *Chyba metody* - měření pomocí fotoelektrického pletysmografu není zcela přesné a je vhodné jako orientační, a to kvůli měření pouze dynamické změny v krevním řečišti. Další věc, která může ovlivnit správnost měření, jsou vlnové délka světelného zdroje a kvalita snímače.

6 Příprava dat

Nyní již máme popsané metody, jak získat dva typy dat, a to data se sentimentem v kapitole (4) a data s tepovou frekvencí (5). Tato kapitola se bude zabývat tím, jak tato data zkombinovat a následně si vytvořit ucelený obrázek nad změřenou tepovou frekvencí a sentimentem extrahovaným z textu.

Určený sentiment je zde nezávislá proměnná, jeho hodnota je dána vyjádřením účastníka experimentu. Závislou proměnnou je tepová frekvence, u které bude hledána závislost na sentimentu příspěvku.

6.1 Diskrétní a spojité veličiny

Měřené hodnoty mohou být vyjádřeny dvěma způsoby. Pokud máme měření, které obsahuje spočitatelné množství hodnot po nespojitých krocích (izolované body v čase), mluvíme o diskrétní veličině. Hodnoty těchto měřeních mohou být použity například pro zjištění počtu či četnosti.

Druhým typem je spojitá veličina, kde hodnota nabývá jakékoli velikosti v určitém intervalu. Takto se dá měřit většina veličin, například váha. Za spojité veličiny jsou považovány i měření tepové frekvence či rychlosti dýchání a to kvůli tomu, že je zde velké nespočitatelné množství výsledků v čase.

Pokud se podíváme na měření v této práci, tak měření tepové frekvence bude bráno jako spojitá veličina, ale zjištění sentimentu je diskrétní. Proto bude nutné najít způsob, jak oba typy dat spojit dohromady.

6.2 Příprava dat

Z jednotlivých experimentů nejsou data ihned vhodná k použití. Je nutné data předzpracovat, aby byla vhodná pro strojové učení a odstranily se chyby v datech. Jako příklad mohou sloužit překlepy v psaní příspěvků na sociální síť Twitter.

6.2.1 Data pro extrakci sentimentu

Data pro sentiment byla sbírána během pilotního (3.2.1) a kvaziexperimentu (3.2.2). Účastníci kvaziexperimentu psali příspěvky v českém jazyce a kromě

opravy gramatiky bylo nutné i příspěvky přeložit do angličtiny, protože většina nástrojů pro extrakci sentimentu funguje nejlépe s angličtinou.

Před použitím dat bylo nutné udělat tyto kroky:

1. Kontrola gramatiky česky psaných příspěvků, tato kontrola byla prováděna ručně s využitím kontroly pravopisu v Google Docs
2. Překlad pomocí Google Translatoru do angličtiny
3. Ruční kontrola, zda příspěvek psaný česky odpovídá přeloženému do angličtiny
4. Kontrola anglického pravopisu pomocí automatické kontroly v Microsoft Excel
5. Extrakce sentimentu pomocí strojového učení, neuronových sítí či hotových nástrojů (4.4)

Využití automatizovaných nástrojů pro překlad a kontrolu gramatiky bylo nutné vzhledem k počtu příspěvků (pilotní experiment přes 2000 a kvaziexperiment přes 1000). Aby se snížil počet chybných překladů, překlepů či gramatických chyb, byla provedena kontrola každého příspěvku ručně.

Využití hotových nástrojů se jeví jako vhodná varianta. Nástroje se chlubí vysokou přesností a také v této práci jde hlavně o nalezení možné souvislosti mezi sentimentem a tepovou frekvencí.

6.2.2 Data z měření srdečního tepu

Stejně jako data pro extrakci sentimentu byla měřena tepová frekvence během pilotního a kvaziexperimentu. Data každého účastníka jsou různá, tj. různá klidová tepová frekvence (5.1.1) i maximální tepová frekvence (5.1.2). Aby se dala data porovnávat mezi jednotlivými subjekty, je nutné naměřená data normalizovat.

6.3 Reprezentace a příprava dat sentimentu na fúzi

Na začátku této kapitoly bylo řečeno, že existují diskrétní a spojité veličiny (6.1) a že extrahovaný sentiment je reprezentován jako diskrétní veličina a naměřená tepová frekvence jako spojitá. Z toho vyplývá, že je nutné převést jedno či druhé měření na druhý typ veličiny, aby byla možná jejich fúze.

6.3.1 Reprezentace dat

Sentiment extrahovaný ze zjištěných dat může nabývat hodnot podle podle použitého nástroje. Data mohou být vyjádřena například pouze jako negativní (-1 nebo N) či jako pozitivní (1 nebo P) a občas se může přidat i kategorie neutrálního sentimentu (0 nebo X). Dále se můžeme setkat s více kategoriemi, kde se přidávají vyjádření o jak moc pozitivní či negativní příspěvek se jedná, a to buď slovně (velmi negativní, negativní, neutrální, pozitivní, velmi pozitivní), či číselně (0, 1, 2, 3, 4, 5). Některé nástroje extrahují sentiment jako spojitou veličinu mezi hodnotami -1 a 1 [35].

6.3.2 Interpolace sentimentu

Abychom byli schopni spojit hodnoty sentimentu, které jsou diskrétní a mají tak platnost pouze v jeden časový okamžik, a hodnot měření tepové frekvence, je nutné data pro sentiment interpolovat. První možností je prodloužení hodnoty sentimentu až do další změny například metodou Zero-order Hold (změna bude skoková) a nebo provést trojúhelníkové prodloužení pomocí First-order Hold (změna bude postupná).

6.3.3 Zero-order Hold

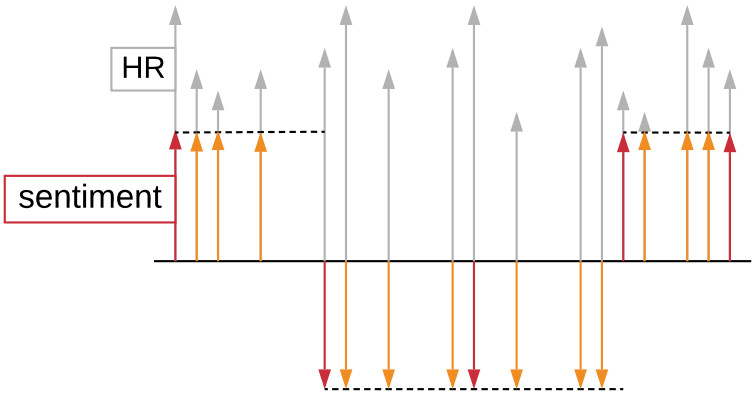
Zero-order Hold (ZOH) je matematický model pro rekonstrukci signálu. Jeho použití tkví v převodu diskrétního signálu na spojitý tak, že udržuje hodnotu signálu do té doby, než je nalezena hodnota další. Jedná se tedy o prodloužení hodnoty v čase. Vzorec pro ZOH vypadá následovně:

$$x_{ZOH} = \sum_{n=-\infty}^{\infty} x(n) rect(\frac{t - \frac{T}{2} - nT}{T}) \tag{6.1}$$

Kde platí že t je okamžitý čas, T je velikost intervalu a rect je obdélníková funkce:

$$rect(x) = \Pi(x) = \begin{cases} 0, & \text{if } |x| > \frac{1}{2} \\ \frac{1}{2}, & \text{if } |x| = \frac{1}{2} \\ 1, & \text{if } |x| < \frac{1}{2} \end{cases} \tag{6.2}$$

Pokud použijeme tedy ZOH na hodnoty sentimentu, bude přiřazená hodnota sentimentu k tepové frekvenci pořád stejná, dokud nedojde ke změně sentimentu. Tato změna může nastat pouze načtením tweetu s jinou hodnotou sentimentu. Spojení s tepovou frekvencí je vidět na obrázku (obrázek 6.1):



Obrázek 6.1: Zero-order hold s tepovou frekvencí

6.3.4 First-order Hold

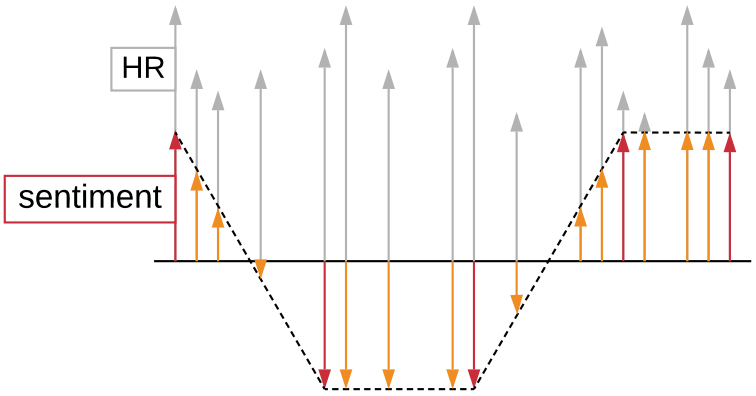
Další možností je použití metody First-order Hold (FOH), která diskrétní signál mění na spojitý po částech jako lineární aproximaci vzhledem k původnímu signálu. Tato metoda se od Zero-order Hold liší v lepším napodobení původní křivky, takže model lépe odpovídá původnímu signálu. Interpolace mezi dvěma hodnotami se řídí následujícím vzorcem:

$$x_{FOH} = \sum_{n=-\infty}^{\infty} x(nT) tri(\frac{t - nT}{T}) \tag{6.3}$$

Kde platí že t je okamžitý čas, T je velikost intervalu a tri je trojúhelníkové funkce:

$$tri(x) = \Lambda(x) = \begin{cases} 1 - |x|, & \text{if } |x| < 0 \\ 0, & \text{jinak} \end{cases} \tag{6.4}$$

Při použití FOH dojde k tomu, že hodnoty mezi jednotlivými měřeními sentimentu jsou následně proloženy přímkou (lineární interpolace) a vzniká trojúhelníkový tvar výsledného signálu. Hodnota sentimentu nyní již ovšem nebude binární, ale bude se měnit vzhledem k interpolaci hodnot sentimentu. Spojení s tepovou frekvencí můžeme vidět na dalším obrázku (obrázek 6.2):



Obrázek 6.2: First-order hold s tepovou frekvencí

6.4 Reprezentace a příprava dat tepové frekvence na fúzi

Data pro tepovou frekvenci jsou vyjádřena v počtu úderů za minutu. Tato data jsou nasbírána jako časová řada a naměřené hodnoty jsou pro každý subjekt individuální. Aby byla data lépe zpracovatelná, je nutné, aby různá měření měla stejný rozsah. Pro změnu měřítka je možné využít metod pro normalizaci.

6.4.1 Normalizace

Normalizace znamená převedení dat z původního rozsahu na nový rozsah tak, že nový rozsah je mezi hodnotami 0 a 1 nebo -1 a 1.

Lineární min-max normalizace

Nejjednodušší formou normalizace je min-max normalizace, kdy se hodnoty lineárně transformují do nového oboru hodnot a zůstává zachována jejich distribuce. Pro použití této metody je nutné znát maximum a minimum z použitých hodnot. Nevýhodou použití min-max normalizace je, že může dojít ke zpracování dat, které mají maximum a minimum výrazně odlišné od střední hodnoty. Důsledkem toho zůstává většina rozsahu nevyužita a pokud se jedná o anomálie, může dojít i k výraznému zkreslení vstupní informace. Následující vzorec vyjadřuje použití min-max normalizace:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{6.5}$$

Hodnoty x jsou původními hodnotami a x' je normalizovaná hodnota.

Nelineární soft-max normalizace

Tato normalizace se provádí pomocí logaritmické funkce. Data, která se nalézají v blízkosti standardní odchylky od průměru, jsou normalizována téměř lineárně, zatímco data, která se nalézají dále, se normalizují nelineárně. Tento přístup nám umožňuje snížit dopad minimálních a maximálních hodnot z min-max normalizace. Další výhodou je, že nemusíme znát minimální a maximální hodnotu dat. Soft-max normalizaci lze provést pomocí následujícího vzorce:

$$x' = \frac{1}{1 - e^{\frac{x-\mu}{\sigma}}} \tag{6.6}$$

Hodnota x' je normalizovaná hodnota, x původní hodnota, μ je aritmetický průměr$a\sigma$ směrodatná odchylka.

Z-score normalizace

Další možnou metodou pro normalizaci je použití z-score normalizace. Průměrná hodnota při použití této metody je rovna 0 a směrodatná odchylka odpovídá hodnotě 1. Pomocí z-score se vyjadřuje vzdálenost mezi normalizovanou hodnotou a střední hodnotou množiny. Pokud je známa hodnota průměru a směrodatné odchylky, můžeme z-score vypočítat následovně:

$$x' = \frac{x - \mu}{\sigma} \tag{6.7}$$

kde x je původní hodnota, x' normalizovaná, μ je aritmetický průměr a σ směrodatná odchylka.

