ALGORITHM FOR DETECTION OF POSITIVE AND NEGATIVE TEXT

David Musil
Master Degree Programme (2), FEEC BUT
E-mail: xmusil49@stud.feec.vutbr.cz

Supervised by: Lukáš Povoda
E-mail: xpovod00@phd.feec.vutbr.cz

Abstract: In the present, obtaining and sorting knowledge from data produced by various sources requires significant effort which is not ensured easily by a human, meaning machine processing is taking place. Purpose of this work was to create a system capable of positive and negative emotion detection from text along with evaluation of its performance. System allows training with use of large amount of data (known as Big Data), exploiting Spark library. Classificator model was created with use of Support Vector Machines. Highest achieved accuracy is 78,05% for Czech, 79,73% for German and 91,88% for English.

Keywords: artificial intelligence, Big Data, emotion detection, text-mining

1 ÚVOD

Dolování znalostí, zejména emocí z textových dat, je poměrně rozsáhlým oborem, těšícím se v současné době značné popularitě. Emoce mají značný dopad na lidské vztahy v mnoha aspektech každodenního života. V moderní společnosti nejen mezilidská komunikace, ale i komunikace mezi člověkem a přístrojem, tíhne k přesunu do světa počítačů a internetu. Množství textových dat drasticky stoupá s nárůstem využití současných trendů v komunikačních technologiích.

Cílem této práce je vytvoření systému, který je schopen identifikovat pozitivní a negativní emoce v jemu předloženém textu, a sice takovým způsobem, aby byla zajištěna možnost jeho trénování pomocí velkých objemů dat (Big Data, přiblíženo v článku [1]), konkrétně s využitím knihovny Spark. Jedná se o víceúčelový clusterový výpočetní framework pod open source licencí vhodný pro zpracování velkých objemů dat.

První kapitola článku představuje hlavní přínos práce a základní poznatky o problematice, druhá kapitola zahrnuje aktuální situaci v oblasti zpracování textových dat a uvádí některé podobné práce. Třetí kapitola objasňuje proces zpracování textu včetně popisu algoritmu umělé inteligence. Čtvrtá kapitola uvádí systémem dosažené výsledky, pátá kapitola nabízí rekapitulaci práce a její zhodnocení.

2 SOUVISEJÍCÍ PRÁCE

Přístup k úkolu dolování znalostí za účelem rozpoznání emocí v textu se u různých prací může značně lišit. Známou možností je vyhledávání pomocí klíčových slov, jež použili například autoři článku [2], kde se nejvyšší dosažená úspěšnost pohybovala okolo 75%. Takovýto systém detekuje emoce za pomoci seznamů slov roztříděných do specifických kategorií. Další přístup využívá umělé inteligence jako kupříkladu v článku [3], kde nejúspěšnější ze série experimentů dosáhl na úspěšnost 86% pomocí algoritmu podpůrných vektorů. V experimentu [4] autoři provádí analýzu sentimentu v datech s využitím algoritmů Naive Bayes a SVM, přičemž maximální úspěšnost se pohybuje v rozmezí 89% až 94%. V neposlední řadě existují hybridní techniky, kde jsou předešle zmíněné přístupy kombinovány. V nedávné době bylo provedeno množství výzkumů na datech z populární sociální sítě Twitter,

kupříkladu článek [5], kde je pro analýzu emocí v příspěvcích využito několik různých typů klasifikátoru včetně Naive Bayes, SVM, k-nejbližších sousedů a rozhodovacích stromů. Za zmínku dále stojí zrychlování samotných výpočtů jako u experimentu [6], kde byl výpočet urychlen algoritmem k-nejbližších sousedů a kombinací více grafických karet až 750-násobně oproti samotnému procesoru.

3 POPIS FUNKCE SYSTÉMU

Pro projekt byla použita databáze textů, ze které jsou čerpána data jak pro trénování, tak pro testování. Jedná se o komentáře uživatelů z internetových serverů, odrážející jejich subjektivní názor na konkrétní produkt nebo dílo, jde tedy o nestrukturovaný text. Každý z komentářů je celistvý text opatřený označením, zda v něm převažuje pozitivní či negativní emoce. Je možno zde nalézt vzorky ve třech jazycích, a sice v češtině, němčině a angličtině. Další rozdělení dat je podle určení pro trénování či testování. Polovina vzorků v každém jazyce je určena pro trénování, druhá pak pro testování. Pro každý z jazyků obsahuje databáze 12 000 vstupů, celkově je tedy k dispozici 36 000 záznamů. Toto množství dat je pouze zlomkem objemů, pro které má být možno systém využívat.

První částí procesu je import dat pro trénování klasifikátoru. Texty obsahují subjektivní názor uživatelů se všemi jejich náležitostmi včetně gramatických chyb, překlepů, apod. Následuje segmentace textu neboli rozčlenění vět na tokeny, sloužící jako základní nositele informace, v tomto případě jednotlivá slova věty. Vstupní text je převeden na tokeny tak, že je využito znaků oddělujících slova jako mezery, tečky, číslice a ostatní znaky, jenž nejsou písmeny, načež je veškerý nyní již rozdělený text převeden na písmena malé abecedy. Tímto krokem úprava textu pro další práci končí, není tedy aplikován lemmatizátor či stemmer, seznam nejčastějších zkratek, filtr stop slov, či slovník pro eliminaci překlepů. Tyto skutečnosti je nutno brát v úvahu při posuzování naměřených výsledků.

Aby bylo možné