7 Strojové učení

Strojové učení je metoda pro použití algoritmů umožňujících učení umělých objektů, přičemž učením se rozumí zlepšování výsledků na základě zkušeností. Metody strojového učení si vytvářejí matematický model například pomocí trénovacích dat a následně se podle něj rozhodují či vytvářejí predikce. Při učení dochází k změnám v modelu, které by měly vést ke zlepšení a přizpůsobení modelu okolnímu prostředí. [4]

V této kapitole jsou rozebrány základní metody strojového učení včetně neuronových sítí. Strojové učení bude využito ke zjištění vazby mezi tepovou frekvencí a sentimentem extrahovaným z textu. K tomuto účelu budou využita data, která vzniknou fúzí dat (6).

7.1 Rozdělení učících se algoritmů

Jak již bylo zmíněno, tak metody strojového učení vytvářejí matematický model. Máme více možností jak tento model tvořit a jak přistupovat k trénovacím datům.

7.1.1 Učení s učitelem

Při strojovém učení s učitelem obsahují trénovací data nejen příznaky, ale také požadovaný výstup. Vytvořený matematický model se iterativně optimalizuje tak, aby dokázal klasifikovat nebo použít regresi na neznámá data. Možné je také využití trénovacích dat, kde části dat chybí požadovaný výstup. Taková metoda se nazývá jako kombinace učení s učitelem a bez učitele. [48]

7.1.2 Učení bez učitele

Učení bez učitele nepoužívá trénovací data. Data tedy musejí splňovat podmínku, že nebyla klasifikována, kategorizována nebo označena. Tyto metody nacházejí spojitosti v jednotlivých datech a musí klasifikovat data bez jakékoli vnější pomoci. [48]

7.1.3 Zpětnovazebné učení

Učení pomocí zpětné vazby má za úkol minimalizovat chybovost vyhodnocování. Modelu tedy není řečeno, jak by měl činnost vykonávat, ale místo toho se snaží zjistit, který krok produkuje co nejlepší výsledek. [57]

7.2 Typy úloh

7.2.1 Klasifikace

Klasifikace je definována jako zařazení nového objektu do určené sady kategorií. Určování probíhá na základně trénovacích dat, která obsahují požadovaný výstup. Jedná se tedy o učení s učitelem. Trénovací data jsou analyzována pomocí nezávislých příznaků (features), které jsou následně rozpoznávány u neznámých dat a dochází ke klasifikaci do určité kategorie. [1]

7.2.2 Regrese

Regrese je také úloha spadající do učení s učitelem. Od klasifikace se liší tak, že neklasifikuje do kategorií, ale podle vzniklého modelu vytváří predikci pro novou číselnou hodnotu. [1]

7.2.3 Shlukování

Tato úloha se snaží o shlukování objektů do skupin (clusterů) tak, aby si objekty z jedné skupiny byly podobnější než objekty z ostatních skupin. Data musejí být popsána stejnou množinou příznaků a nenesou informaci o příslušnosti k požadované skupině. Shlukování spadá do kategorie učení bez učitele. [32]

7.3 Modely strojového učení

7.3.1 Support Vector Machine

Support vector machine (SVM) [9] je model strojového učení s učitelem určený pro klasifikaci a regresi. Učební algoritmus pro SVM je nepravděpodobnostní, binární, lineární klasifikátor, který se snaží nalézt nadrovinu rozdělující trénovací data do dvou opačných poloprostorů. Hledá se taková nadrovina, která je od trénovacích dat co nejvzdálenější. Pro popis této nadroviny se používají nejbližší body nazvané podpůrné vektory (support

vectors) a je lineární funkcí v prostoru příznaků. Pokud máme množinu trénovacích dat ($\vec{x_i}$, y_i), kde $\vec{x_i}$ je vektor a y_i je informace od učitele nabývající hodnot -1 až 1, poté můžeme zapsat rovnici nadroviny následovně:

$$\vec{w} * \vec{x} + b = 0 \tag{7.1}$$

kde $\vec{w}$ je normála nadroviny.

7.3.2 Naive bayes

Metoda bayesovské klasifikace vychází z Bayesovy věty o podmíněných pravděpodobnostech. Bayesovský klasifikátor zařazuje objekt do nejpravděpodobnější třídy pomocí vektoru příznaků. Bayesův vztah pro výpočet podmíněné pravděpodobnosti vychází z předpokladu, že platí hypotéza Y, pokud pozorujeme výskyt X.

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \tag{7.2}$$

Naive bayesovský klasifikátor [36] vychází z Bayesova teorému. Stejně jako u SVM se jedná o metodu učení s učitelem. Naive bayes využívá následující vzorec:

$$P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \tag{7.3}$$

- *P(h)* - apriorní pravděpodobnost jevu h
- *P(D)* - pravděpodobnost výskytu dat D bez znalosti pravděpodobnosti jevu
- *P(h/D)* - aposteriorní pravděpodobnost, pravděpodobnost jevu h, pokud jsou přítomna data D
- *P(D/h)* - pravděpodobnost výskytu dat D, jestliže nastal jev h

Klasifikátor funguje tak, že hledá hypotézu s nejvyšší aposteriorní pravděpodobností.

7.3.3 Rozhodovací strom (Decision tree)

Rozhodovací strom [1] je metoda primárně založená na klasifikaci podle vstupních dat. V nelineární hierarchické struktuře se vyskytují uzly, hrany a listy. Uzly využívají hodnoty zkoumaných atributů a rozdělují model na jednotlivé podstromy, přičemž atribut pomocí své hodnoty popisuje sledovaný

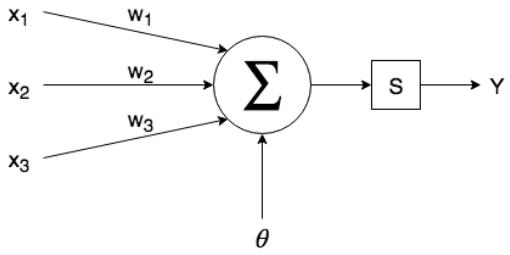
objekt. Uzly tedy připomínají otázky, podle kterých se model rozhoduje. Následují hrany, které představují odpovědi na otázky, a podle hodnoty atributu může model zvolit, jakým směrem se vydá. Poslední částí jsou listy, kde se nacházejí rozhodnutí.

Jedním z možných algoritmů pro tvorbu rozhodovací stromu je ID3 (Iterative Dichotomiser 3). Tento algoritmus vytváří rozhodovací strom podle počtu hodnot dělících atributů, přičemž vybírá ty atributy, které zvyšují informační zisk stromu snižováním entropie. Entropie nám udává míru neuspořádanosti stromu v intervalu 0 až 1 a cílem je její minimalizace, a tím dosažení přesnějších výsledků. Algoritmus pracuje v těchto krocích:

1. *Vytvoření rodičovského uzlu* - uzel obsahuje všechny záznamy
2. *Testování ukončovacích podmínek* - pokud uzel splňuje ukončovací podmínky, je tedy listem a algoritmus končí
3. *Výběr vhodného atributu* - výběr atributu s největším informačním přínosem, který nejlépe rozděluje jednotlivé záznamy
4. *Rozdělení záznamů do nových uzlů* - záznamy jsou rozděleny do nových uzlů, které se stávají rodiči, a algoritmus se opakuje

7.3.4 Neuronové sítě

Neuronové sítě vznikly jako metoda pro paralelní zpracování dat a jejím vzorem je chování v biologických strukturách. Neuronová síť je tvořena orientovaným grafem, který se skládá z neuronů (vrcholů) a spoji mezi neurony - synapsemi (hrany). Neurony mají libovolný počet vstupů, ale pouze jeden výstup a navzájem si předávají signály, které jsou transformovány pomocí aktivačních funkcí. Umělý neuron si můžeme ukázat na modelu, který popsali McCulloch a Pitts [37] (obrázek 7.1):



Obrázek 7.1: Perceptron

$$Y = S(\sum_{i=1}^{N}(w_i x_i) + \Theta) \tag{7.4}$$

- x_i - vstupy neuronu
- w_i - váhy synapsí
- Θ - práh
- S - aktivační funkce neuronu
- Y - výstup neuronu

Hodnota Θ nám určuje prahovou hodnotu aktivace; pokud není této hodnoty dosaženo, neuron zůstává neaktivní.

Aktivační funkcí používaných v neuronových sítí je mnoho, mezi základní patří:

- *Funkce jednotkového skoku*
- *Funkce signum*
- *Lineární funkce*
- *Hyperbolický tangent*
- *Sigmoidální funkce*

Nejjednodušším modelem neuronové sítě je perceptron, který patří do dopředných sítí. Skládá se pouze z jednoho neuronu a umožňuje nám klasifikovat množiny, které jsou lineárně oddělitelné.

Mimo dopředných neuronových sítí, které mají výstup jedné vrstvy připojen na vstup druhé vrstvy, rozeznáváme ještě sítě se zpětnou vazbou (rekurentní sítě), kde se signál z výstupu sítě vrací na její vstup. [63]

7.3.5 Genetické algoritmy

Genetické algoritmy se snaží napodobovat biologickou evoluci tak, jak jí popsal Charles Darwin. Každý jedinec je tvořen atributy (geny) a z hlediska vývoje jsou nejcennější ty atributy, které zvyšují šanci jedince na přežití. Míra adaptace v prostředí se vyjadřuje kvantitativní číselnou hodnotou (hodnotící či fitness funkce). Šance na přežití mají ti nejlepší jedinci a ti, kteří měli štěstí. Tito jedinci se stává rodiči další generace a dochází ke křížení genů rodičů a náhodným chybám (mutacím). Tento cyklus se neustále opakuje. [21]

1. *Vytvoření první generace* - první generace je vytvořena náhodným generováním
2. *Vyhodnocení první generace* - zjištění adaptace jedinců
3. *Selekce* - výběr přeživších jedinců
4. *Křížení a mutace* - vytvoření párů z přeživších a vytvoření nové generace
5. *Vyhodnocení nové generace* - zjištění adaptace nových jedinců
6. *Ukončovací podmínka* - ukončovací podmínkou může být například kvalita jedinců či počet vytvořených generací, pokud není splněna, vrací se algoritmus na bod 3

7.4 Měření přesnosti

Aby bylo možné určit přesnost jednotlivých metod pro strojové učení, je nutné použít určité metriky. Nejprve je nutné zjistit, zda byl objekt zařazen podle našeho očekávání či nikoliv, celkově mohou nastat 4 případy:

- *správně pozitivní (TP)* - kolik výsledků bylo správně pozitivních
- *falešně pozitivní (FP)* - kolik výsledků bylo chybně klasifikováno jako pozitivní
- *falešně negativní (FN)* - kolik výsledků bylo chybně klasifikováno jako negativní
- *správně negativní (TN)* - kolik výsledků bylo správně negativních

Následně je možné tyto případy vyjádřit v tabulce známé jako matice záměn (confusion matrix) (tabulka 7.1) [56]. Tato matice je určená pro vizualizaci efektivity algoritmu.

		Skutečnost	
		Pozitivní	**Negativní**
Předpověď	**Pozitivní**	TP	FP
	Negativní	FN	TN

Tabulka 7.1: Matice záměn

Pokud známe tyto hodnoty, můžeme je využít k výpočtu jednotlivých metrik.

7.4.1 Přesnost

Přesnost (accuracy) je určena podílem pravdivých výsledků vůči všem výsledkům a určuje jak blízko je výsledek měření k pravé hodnotě veličiny.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.5}$$

7.4.2 Preciznost

Preciznost (precision) se určuje jako podmíněná pravděpodobnost, že daný objekt byl zařazen do správné skupiny. Hodnota preciznosti určuje, jak moc se dá věřit klasifikaci jako pozitivní.

$$precision = \frac{TP}{TP + FP} \tag{7.6}$$

7.4.3 Úplnost

Úplnost (recall) je určena jako podíl počtu správné pozitivně identifikovaných objektů a podílu všech relevantních objektů. Hodnota úplnosti nám říká, jaký podíl všech pozitivních objektů odhalíme.

$$recall = \frac{TP}{TP + FN} \tag{7.7}$$

7.4.4 F-míra

Hodnoty přesnosti a úplnosti lze zkombinovat do harmonického průměru nazvaného F-míra (F1 score).

$$F1\ score = \frac{2 * TP}{2 * TP + FP + FN} \tag{7.8}$$

7.4.5 Křížová validace

Křížová validace je model, pomocí kterého lze určit, jak se modely strojového učení chovají na nových a nezávislých datech. Jedním z možných modelů může být například k-folding [18]. Vstupní množina dat se rozdělí podle velikosti k na jednotlivé podmnožiny a jedna z těchto podmnožin je brána jako testovací data a zbytek jako data trénovací. Po natrénování a otestování modelu se natrénuje nový model s následujícími trénovacími a testovacími daty.

7.5 Problémy strojového učení

7.5.1 Přeurčení

Přeurčení je problém strojového učení, který nastává, pokud je model až příliš dobře natrénován trénovacími daty. Model se totiž natrénuje i podle nejméně významných trénovacích dat jako jsou šum či nepřesná měření a následně ztrácí schopnost generalizace. Tímto dochází následně ke špatným výsledkům na reálných datech. Přeurčení se dá omezit, pokud pro lineární data použijeme lineární model či použitím parametru maximální hloubky u rozhodovacích stromů. [6]

7.5.2 Podurčení

Druhým problémem je podurčení, které ovšem nastává, pokud máme trénovacích dat málo. Pokud model nemá dostatek dat k učení, dochází k přílišné generalizaci, což vede ke špatným výsledkům. Bohužel podurčení se dá odstranit pouze použitím větší sady trénovacích dat. [6]

7.6 Ensembling

Ensembling [40] byl zmíněn v jednom z článků, který byl inspirací pro tuto práci a vedl ke zlepšení výsledků [52]. Ensembling je proces, při kterém je využívána kombinace modelů pro dosažení požadovaného výstupu. Předpoklad je takový, že kombinací různých modelů dojde k omezení odchylek a tendencí. Tendence nám značí, jak moc se liší predikované hodnoty od měřené a odchylka říká, jak moc jsou od sebe jednotlivé predikce vzdálené. Výsledná chyba může být zapsána takto:

$$E = tendence^2 + odchylka + neodstranitelná\ chyba \tag{7.9}$$

K snížení chyby můžeme využít následující metody [59]:

- *Bagging* - snižuje odchylku přes použití podobných modelů a následné zprůměrování jejich výsledků [5]
- *Boosting* - boosting sekvenčně kombinuje slabé prediktory (přesnost stačí lehce nad 50%) k dosažení lepších výsledků. Funguje na principu vah, kdy každému vzorku je přidělena váha, a váha stoupá, pokud byl klasifikován špatně. Výsledky jsou poté produkovány váženou většinou (klasifikace) nebo váženým součtem (regrese) [15]

- *Stacking* - stacking používá více vstupních modelů a následně výstupní model. Výstupní model se učí podle výstupů ze vstupních modelů. Tato metoda může snižovat jak odchylku, tak tendence [64]

7.6.1 Random Forest

Random Forest [60] je jednou z metod využívající boosting. Pomocí jednotlivých náhodných rozhodovacích stromů jsou vytvářeny dílčí výsledky, které jsou následně zprůměrovány, a tím je vytvořen konečný výsledek. Podobně jako u rozhodovacího stromu 7.3.3 se model dělí pomocí uzlů, ale změna tkví v tom, že množina atributů je náhodně daná a opět je vybírán nejlepší možný kandidát. Samotné stromy tedy nevrací nejlepší možný výsledek, ovšem jejich zprůměrováním lze vytvořit spolehlivý klasifikátor. Výhodou této metody je i klasifikace do více než dvou tříd a možnost paralelizace výpočtu. Nevýhodou může nastat u úloh s velkým množstvím atributů, kde je jich málo relevantních. V tomto případě nemusejí být vybrány správné atributy.

7.6.2 Adaboost

AdaBoost [15] je metoda využívající boosting pro lineární kombinaci slabých klasifikátorů v jeden silný binární klasifikátor H(x). Adaboost v každém kroku svého učení přidává do silného klasifikátoru jeden slabý tak, aby minimalizoval odhad horní chyby klasifikátoru. V metodě jsou přiřazeny trénovací množině váhy D_t, které jsou inicializovány rovnoměrně. Při každé iteraci je váha špatně zařazeným vzorkům zvýšena a správně zařazeným snížena, a tím je i ovlivněn výběr dalšího klasifikátoru. Díky tomuto přístupem je AdaBoost schopen se přizpůsobit i těžko klasifikovatelným datům. Algoritmus pro AdaBoost ve zjednodušené formě vypadá následovně:

1. Nalezení nejlepšího možného klasifikátoru pro dané váhy D_t
2. Chyba klasifikátoru musí být menší než 0,5
3. Výpočet koeficientu slabého klasifikátoru v lineární kombinaci H(x)
4. Aktualizace vah D_t

8 Analýza, návrh a implementace

V této kapitole je popsán vytvořený program od analýzy, teoretického návrhu až po implementaci. Také je zde popsán výběr jednotlivých metod a důvody jejich volby.

V první části je popsána analýza vstupního text (8.1) a tepovou frekvenci (8.2). Je nutné zjistit, jak se vstupním textem zacházet, abychom dostali relevantní data pro extrakci sentimentu (8.1.1). Také musí být vybrán vhodný nástroj pro extrakci sentimentu (8.1.3) a určeno, jak zacházet s neutrálním sentimentem (8.1.4). U dat pocházející z měření tepové frekvence se musí analyzovat doba měření a případně ořezat data. Také je nutné zvolit metodu pro normalizaci (8.2.1).

V druhé části (8.3) je navrženo jak data se sentimentem a tepovou frekvencí spojit (8.3.1), jak spojená data vybalancovat (8.3.3) a rozdělit data na trénovací a testovací pro strojové učení (8.3.4). Následně zde nalezneme výběr metod strojového učení (8.3.5) a zaznamenávané metriky (8.3.7). V neposlední řadě je také navrženo přidání dalších vstupních dat (8.3.8).

V poslední části je popsána struktura projektu (8.4.1), jak zacházet se vstupními daty (8.4.2), provádět extrakci sentimentu (8.4.4) a jak programově spojit data (8.4.5). Také zde nalezneme popis použití metod strojového učení (8.4.6) včetně křížové validace a ukládání výsledků.

8.1 Analýza textu a extrakce sentimentu

Nejprve se zaměříme na analýzu sentimentu. Jako vstupní data pro sentiment byly zvoleny tweety z kvaziexperimentu (3.2.2), kterých bylo 1012 od sedmi subjektů experimentu. V kvaziexperimentu se počítalo s binárním vyjádřením pocitů (uživatelé označovali své tweety buď jako pozitivní, nebo negativní pomocí #p či #n), proto se nadále také bude počítat s binárním rozdělením sentimentu.

8.1.1 Předzpracování textu

Předzpracování příspěvků z Twitteru probíhalo jejich manuálním vyčištěním dvěma dobrovolníky (dva muži). Vyčištění mělo za úkol odhalit a opravit

gramatické chyby a překlepy. Následně byly příspěvky zkontrolovány pomocí slovníku v MS Excel.

Předzpracovaný text byl použit ke strojovému překladu v *Google Docs (dále označené GD)* s využitím funkce "GoogleTranslate", ovšem tyto překlady nebyly dostatečně kvalitní. Jako příklad lze uvést větu "Jsem na italštině", které byla přeložena jako "I am Italian". Z tohoto důvodu byl využit doplněk *GoogleTranslate (dále označené GT)* pro prohlížeč Chrome, který celou tabulku dokázal přeložit přímo Google Translatorem a takto vzniklé překlady byly kvalitnější.

Překlady z *Google Translatoru (GT)* byly následně ručně zkontrolovány a upraveny tak, aby výsledek odpovídal české předloze. Tato část opět byla provedena ručně dvěma dobrovolníky. Poslední částí byla kontrola překladů pomocí slovníku v MS Excel.

8.1.2 Ruční extrakce sentimentu

Abychom věděli, že extrahovaný sentiment bude relevantní k sentimentu vyjádřenému v příspěvcích, byl sentiment extrahován také ručně. Ruční extrakci sentimentu prováděli tři dobrovolníci (dva muži a jedna žena, věkové rozpětí 24 až 38 let). Extrakce probíhala jako zaslepená, aby nedošlo k ovlivnění hodnotitelů, tj. každý z dobrovolníků nevěděl nic o hodnoceném subjektu, dostal pouze texty a odevzdal texty s ohodnoceným sentimentem. Pro zjištění sentimentu byly využity příspěvky psané v češtině a hodnocení probíhalo binárně - 0 pro záporný sentiment a 1 pro kladný.

8.1.3 Výběr nástrojů pro extrakci sentimentu

Pro extrakci sentimentu byly vybrány dva nástroje kvůli ověření správnosti extrakce. Prvním z nich je CoreNLP (4.4.1) vybraný vzhledem k rozšířenosti a velikosti naučených textů a frází. Druhým z nich je Vader (4.4.2) jako jednodušší nástoroj, založený na slovnících pravidlech. Pro exrakci sentimentu pomocí CoreNLP bylo využito překladů z *Google Docs (GD)* a pro extrakci sentimentu pomocí Vader překladů z *Google Translator (GT)* a *Google Docs (GD)* (viz níže 8.1.4).

8.1.4 Neutrální sentiment

Oba tyto nástroje extrahují kromě pozitivního a negativního ohodnocení sentimentu i sentiment neutrální. Jelikož experiment byl koncipován tak, aby příspěvek byl označen jako pozitivní, či negativní, je nutné nalezení dodatečné polarity při nalezení neutrálního sentimentu.

U CoreNLP každá z pěti kategorií sentimentu (6.3.1) získává při ohodnocení pravděpodobnost, s jakou má daná věta konkrétní váhu sentimentu. Pokud je sentiment neutrální, tak druhá nejvyšší pravděpodobnost je buď negativní, nebo pozitivní, tudíž můžeme neutrální hodnotu nahradit hodnotou na základě druhé nejvyšší pravděpodobnosti.

Pro změnu u Vader byl při neutrálním sentimentu použit překlad pomocí *Google Docs (GD)* či upravené překlady s využitím synonym nebo parafrází, a tím získána další hodnota sentimentu.

8.1.5 Porovnání nástrojů

Pro určení vhodnosti použití nástroje na extrakci sentimentu je nutné ověřit míru shody extrahovaného sentimentu. Jako referenční hodnota nám budou sloužit data z manuální extrakce sentimentu. Srovnání je přehledně vyčísleno v následujících dvou tabulkách (tabulka 8.1) a (tabulka 8.2).

		Metoda extrakce sentimentu			
		Hashtag	Ruční	CoreNLP	Vader
Počet výskytů sentimentu	Kladný	808	729	348	709
	Záporný	204	283	664	303

Tabulka 8.1: Počet výskytů jednotlivých sentimentů při použití nástrojů

	Hashtag	Ruční extrakce	CoreNLP	Vader
Hashtag	1	0,88	0,43	0,77
Ruční extrakce	0,88	1	0,52	0,78
CoreNLP	0,43	0,52	1	0,50
Vader	0,77	0,78	0,50	1

Tabulka 8.2: Shoda mezi jednotlivými metodami pro extrakci sentimentu

První tabulka (tabulka 8.1) ukazuje celkový počet kladně a záporně extrahovaných sentimentů přes všechny subjekty. Je vidět, že míra shody mezi označením příspěvků od uživatelů a ruční extrakcí je vysoká (88%) a uživatelé jsou velmi často pozitivní. Dále je na první pohled zřejmé, že CoreNLP přiřazuje negativní sentiment ve více případech než Vader nebo manuální extrakce sentimentu. V druhé tabulce (tabulka 8.2) vidíme poměr shody mezi jednotlivými metodami extrakce sentimentu. Jak můžeme vidět, shoda mezi CoreNLP a ostatními způsoby extrakce sentimentu je zhruba 50%. V případě použití knihovny Vader je shoda s manuální extrakci na 78% a s

označením od uživatelů na 77%. Z toho důvodu byl Vader vybrán jako nástroj pro finální extrakci sentimentu.

8.2 Analýza tepové frekvence

Tepová frekvence je u každého uživatele jiná, viz tabulka (tabulka 8.3). Aby měla data od všech uživatelů stejnou vypovídající hodnotu, je nutná jejich normalizace. A protože měření probíhalo v některých případech i mimo dobu sběru sentimentu, bude potřeba data filtrovat.

		Tepová frekvence (údery za min)			
		Maximální	Minimální	Průměrná	Medián
Subjekt	101	155	45	75	72
	102	167	55	87	86
	103	147	39	67	64
	104	163	49	78	78
	201	145	53	78	75
	203	140	58	91	92
	204	169	53	90	90

Tabulka 8.3: Hodnoty tepové frekvence napříč uživateli

8.2.1 Předzpracování dat tepové frekvence

První částí předzpracování tepové frekvence je její normalizace, kde byla zvolena min-max normalizace (6.4.1). Tuto normalizaci jsem zvolil vzhledem k tomu, že zachovává distribuci hodnot a celkový rozsah vyskytující se tepové frekvence je poměrně malý. Jako ukázku si můžeme vzít měření subjektu 101, kde je minimální tepová frekvence 45, maximální 155 s průměrnou tepovou frekvencí 75,28 úderů za minutu a směrodatná odchylka je 55.

Data pro tepovou frekvenci byla měřena s výjimkami 24 hodin denně, ovšem data pro sentiment byla zaznamenávána pouze od 8:00 do 22:00. Podle časů jednotlivých tweetů byla nakonec vytvořena každodenní okna, podle kterých byla tepová frekvence ořezána. Kdybychom tepovou frekvenci neořezali, zpracovávali bychom i data mimo sběr sentimentu, a tedy nevíme, co uživatel prožíval a jaké byly jeho pocity. Pravidla pro výběr dat tepové frekvence jsou následující:

1. první den experimentu je okno pro výběr tepové frekvence mezi 17:00 a 23:15 hodinami (začátek experimentu je vždy v 18:00 hodin)

2. běžné dny experimentu je okno pro výběr dat mezi 7:00 a 23:15 hodinami

3. poslední den experimentu je okno pro výběr dat nastaveno mezi 7:00 a 19:00 hodinami (konec experimentu je vždy v 18:00 hodin)

Konec mimo první a poslední den měření ve 23:15 byl zvolen vzhledem k občasnému výskytu tweetů kolem 23 hodiny.

8.3 Návrh implementace

V této části již máme připravena data a budeme se zabývat tím, jak data spojit, jaké použít metody strojového učení či jak zaznamenávat výsledky.

8.3.1 Návrh fúze dat

Nyní jsou již data sentimentu (8.1), tak tepové frekvence (8.2) připravena pro jejich fúzi. Zde nastává problém, že časové značky dat sentimentu a tepové frekvence mají jinou granularitu. Je tedy nutné nalézt metodu, která najde nejbližší časové značky v obou datových setech a na základě nichž bude možné data spojit.

Abychom nemuseli metodu implementovat vlastními silami, použijeme metodu merge_asof z knihovny pandas, používané po celou dobu pro analýzu a úpravu dat. Pro její správnou činnost musí být data seřazena. Metoda nabízí tři možnosti, jak lze data spojit:

- *backward* - doplňuje data odzadu (zprava do leva), až do další změny (na začátku dat mohou vzniknout nedoplněná data)
- *forward* - doplňuje data odpředu (zleva do prava), až do další změny (na konci dat mohou vzniknout nedoplněná data)
- *nearest* - doplňuje data k nejbližší časové značce

Vhodná varianta je *nearest* (9.4), která spojuje data pomocí nejbližší časové značky, což odpovídá metodě interpolace dat s názvem Zero-Order Hold 6.3.3.

8.3.2 Zaznamenávání výsledků

Při spouštění velkého počtu experimentů rostou nároky na zaznamenávání a zpracování výsledků. Pouze jejich záznam do CSV souboru přestává stačit a potřeba rychle vyhledávat ve výsledcích a porovnávat je vede k použití specializovaného nástroje.

Jedním z nich je například MLFlow[1], který dokáže jak zaznamenávat výsledky, tak i vytvářet vlastní spustitelné projekty s experimenty či ukládat modely strojového učení. Další výhodou zaznamenávání v MLFlow je okamžitá možnost vizualizace a porovnání jednotlivých výsledků (obrázek 8.1).

Default > Comparing 3 Runs

Run ID:	5aaaed43e9ec4377a88106...	e1e4fb4cb7a648e286f548...	5360a673a1b04f3bbaaee9...
Run Name:			
Start Time:	2019-05-10 16:32:50	2019-05-10 11:46:45	2019-05-09 20:05:35
Parameters			
Setting	DecisionTreeClassifier(clas...	AdaBoostClassifier(algorith...	SVC(C=1.0, cache_size=20...
Metrics			
Accuracy	0.668	0.645	0.638
Accuracy_stdev	0.045	0.053	0.05
F1	0.777	0.75	0.743
F1_stdev	0.045	0.056	0.054

Obrázek 8.1: Porovnání tří experimentů v MLFlow

8.3.3 Balancování dat

Aby metody strojového učení fungovaly správně, je pro učení s učitelem nutné mít vybalancovaná vstupní data. Očekává se, že počet vstupních dat pro každou třídu bude stejný. Pokud by tak nebylo, strojové učení by následně bylo zaujaté kategorizovat data do té třídy, která je nejvíce zastoupená. Po vybalancování tříd jsou data připravena pro použití strojového učení.

[1] https://mlflow.org/

8.3.4 Rozdělení dat

Data pro strojové učení je potřeba rozdělit na trénovací a testovací. Jako základní poměr byl zvolen 75:25, tj. 3 trénovací záznamy na 1 testovací. Aby trénovací a testovací množina obsahovala data od všech subjektů, budou data pro trénování a testování vybírána postupně od každého subjektu.

8.3.5 Metody strojového učení

Pro aplikaci metod strojového učení byla vybrána knihovna scikit-learn, která obsahuje již připravené metody pro klasifikaci, regresi a další. Výhodou použití této knihovny je fakt, že všechny metody strojového učení zde mají standardizované vstupní parametry a lze je následně snadno zaměnit za jiné. Mezi vybrané metody patří SVM (7.3.1), Naive Bayes (7.3.2), Rozhodovací strom (7.3.3), Random Tree (7.6.1) a AdaBoost (7.6.2). Metody strojového učení jsou vybrány vzhledem k existujícím výzkumům (2.1) a jsou běžně používané pro binární klasifikaci.

8.3.6 Křížová validace

Pro ověření modelu strojového učení nám slouží křížová validace (7.4.5). Počet částí křížové validace je roven 4, a odpovídá tedy rozdělení dat (8.3.4). Program proběhne 4x a během každého běhu se posouvá okénko s trénovacími a testovacími daty, a tak dojde k ověření metody strojového učení nad celým datasetem.

8.3.7 Zaznamenávané metriky

Aby bylo možné jednotlivé metody mezi s sebou porovnávat, je nutné vybrat metriky, které se budou během běhu programu zaznamenávat. Vybrány proto byly: matice záměn (7.4), přesnost (7.4.1), preciznost (7.4.2), úplnost (7.4.3) a F-míra, která je nejdůležitější metrikou pro binární klasifikaci a představuje rovnováhu mezi precizností a úplností (7.4.4), čímž nám dá nejlepší představu o modelu. Kvůli použití křížové validace byla data pro tabulku záměn postupně sčítána, ale pro ostatní metriky byl vyhodnocen průměr měření a směrodatná odchylka.

8.3.8 Přidání vstupních proměnných

Doposud se počítalo pouze s uplatněním jednoho příznaku (feature) pro strojové učení, a to byla normalizovaná tepová frekvence. Ovšem při experimentu

o sobě zaznamenali účastníci i další údaje jako jsou věk a pohlaví. A také je možnost použití ID účastníka jako vstupu. Přidáním dalších příznaků (features) můžeme docílit vylepšení výsledků stávající metody strojového učení.

Všechny tyto hodnoty byly jako nové sloupce přidány přímo do CSV souboru obsahujícího reprezentaci sentimentu, věk, pohlaví zakódované číselným údajem (kde 0 znamená muž a 1 je žena) a ID účastníka.

Pokud máme větší počet účastníků experimentu (N) s různými ID, nelze pouze přidat sloupeček s příznakem ID (feature ID). Metody strojového učení by se snažily nalézt spojitost mezi těmito daty, ať již by se jednalo o jejich řazení či by vnitřně mohly počítat například průměr, kde by ID dvou účastníků mohlo implikovat ID třetího účastníka. Tento problém řeší metoda one-hot encoding.

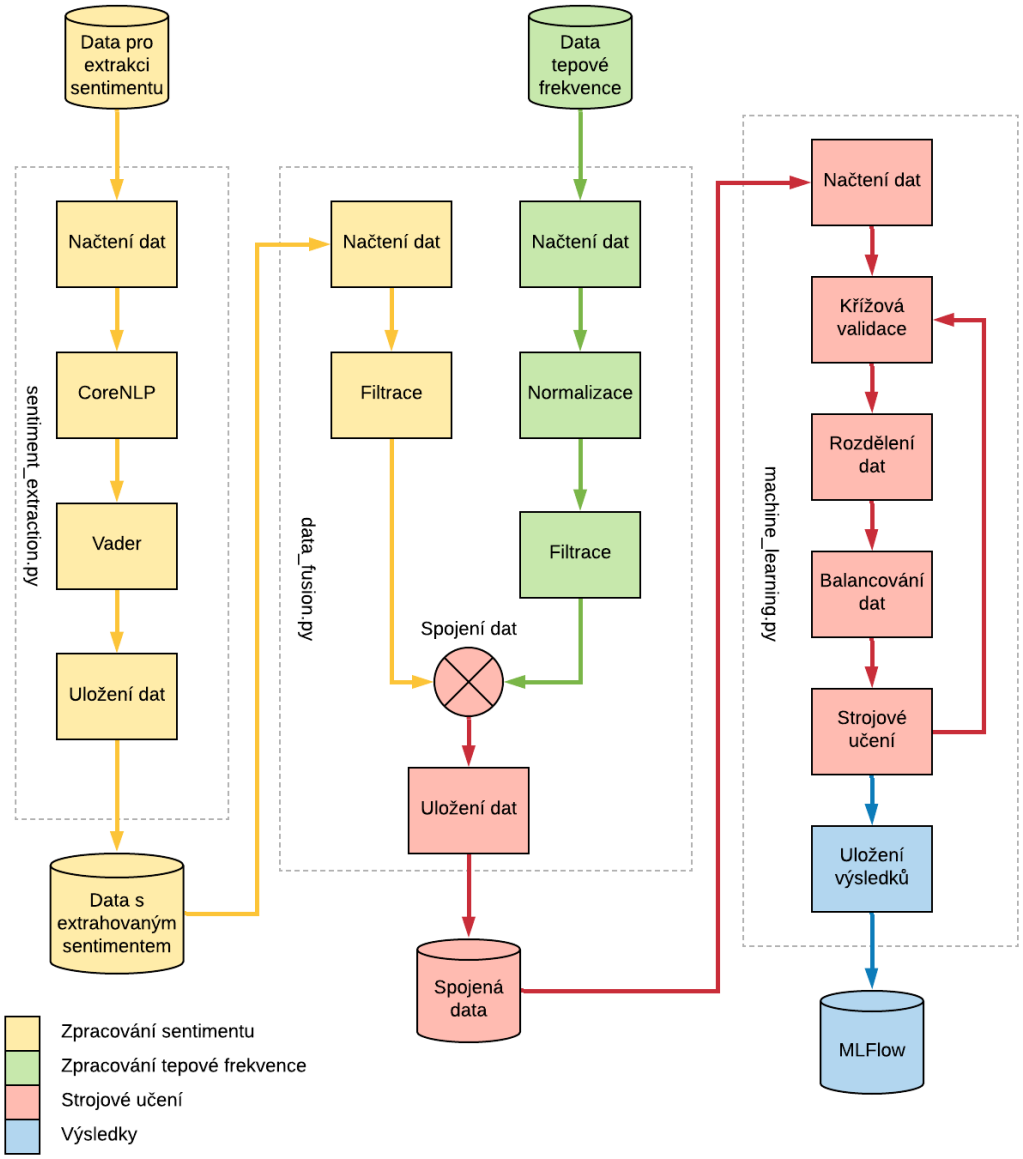
Pro každé ID účastníka je vytvořen speciální sloupec v datech (zde tedy N sloupců), který obsahuje 1, pokud ID účastníka odpovídá a 0, pokud ne. Zjednodušeně účastník s ID 101 (první ID) bude mít ve sloupci "User 1" samé 1 a ve zbývajících 0. Obdobně pak následující účastníci (tabulka 8.4).

ID účastníka	Tep. frekv.	User 1	User 2	User 3	User 4	User 5
101	74	1	0	0	0	0
101	85	1	0	0	0	0
101	41	1	0	0	0	0
102	95	0	1	0	0	0
102	86	0	1	0	0	0
102	79	0	1	0	0	0
103	61	0	0	1	0	0
103	69	0	0	1	0	0
103	112	0	0	1	0	0

Tabulka 8.4: Ukázka použití one-hot encoding pro pět účastníků experimentu

8.4 Implementace

V této části se budu věnovat samotné tvorbě programu. Celý program zde bude rozebrán od načtení dat po uložení výsledků spolu s ukázkami kódu. Jak program funguje je zobrazeno v diagramu datových toků (obrázek 8.2).



Obrázek 8.2: Diagram datových toků programu

8.4.1 Struktura projektu

Projekt byl vytvořen v jazyce Python 3.7, který byl zvolen kvůli dostupnosti knihoven pro strojové učení. Souborová struktura je následující:

- *složka data* - složka pro vstupní soubory dat a ukládání mezivýsledků
  - *fitbit101.csv až fitbit204.csv* - data s tepovou frekvencí
  - *twitter101.csv až twitter204.csv* - data pro extrakci sentimentu
  - *sentiment.csv* - data s extrahovaným sentimentem
  - *fusion_hr_sentiment.csv* - spojená data tepové frekvence a sentimentu
- *main.py* - obsluha celého programu
- *sentiment_extraction.py* - extrakce sentimentu
- *data_fusion.py* - zpracování a filtrace tepové frekvence, filtrace sentimentu a spojení dat
- *machine_learning.py* - strojové učení a zaznamenávání výsledků

Pro běh programu jsou potřebné následující knihovny: pycorenlp (0.3.0), mlflow (0.9.0), sklearn (0.21.0), vaderSentiment (3.2.1) a pandas (0.24.2).

Dále je také potřeba spuštěná instance CoreNLP [2] na portu 9000 a MLFlow [3] na portu 5000.

Následně je možné spustit program pomocí main.py a výsledky nalezneme v konzoli a také uložené v systému MLFlow.

8.4.2 Vstupní data

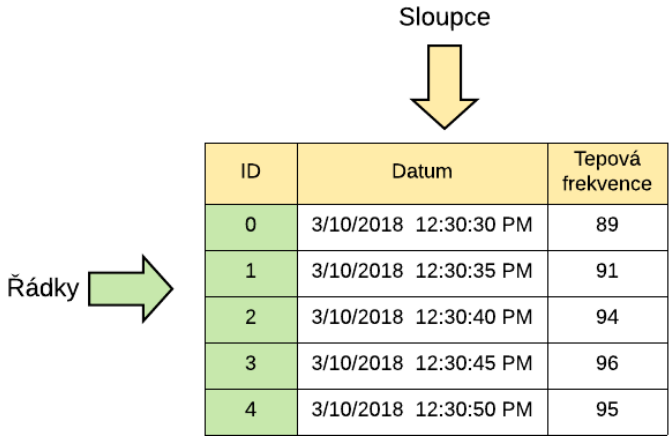
Data pro diplomovou práci byla dodána zadavatelem. Jednalo se o sedm CSV souborů obsahujících záznamy z měření tepové frekvence, kde název souboru se skládal z prefixu "fitbit" a ID účastníka experimentu. V souborech je možné nalézt v prvním sloupečku datum měření a ve druhém hodnotu tepové frekvence.

Druhými vstupními daty je sentiment ze sítě Twitter, která se nalézala v Google Docs Spreadsheet. Zde také docházelo k překladům a následně byla data exportována do CSV souborů s názvem "twitter" a ID účastníka. Všechny soubory jsou uloženy v projektu ve složce data.

8.4.3 Načtení dat

Data jsou v projektu uchovávána v Pandas DataFrame, což je dvoudimenzionální tabulka uchovávající záznamy s popisem jednotlivých os (obrázek 8.3).

[2] https://stanfordnlp.github.io/CoreNLP/download.html
[3] https://github.com/mlflow/mlflow



Obrázek 8.3: Ukázka Pandas DataFrame

Načtení dat bude ukázáno pro data s tepovou frekvencí. Zde můžeme vidět pojmenování jednotlivých sloupců a nastavení typů dat a parsování dat s následným uložením do struktury hr.

```
import pandas as pd
...
# Pripravu dataframe
hr = pd.DataFrame()
hr_names=['Date␣Time', 'HR']
hr_dtypes = {'Date␣Time': 'str', 'HR': 'int64'}
hr_parse_dates = ['Date␣Time']
...
# Pro kazdy soubor se nactou data a pridaji se do
   ↪ dataframe
for d in data:
    dx = pd.read_csv('data/fitbit' + d + '.csv', header =
       ↪ None, names = hr_names,
                    dtype = hr_dtypes, parse_dates =
                       ↪ hr_parse_dates)
    hr = hr.append(dx)
```

8.4.4 Extrakce sentimentu

Po načtení dat, která je obdobná, jako je popsáno výše (8.4.2), můžeme přejít k extrakci sentimentu. Pro extrakci sentimentu byly zvoleny nástroje CoreNLP ve verzi 3.9.2 a Vader s verzí 3.2.1 (8.1.3). Nyní se podíváme na

to, jak tyto nástroje použít a následně si ukážeme jak nakládat s neutrálním sentimentem.

Pro CoreNLP je nutné mít nainstalovanou Java 1.8 a po stažení lze spustit CoreNLP následujícím příkazem a poté naslouchá na portu 9000:

```
java -mx4g -cp "*" edu.stanford.nlp.pipeline.
   ↪ StanfordCoreNLPServer -port 9000
```

Následně můžeme importovat třídu CoreNLP do Pythonu a začít používat CoreNLP následujícím způsobem:

```
from pycorenlp import StanfordCoreNLP
...
nlp = StanfordCoreNLP('http://localhost:9000')
...
res = nlp.annotate(row['Text␣EN␣GTE'],
       properties={
           'annotators': 'sentiment',
           'outputFormat': 'json',
           'timeout': 50000,
})
...
sent_val = int(res['sentences'][0]['sentimentValue'])
```

Pokud nalezneme neutrální sentiment (mající hodnotu 2), je vybrán sentiment s druhou nejpravděpodobnější hodnotou:

```
sent_prob_all = res['sentences'][0]['sentimentDistribution'
   ↪ ]

if sent_val == 2:
    max2 = sorted(sent_prob_all)[-2] # 2. nejvyssi cislo
    index_of_max2 = sent_prob_all.index(max2)
    sent_val = index_of_max2
```

Vader pro svůj běh nepotřebuje mít spuštěný žádný server a stačí pouze importovat příslušnou knihovnu. Po importu do Pythonu jej lze použít například následovně:

```
from vaderSentiment.vaderSentiment import
    ↪ SentimentIntensityAnalyzer
...
analyser = SentimentIntensityAnalyzer()
vader_result = []
...
snt = analyser.polarity_scores(row['Text␣EN␣GTE'])
if snt['compound'] > 0:
    vader_result.append(1)
elif snt['compound'] < 0:
    vader_result.append(0)
```

Hodnota compound nám udává hodnotu sentimentu v rozmezí od -1 pro negativní sentiment až do 1 pro pozitivní sentiment. Stejně jako u CoreNLP může nastat případ, kdy je hodnota sentimentu neutrální. V tomto případě ovšem nelze vzít hodnotu druhého nejpravděpodobnější sentimentu, a tak bylo využito ve sporných případech překladu z *Google Docs (GD)* či byl překlad parafrázován.

8.4.5 Fúze dat

Nyní již máme extrahovaný sentiment a budeme pokračovat jeho filtrací. Filtrace má shodná pravidla jako filtrace tepové frekvence (8.2.1). Takto již máme připraven sentiment pro fúzi dat.

Dalším krokem je načtení dat pro tepovou frekvenci (8.4.2) a dále je nutné data normalizovat. Jak již bylo zmíněno (8.2.1), bude použita min-max normalizace, kterou lze provést následovně:

```
hr['HR␣norm'] = hr.groupby('Subject')['HR'].transform(
    ↪ lambda x: (x - x.min()) / (x.max()- x.min()))
```

Do DataFrame hr, který obsahuje absolutní hodnoty tepové frekvence, se takto přidá další sloupec "HR norm", kde se pomocí groupby vyberou jednotlivé subjekty a následně je proveden výpočet nové normalizované hodnoty tepové frekvence.

Následně můžeme opět provést filtraci dat, ke které nám stejně jako v případě sentimentu slouží funkce filter_days (8.2.1). Takto máme připravenou tepovou frekvenci pro fúzi dat.

Spojení dat probíhá pomocí následujícího algoritmu; je využita funkce merge_asof z balíku pandas (8.3.1):

```
data = ['101','102','103','104','201','203','204']
merged_data = pd.DataFrame([])
for d in data:
    hr_merge = hr[hr['Subject'] == int(d)]
    se_merge = se[se['Subject'] == int(d)]
    merged = pd.merge_asof(hr_merge, se_merge, left_on = '
        ↪ Date Time', right_on= 'Real Date Time', direction
        ↪ ='nearest')
    merged_data = merged_data.append(merged)
```

Pro každý měřený subjekt (data) jsou vybrány hodnoty tepové frekvence (hr_merge) a sentimentu (se_merge). Následně jsou oba typy dat spojeny pomocí vybrané metody uvedené výše. Takto spojená data jsou následně uložena (merged_data) a jsou tedy připravena pro použití strojového učení.

8.4.6 Strojové učení

Data pro implementaci strojového učení jsou již připravena, v této části je popsáno, jak jsou data rozdělena na učící a testovací, metody strojového učení, jejich použití a zaznamenávání výsledků.

Křížová validace a rozdělení dat

Data pro strojové učení máme sice připravena, ale nyní je důležité je rozdělit na trénovací a testovací data. Během křížové validace (8.3.6) dojde pokaždé k rozdělení části dat na trénovací a testovací v poměru 75:25 (8.3.4). Vždy se vybere okénko pro testovací data, zbytek je určen jako data trénovací. V dalším průběhu se okénko posune.

```
# Velikost krizove validace
k_fold = 4
i = 0
while i < k_fold:
    learn_list = []
    test_list = []

    # Rodeleni dat po subjektech
    group = ml_data.groupby('Subject_x')
    for ml_group in group:
```

```
        # Zjisteni velikosti okenka podle subjektu
        group_data = ml_group[1]
        window = int(round(len(group_data)/k_fold))
        start = i * window
        end = (i+1) * window

        # Rozdeleni dat
        for index, row in group_data.iterrows():
            if index > start and index <= end:
                test_list.append(row)
            else:
                learn_list.append(row)
    i = i + 1
```

Data jsou rozdělena podle subjektů a pro každý subjekt je vypočítána velikost okénka (window). Velikost okénka je zaokrouhlená velikost celkového počtu dat subjektu vydělená velikostí křížové validace (k_fold). V poslední části jsou data rozdělena na trénovací a testovací podle jejich indexu. Takto připravená data mohou následně podstoupit vybalancování.

Metody strojového učení

Pro aplikaci metod strojového učení byla vybrána knihovna scikit-learn, která obsahuje již připravené metody pro klasifikaci, regresi a další. Výhodou použití této knihovny je fakt, že všechny metody strojového učení zde mají standardizované vstupní parametry a lze je následně jednoduše vyměňovat za jiné. Zde je ukázka použití knihovny s SVM:

```
from sklearn.svm import SVC
...
clf = svm.SVC(kernel='rbf', gamma='scale')
while i < k_fold:
    ...
    clf.fit(x_learn, y_learn)
    clf_predictions = clf.predict(x_test)
```

Použití scikit-learn pro storojové učení je velmi jednoduché, nejprve se zvolí a nastaví klasifikátor (clf) a následně je zavolána funkce klasifikátoru fit, které jsou předána data obsahující příznaky (features) a výsledky. Poté

klasifikátor vyhodnotí trénovací sadu dat díky funkci predict a výsledky jsou uloženy v clf_predictions pro další zpracování.

Zaznamenávání výstupů

Nyní si ukážeme, jak zaznamenávat vybrané metriky (8.3.7) do systému MLFlow. Stačit nám k tomu bude knihovna mlflow a spuštěný server na portu 5000. Zde je ukázka se záznamem přesnosti:

```
import mlflow.sklearn
from mlflow import log_metric, log_param
...
mlflow.set_tracking_uri("http://127.0.0.1:5000")
...
log_param("Setting", clf_log)
log_metric("Accuracy", accuracy_avg)
```

Následně lze například v nejjednodušší formě logovat nastavené parametry klasifikátoru uložené v clf_log nebo průměrnou přesnost, která se nachází v proměnné accuracy_avg.

Všechny vybrané metriky pro zaznamenávání již obsahuje knihovna scikit-learn implementované. Během programu jsou zaznamenávány mezivýsledky a na konci jsou odeslány do MLFlow a vytištěny do konzole.

9 Výsledky

Nyní již máme celé zpracování dat od extrakce sentimentu po implementaci modelů strojového učení hotové a můžeme zhodnotit a diskutovat jednotlivé výsledky.

Základní nastavení modelů strojového učení vychází nejprve z tutoriálů na scikit-learn[1]. Po získání více zkušeností s použitím konkrétních modelů a výběru nejvhodnějších z nich na základě průběžných výsledků, došlo k úpravě nastavení modelů tak, abychom dostali co nejlepší výsledky.

Nejprve byly všechny modely natrénovány za použití základního výběru příznaku (feature), tj. normalizované tepové frekvence a to s použitím vybalancovaní celého datasetu (9.1), či pouze balancovanými trénovacími daty (9.2). Pokud byla použita nevybalancovaná data, modely strojového učení předpovídaly zařazení pouze do jedné kategorie. Což je způsobeno nevyváženým poměrem (zhruba 2:1) extrahovaného pozitivního a negativního sentimentu (tabulka 8.1).

Na základě výsledků těchto dvou postupů došlo k vylepšení metod (9.3), konkrétně k úpravě dvou nejlepších metod strojového účení, a to SVM (9.3.1), AdaBoost (9.3.2) a Rozhodovacího stromu (9.3.3).

Součástí výsledků je také zhodnocení vlivu různých nastavení (8.3.1) spojení dat (sentimentu a tepové frekvence) funkce merge_asof (9.4).

Následně byl metodou AdaBoost natrénován model za použití dalších příznaků (features) (9.5). Tudíž je možné zhodnotit jejich vliv na zvolenou metodu strojového učení v porovnání s předchozími výsledky.

9.1 Výsledky s balancovanými daty

Následující tabulka 9.1, ukazuje výsledky metod strojového učení s kompletně vybalancovanými daty, tj. třídy s pozitivním i negativním sentimentem jsou v trénovacích i testovacích datech stejně velké.

[1] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

	Naive-Bayes	Decision T.	Random F.	AdaBoost	SVM
Přesnost	0.5437	0.5659	0.5657	0.5541	0.5541
Preciznost	0.5433	0.5607	0.5602	0.5382	0.5387
Úplnost	0.5574	0.5962	0.5986	0.7666	0.7582
F-míra	0.5473	0.5767	0.5775	0.6314	0.6286
TP	70502	75430	75708	97416	96472
FP	59348	59195	59518	83539	82503
FN	57510	52582	52304	30596	31540
TN	68664	68817	68494	44473	45509

Tabulka 9.1: Výsledky s vybalancovanými daty

Směrodatná metrika pro nás bude F-míra, která je při binární klasifikaci brána jako míra přesnosti testu (8.3.7). Z tabulky je patrné, že F-míra se při použití kompletně vybalancovaných dat pohybuje mezi 54 a 63 %, což lze považovat za velmi nízký výsledek, vzhledem k očekávaní F-míry přes 70 %. Jedním z důvodů může být příliš velké omezení dat balancováním, kde při extrakci pomocí Vaderu máme záporný sentiment pouze v 303 případech extrakce sentimentu vůči 709 případům kladného sentimentu (tabulka 8.1).

9.2 Výsledky s balancovanými trénovacími daty

Výsledky s pouze vybalancovanými trénovacími daty zobrazuje tabulka 9.2. Metody strojového učení jsou natrénovány na vybalancovaných trénovacích datech, přičemž testovací data jsou ponechána původní.

	Naive-Bayes	Decision T.	Random F.	AdaBoost	SVM
Přesnost	0.5536	0.5813	0.5822	0.6421	0.6382
Preciznost	0.7329	0.7454	0.7451	0.7291	0.7294
Úplnost	0.5583	0.5973	0.5996	0.7674	0.7587
F-míra	0.6318	0.6624	0.6637	0.7474	0.7429
TP	165821	177387	178103	227489	224775
FP	59348	59195	59518	83539	82503
FN	129656	118090	117374	67988	70702
TN	68664	68817	68494	44473	45509

Tabulka 9.2: Výsledky s vybalancovanými trénovacími daty

V obou případech je vidět, že nejlépe vycházejí metody AdaBoost a SVM s F-mírou kolem 75 % respektive 74 %, kde F-míra nám představuje přesnost

testu. Proto se budeme dále snažit vylepšit právě tyto dvě metody strojového učení.

9.3 Vylepšení metod

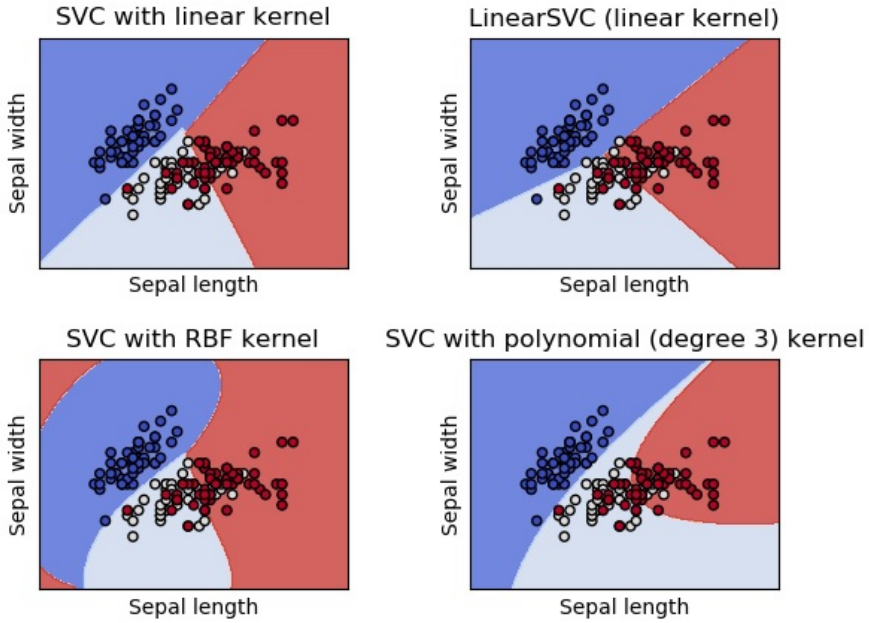
V následující části si ukážeme, jaké jsou možnosti pro dosažení zlepšení predikce jednotlivých metod strojového učení. Toho dosáhneme pomocí nastavení vhodnějších parametrů konkrétního algoritmu, čímž vylepšíme výstupní metriky modelů.

9.3.1 Úprava SVM

SVM metoda z knihovny scikit-learn[2] dovoluje nastavit rozdělení dat pomocí čtyř různých funkcí (obrázek 9.1):

- *lineární* - rozdělení pomocí lineární funkce
- *polynomiální* - rozdělení polynomiální funkcí, můžeme nastavit stupeň polynomu
- *sigmoid* - rozdělení hyperbolickou funkcí (tanh)
- *RBF*[3] - rozdělení pomocí radiální bázové funkce

[2]https://scikit-learn.org/stable/modules/svm.html
[3]Radial Basis Function



Obrázek 9.1: SVM ukázka rozdělení dat podle funkcí

Pro optimalizaci modelu tedy postupně použijeme všechny čtyři možnosti a polynomiální funkce zkusíme nastavit stupeň polynomu na 2 a 3 a porovnáme dosažené výsledky (tabulka 9.3). Díky rozdělení pomocí různých funkcí můžeme získat novou hranici mezi klasifikovanými třídami jak je vidět na obrázku (obrázek 9.1).

	Lineární	Poly d=2	Poly d=3	RBF	Sigmoid
Přesnost	0.4827	0.5172	0.3502	0.6382	0.4651
Preciznost	0.7412	0.5524	0.7206	0.7294	0.6674
Úplnost	0.3925	0.2205	0.1131	0.7587	0.4670
F-míra	0.5107	0.3108	0.1934	0.7429	0.5488
TP	165821	27843	33633	224775	137899
FP	59348	23352	13301	82503	68942
FN	116767	100169	261844	70702	157578
TN	87658	104660	114711	45509	59070

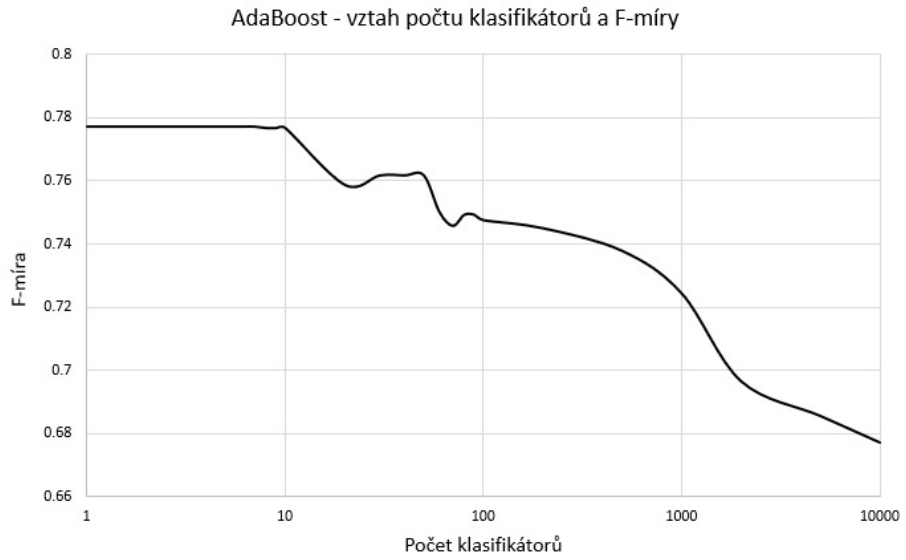
Tabulka 9.3: Úprava SVM

Jak je z výsledků F-míry patrné, nepodařilo se vylepšit predikci pomocí metody SVM použitím jiné rozdělovací funkce. Tudíž nejlepších hodnot dosahuje model s použitím rozdělovací funkce RBF. I přesto, že je dosažená F-míra vyšší než očekávaných 70 %, pokusíme se jí ještě zlepšit úpravou

druhého modelu využívající AdaBoost metodu strojového učení.

9.3.2 Úprava AdaBoost

Vzhledem k tomu, že AdaBoost je ensamble metoda (7.6.2) a kombinuje různé druhy klasifikátorů, je jeden z nejdůležitějších parametrů jejich počet, který AdaBoost používá. Testování můžeme vidět v následujícím grafu (obrázek 9.2).



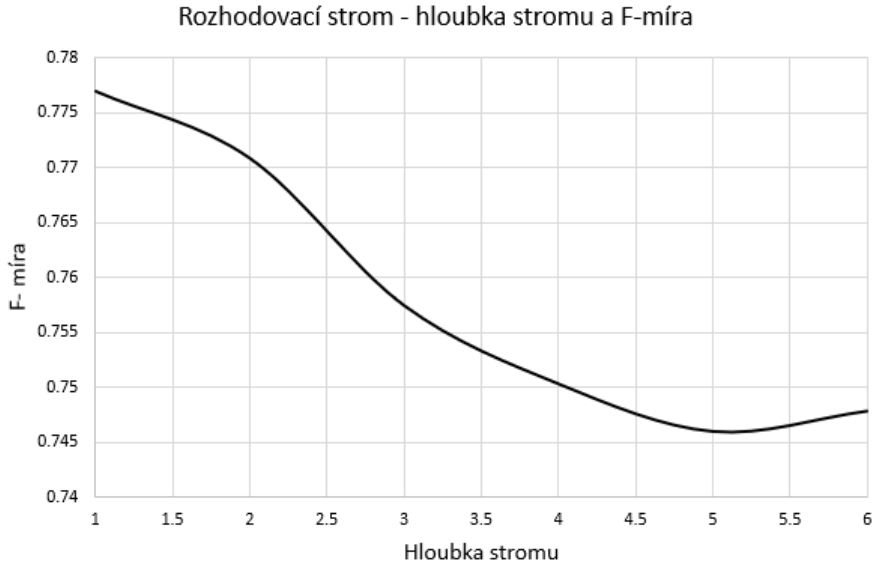
Obrázek 9.2: Vztah počtu klasifikátorů a F-míry v metodě AdaBoost

Jak je z grafu zřejmé, nejlépe funguje AdaBoost pouze s jedním klasifikátorem, kdy je F-míra necelých 78 % a při použití více než 10 klasifikátorů F-míra postupně klesá nejprve o jednotky procent. A poté až o více než 10 % při enormních počtech klasifikátorů.

V případě, že provedeme reverzní inženýring zdrojového kódu AdaBoost metody knihovny scikit-learn, najdeme tam, že základním klasifikátorem, který AdaBoost používá je rozhodovací strom s hloubkou 1. Pokusíme se tedy o další vylepšení výsledků úpravou parametrů metody rozhodovacího stromu.

9.3.3 Úprava rozhodovacího stromu

U rozhodovacího stromu můžeme měnit hloubku stromu, v úvodní části výsledků nebyla hloubka stromu nastavena, zkusíme se tedy podívat na to, jak hloubka stromu ovlivní výsledky (obrázek 9.3).



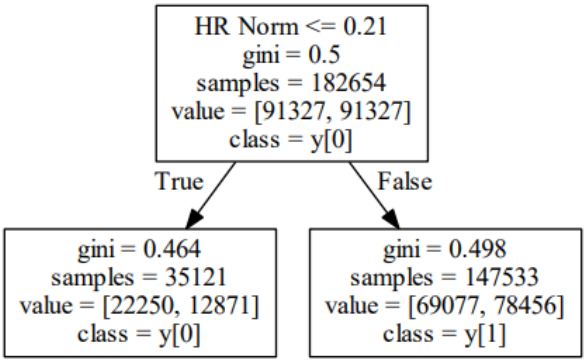
Obrázek 9.3: Vztah hloubky stromu a F-míry v metodě rozhodovací strom

Z obrázku je patrné, že s rostoucí hloubkou rozhodovacího stromu klesá F-míra a nejlepších výsledků dosahujeme s hloubkou 1. Nyní se podíváme na porovnání výsledků Rozhodovacího stromu s hloubkami 1 a 3 a metody Adaboost s počtem klasifikátorů 1 a 100 (tabulka 9.4).

	Rozhodovací strom		AdaBoost	
	h=1	h=3	k=1	k=100
Přesnost	0.6682	0.6521	0.6682	0.6421
Preciznost	0.7259	0.7305	0.7259	0.7291
Úplnost	0.8370	0.7871	0.8370	0.7674
F-míra	0.7769	0.7574	0.7769	0.7474
TP	247853	233345	247853	227489
FP	92849	85191	92849	83539
FN	47624	62132	47624	67988
TN	35163	42821	35163	44473

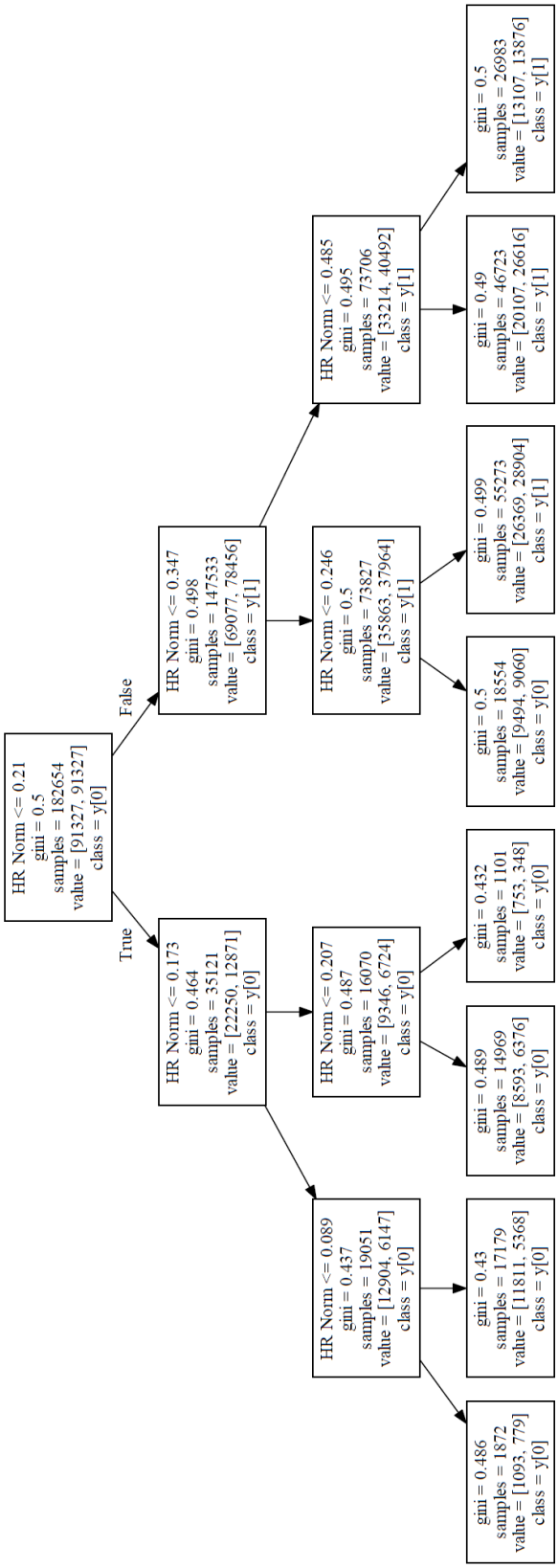
Tabulka 9.4: Porovnání nejlepší AdaBoost a rozhodovací strom

Jak Rozhodovací strom s hloubkou 1 tak AdaBoost s počtem klasifikátorů 1 mají nyní stejné výsledky a je tedy vidět, že používají stejný klasifikátor. Můžeme se tedy pokusit rozhodovací strom zobrazit a podívat se detailněji na rozhodovací pravidla. Na to se podíváme v dalších obrázcích (obrázek 9.4, obrázek 9.5).



Obrázek 9.4: Vizualizace rozhodovacího stromu s hloubkou 1

Zde vidíme, že strom se rozhoduje pouze na základě jediného pravidla tak, že klasifikuje normalizovanou tepovou frekvenci (HR Norm) $HR\,Norm <= 0,21$ negativně ($class = y[0]$) a $HR\ Norm > 0,21$ pozitivně ($class = y[1]$). Tento výsledek je zajímavý tím, že by bylo teoreticky možné, předchozí metody strojového učení efektivně nahradit pouze jednou podmínkou.



Obrázek 9.5: Vizualizace rozhodovacího stromu s hloubkou 3

Pokud se podíváme na strom s hloubkou 3 tak vidíme, že hodnota k dělení mezi kladnou a zápornou klasifikací dochází na hodnotě 0.246 normalizované tepové frekvence (HR Norm). Ovšem strom s touto hloubkou nedosahuje predikce s takovou výší F-míry jako strom s hloubkou 1. Můžeme tedy říci, že rozhodovací strom přiřazuje záporný sentiment (0) nižší tepové frekvenci, což je v rozporu s očekáváním z (2.3).

9.4 Ověření způsobu spojení dat merge_asof funkce

V kapitole 8.3.1 je popsáno, že funkce merge_asof má 3 možnosti jak spojit data. Nyní se podíváme na výsledky modelu natrénovaném pomocí rozhodovacího stromu hloubky jedna při použití všech 3 způsobů spojení dat (tabulka 9.5).

	Backward	Forward	Nearest
Přesnost	0.5527	0.6468	0.6544
Preciznost	0.7051	0.7094	0.7138
Úplnost	0.6722	0.8331	0.8400
F-míra	0.6280	0.7642	0.7706
TP	37712	28443	28454
FP	88596	100325	99558
FN	98642	48522	46811
TN	193416	244085	248666

Tabulka 9.5: Porovnání různých směrů zrpacování v merger_asof funkci

Z tabulky je zřejmé, že nejvyšší F-míru získáme pomocí způsobu spojení nearest, který přiřazuje tepovou frekvenci k nejbližšímu extrahovanému sentimentu na základě časových značek. Způsob spojení nearest byl použit ve všech výše zmíněných experimentech.

9.5 Více vstupních proměnných

Jak bylo zmíněno v (8.3.8), můžeme k normalizované tepové frekvenci přidat další příznaky (features), které ovlivní natrénování modelu strojového učení. Kromě již použité tepové frekvence jsme jako další příznaky zvolili pohlaví, věk a ID uživatele zakódované pomocí one-hot encoding (8.3.8). Přidání dalších vstupních příznaků proměnných jsem vyzkoušel při trénování modelu

pomocí metody AdaBoost s maximálně 10 klasifikátory. Výsledky můžeme vidět v následující tabulce (tabulka 9.6).

Jak je z výsledků patrné, přidání dalších vstupních proměnných nepřináší žádný užitek a oproti použití pouze normalizované tepové frekvence dostáváme v F-míře o zhruba 10 % horší výsledky. K vylepšení výsledků by mohlo pomoci použití další metody měření a to například pomocí galvanické změny odporu kůže.

	Pohlaví	Věk	ID uživatele	ID + věk	ID + pohlaví	Věk + pohlaví	Věk + pohlaví + ID
Přesnost	0.6308	0.6029	0.6009	0.6009	0.6134	0.6029	0.6134
Preciznost	0.7244	0.7370	0.7563	0.7563	0.7608	0.7370	0.7608
Úplnost	0.7396	0.6523	0.6188	0.6188	0.6455	0.6523	0.6455
F-míra	0.7261	0.6896	0.6796	0.6796	0.6980	0.6896	0.6980
TP	220500	194167	183972	183972	191362	194167	191362
FP	81357	66864	57493	57493	59618	66864	59618
FN	74977	101310	111505	111505	104115	101310	104115
TN	46655	61148	70519	70519	68394	61148	68394

Tabulka 9.6: Přidání vstupních proměnných

10 Závěr

V této diplomové práci je realizován program pro zjištění korelace mezi sentimentem extrahovaným z textu a sentimentem predikovaným z tepové frekvence měřené pomocí nositelné elektroniky.

Bylo vyzkoušeno více možností jak extrahovat sentiment z textu a jako nejlepší možnost se ukázalo použití knihovny Vader. Lepší výsledky Vader si vysvětluji tím, že se řídí pomocí slovníku a pravidel a nehledá v krátkých tweetech zbytečné složitosti.

Také bylo ukázáno, že nejlepší možnost pro spojení dat je spojovat tepovou frekvenci k extrahovanému sentimentu na základě nejbližších časových značek. Zde lze poukázat na to, že lidé píšící příspěvky neprožívají daný zážitek pouze do chvíle odeslání tweetu, ale setrvávají v daném rozpoložení déle.

Vyzkoušeno bylo i přidání dalších vstupních proměnných, které ovšem v tomto případě výsledky pouze zhoršovalo. Následně bylo vyzkoušeno více metod strojového učení z nichž nejlépe vychází rozhodovací strom s hloubkou 1.

Z hlediska výpočetní náročnosti běží predikce pomocí SVM modelu zhruba 60 minut, oproti ostatním metodám (AdaBoost, Rozhodovací strom, Random Forest), kde pro výpočet stačí 2 minuty, tedy 30x méně.

Výsledky ukazují, že při použití vybalancovaných trénovacích a testovacích dat dostaneme F-míru 64 %. Pro pouze vybalancovaná trénovací data pak F-míru 77 %. Zajímavé je, že k tomuto výsledku stačí pouze jedna podmínka v modelu Rozhodovacího stromu, přičemž nižší tepová frekvence je hodnocena jako negativní. To lze vysvětlit například tím, že skupinou účastníků experimentu byli studenti, kteří když se doma učili nebo nudili byli negativní, přičemž nevykonávali žádnou fyzickou aktivitu, tudíž jejich tepová frekvence byla nízká. Na druhou stranu lze usoudit, že pokud měli nějakou zajímavou činnost (cestování po městě či na výlet) byli spokojenější, ale jejich tepová frekvence byla vyšší a tudíž hodnocená kladně.

Bylo by proto vhodné rozšířit experiment o další účastníky hlavně z jiné věkové kategorie, abychom mohli potvrdit korelaci mezi extrahovaným sentimentem a sentimentem predikovaným z tepové frekvence.

Dále by bylo vhodné diplomovou práci rozšířit o použití First-order hold, vyzkoušet další možnosti extrakce sentimentu, použít jiné metody strojového učení či zjistit další vstupní příznaky účastníků experimentu nebo vytvořit příznaky čistě technické odvozené z již získaných dat.

Seznam obrázků

Seznam tabulek

Literatura

[1] Alpaydin, E. *Introduction to machine learning.* MIT Press, 2010. ISBN 9780262012430.

[2] Atkinson – Hilgard – Nolen-Hoeksema, S. *Atkinson and Hilgard's Introduction to Psychology.* Cengage Learning, 2014. Dostupné z: https://www.studocu.com/en/document/anton-de-kom-universiteit-van-suriname/sociale-psychologie/book-solutions/atkinson-hilgards-introduction-to-psychology-15th-edition/1162532/view. ISBN 9781844807284.

[3] BINDER, S. Průběh pulsní vlny v závislosti na elasticitě cévního systému na arteria radialis. 2009. Dostupné z: https://theses.cz/id/i69cnp/.

[4] Bishop, C. M. *Pattern recognition and machine learning.* Springer, 2006. ISBN 9780387310732.

[5] Breiman, L. Bagging predictors. *Machine Learning.* 8 1996, 24, 2, s. 123–140. ISSN 0885-6125. doi: 10.1007/BF00058655. Dostupné z: http://link.springer.com/10.1007/BF00058655.

[6] Burnham, K. P. – Anderson, D. R. – Burnham, K. P. *Model selection and multimodel inference : a practical information-theoretic approach.* Springer, 2002. ISBN 9780387224565.

[7] Čermák, V. Analysis of stock market sentiment with social media. 6 2018. Dostupné z: https://dspace.cuni.cz/handle/20.500.11956/99637.

[8] Čihák, R. *Anatomie.* Grada, třetí edition, 2016. ISBN 80-86317-34-5.

[9] Cortes, C. – Vapnik, V. Support-vector networks. *Machine Learning.* 9 1995, 20, 3, s. 273–297. ISSN 0885-6125. doi: 10.1007/BF00994018. Dostupné z: http://link.springer.com/10.1007/BF00994018.

[10] Didier C. Combatalade. Basics of HEART RATE VARIABILITY Applied to Psychophysiology. Technical report, 2010. Dostupné z: http://www.emfandhealth.com/HRVThoughtTechnology.pdf.

[11] DiNardo, J. Natural Experiments and Quasi-Natural Experiments. In *The New Palgrave Dictionary of Economics.* London: Palgrave Macmillan UK, 2016. s. 1–12. doi: 10.1057/978-1-349-95121-5{_}2006-1. Dostupné z: http://link.springer.com/10.1057/9780230280816_18https:

//doi.org/10.1057/978-1-349-95121-5_2006-1. ISBN 978-1-349-95121-5.

[12] Edwards, S. *The heart rate monitor book*. Polar CIC, 1992. ISBN 0963463306.

[13] Esuli, A. – Esuli, A. – Sebastiani, F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *IN PROCEEDINGS OF THE 5TH CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'06*. 2006, s. 417–422. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217.

[14] Fox, S. M. – Haskell, W. L. Physical activity and the prevention of coronary heart disease. *Bulletin of the New York Academy of Medicine*. 8 1968, 44, 8, s. 950–67. ISSN 0028-7091. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/5243890http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1750298.

[15] Freund, Y. – Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. 8 1997, 55, 1, s. 119–139. ISSN 0022-0000. doi: 10.1006/JCSS.1997.1504. Dostupné z: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[16] Froelicher, V. F. – Myers, J. *Exercise and the heart*. Saunders Elsevier, 2006. ISBN 9781416003113.

[17] Hampton, J. R. *EKG strucne, jasne, prehledne*. Grada, 2013. Dostupné z: https://www.grada.cz/ekg-strucne-jasne-prehledne-7189/?fbclid=IwAR2OziLQR-PF1ujtswRIpwtmyuIaItZ7a1ZlcVnQH76fuOx7rjFXLY_Nfhg. ISBN 9788024742465.

[18] Hastie, T. – Tibshirani, R. – Friedman, J. H. J. H. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2009. ISBN 9780387848587.

[19] Healey, J. A. – Picard, R. W. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. Technical report, 2004. Dostupné z: https://www.hpl.hp.com/techreports/2004/HPL-2004-229.pdf.

[20] Hercig, T. Aspects of sentiment analysis: technical report no. DCSE/TR-2015-04. 2015. Dostupné z: https://otik.uk.zcu.cz/handle/11025/21537.

[21] Honzík, P. *Strojové učení*. FEKT VUT Brno, 2006.

[22] Hulley, S. B. et al. *Designing clinical research*. Lippincott Williams & Wilkins, 2007. ISBN 9781608318049.

[23] Jan, S. Generování hudby pomocí neuronových sítí. 5 2017. Dostupné z: https://dspace.cvut.cz/handle/10467/70152.

[24] Kim, S. M. – Electrical, U. – Engineering, I. *Recognising Emotions and Sentiments in Text*. University of Sydney, 2011. Dostupné z: https://books.google.cz/books?id=hHFPMwEACAAJ.

[25] Kliegman, R. et al. *Nelson textbook of pediatrics*. Elsevier, 2019. ISBN 9781455775668.

[26] Kohlíková, E. *Fyziologie člověka: učební texty pro terérskou čkolu FTVS UK v Praze*. Univerzita Karlova, Fakulta tělesné výchovy a sportu, 2004. ISBN 80-86317-34-5.

[27] Kolkur, S. – Dantal, G. – Mahe, R. Study of Different Levels for Sentiment Analysis. *International Journal of Current Engineering and Technology*. 2015, 5. Dostupné z: https://inpressco.com/wp-content/uploads/2015/03/Paper32768-770.pdf.

[28] Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*. 5 2012, 5, 1, s. 1–167. ISSN 1947-4040. doi: 10.2200/S00416ED1V01Y201204HLT016. Dostupné z: http://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016.

[29] Liu, D. – Ulrich, M. Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors. Technical report, 2013. Dostupné z: http://cs229.stanford.edu/proj2013/LiuUlrich-ListenToYourHeart-StressPredictionUsingConsumerHeartRateSensors.pdf.

[30] Liu, Y. – Bi, J. W. – Fan, Z. P. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications*. 9 2017, 80, s. 323–339. ISSN 09574174. doi: 10.1016/j.eswa.2017.03.042. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0957417417301951.

[31] Loper, E. – Bird, S. NLTK. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* -, 1, s. 63–70, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. Dostupné z: http://portal.acm.org/citation.cfm?doid=1118108.1118117.

[32] Lukasová, A. – Šarmanová, J. *Metody shlukové analýzy*. SNTL, 1985.

[33] Lundberg, U. et al. Psychophysiological stress and emg activity of the trapezius muscle. *International Journal of Behavioral Medicine*. 12 1994, 1, 4, s. 354–370. ISSN 1070-5503. doi: 10.1207/s15327558ijbm0104{_}5. Dostupné z: http://link.springer.com/10.1207/s15327558ijbm0104_5.

[34] Maleňák, F. Mobilní systém pro monitorování sportovní aktivity. 6 2015. Dostupné z: https://dspace.vutbr.cz/xmlui/handle/11012/38895?locale-attribute=cs.

[35] Manning, C. D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. Technical report, 2014. Dostupné z: http://aclweb.org/anthology/P14-5010.

[36] Maron, M. E. – E., M. Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*. 7 1961, 8, 3, s. 404–417. ISSN 00045411. doi: 10.1145/321075.321084. Dostupné z: http://portal.acm.org/citation.cfm?doid=321075.321084.

[37] McCulloch, W. S. – Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*. 12 1943, 5, 4, s. 115–133. ISSN 0007-4985. doi: 10.1007/BF02478259. Dostupné z: http://link.springer.com/10.1007/BF02478259.

[38] Mikolov, T. et al. Efficient Estimation of Word Representations in Vector Space. 1 2013. Dostupné z: http://arxiv.org/abs/1301.3781.

[39] Mikolov, T. *Statistical language models based on neural networks*. PhD thesis, Brno University of Technology, 2012. Dostupné z: http://www.fit.vutbr.cz/~imikolov/rnnlm/thesis.pdf.

[40] Opitz, D. – Maclin, R. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*. 8 1999, 11, s. 169–198. ISSN 10769757. doi: 10.1613/jair.614. Dostupné z: https://jair.org/index.php/jair/article/view/10239.

[41] Parthasarathy, G. – Tomar, D. C. A survey of sentiment analysis for journal citation. *Indian Journal of Science and Technology*. 2015, 8, 35. ISSN 09745645. doi: 10.17485/ijst/2015/v8i35/55134. Dostupné z: www.indjst.org.

[42] Pennington, J. et al. Glove: Global vectors for word representation. *IN EMNLP*. 2014. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.671.1743.

[43] Plaminek, J. *Seberizeni : prakticky atlas managementu cilu, casu a stresu.* Grada, 2004. Dostupné z: https://is.muni.cz/publication/658316. ISBN 8024706717.

[44] Plhakova, A. *Ucebnice obecne psychologie.* Academia, 2005. ISBN 8020013873.

[45] Quinlan, J. R. J. R. – Ross, J. *C4.5 : programs for machine learning.* Morgan Kaufmann Publishers, 1993. Dostupné z: https://dl.acm.org/citation.cfm?id=152181. ISBN 1558602380.

[46] Rachman, F. H. – Sarno, R. – Fatichah, C. CBE: Corpus-based of emotion for emotion detection in text document. In *Proceedings - 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2016*, s. 331–335. IEEE, 2017. doi: 10.1109/ICITACEE.2016.7892466. Dostupné z: http://ieeexplore.ieee.org/document/7892466/. ISBN 9781509014347.

[47] Robergs, R. A. – Landwehr, R. THE SURPRISING HISTORY OF THE "HRmax=220-age" EQUATION. *An International Electronic Journal.* 2002, 5. Dostupné z: https://www.asep.org/asep/asep/Robergs2.pdf.

[48] Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology.* 1980, 39, 6, s. 1161–1178. ISSN 00223514. doi: 10.1037/h0077714. Dostupné z: http://content.apa.org/journals/psp/39/6/1161.

[49] Salai, M. – Vassányi, I. – Kósa, I. Stress detection using low cost heart rate sensors. *Journal of Healthcare Engineering.* 2016, 2016, s. 1–13. ISSN 20402309. doi: 10.1155/2016/5136705. Dostupné z: http://www.hindawi.com/journals/jhe/2016/5136705/.

[50] Salamon, J. – Moucek, R. Link between Sentiment and Human Activity Represented by Footsteps - Experiment Exploiting IoT Devices and Social Networks. In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, s. 450–457, 2016. doi: 10.5220/0005818204500457. Dostupné z: http://www.scitepress.org/Papers/2016/58182/58182.pdfhttp://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005818204500457. ISBN 978-989-758-170-0.

[51] Shi, Y. et al. Galvanic skin response (GSR) as an index of cognitive load. In *CHI '07 extended abstracts on Human factors in computing systems - CHI '07*, s. 2651, New York, New York, USA, 2007. ACM Press. doi: 10.1145/1240866.1241057. Dostupné z:

http://portal.acm.org/citation.cfm?doid=1240866.1241057. ISBN 9781595936424.

[52] Simjanoska, M. et al. ECG-derived Blood Pressure Classification using Complexity Analysis-based Machine Learning. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, s. 282–292. SCITEPRESS - Science and Technology Publications, 2018. doi: 10.5220/0006538202820292. Dostupné z: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006538202820292. ISBN 978-989-758-281-3.

[53] Slaměník, I. *Emoce a interpersonální vztahy*. Grada, 2011. ISBN 9788024733111.

[54] Socher, R. et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, s. 1631–1642. Dostupné z: https://aclanthology.info/papers/D13-1170/d13-1170.

[55] Špinar, J. – Vítovec, J. Tepová frekvence a kardiovaskulární onemocnění. *Interní medicína pro praxi*. 2009, 7, s. 315–318. Dostupné z: https://www.internimedicina.cz/pdfs/int/2009/07/02.pdf.

[56] Stehman, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*. 10 1997, 62, 1, s. 77–89. ISSN 0034-4257. doi: 10.1016/S0034-4257(97)00083-7. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0034425797000837?via%3Dihub.

[57] Sutton, R. S. – Barto, A. G. *Reinforcement learning : an introduction*. MIT Press, 1998. Dostupné z: https://books.google.co.uk/books?id=CAFR6IBF4xYC. ISBN 0262193981.

[58] Taelman, J. et al. Influence of mental stress on heart rate and heart rate variability. In *IFMBE Proceedings*, 22, s. 1366–1369. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-89208-3{_}324. Dostupné z: http://link.springer.com/10.1007/978-3-540-89208-3_324. ISBN 9783540892076.

[59] TAVISH SRIVASTAVA. Basics of Ensemble Learning Explained in Simple English, 2015. Dostupné z: https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/.

[60] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, s. 278–282. IEEE

Comput. Soc. Press, 1995. doi: 10.1109/ICDAR.1995.598994. Dostupné z: http://ieeexplore.ieee.org/document/598994/. ISBN 0-8186-7128-9.

[61] Tsytsarau, M. – Palpanas, T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*. 5 2012, 24, 3, s. 478–514. ISSN 13845810. doi: 10.1007/s10618-011-0238-6. Dostupné z: http://link.springer.com/10.1007/s10618-011-0238-6.

[62] Vokurka, M. – Hugo, J. *Velký lékařský slovník*. Maxdorf, 7. vyd. edition, 2007. ISBN 978-80-7345-130-1.

[63] Volná, E. *Nuronové sítě 1*. Ostravská univerzita v Ostravě, 2008.

[64] Wolpert, D. H. Stacked generalization. *Neural Networks*. 1 1992, 5, 2, s. 241–259. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80023-1. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0893608005800231.

[65] Yamakoshi, T. et al. Feasibility study on driver's stress detection from differential skin temperature measurement. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, s. 1076–1079. IEEE, 8 2008. doi: 10.1109/IEMBS.2008.4649346. Dostupné z: http://ieeexplore.ieee.org/document/4649346/. ISBN 978-1-4244-1814-5.

[66] Yousif, A. et al. A survey on sentiment analysis of scientific citations, 2017. ISSN 15737462. Dostupné z: https://doi.org/10.1007/s10462-017-9597-8.

[67] Zhang, L. – Wang, S. – Liu, B. Deep Learning for Sentiment Analysis : Survey. *CoRR*. 2018, abs/1801.0. Dostupné z: https://arxiv.org/ftp/arxiv/papers/1801/1801.07883.pdfhttp://arxiv.org/abs/1801.07883.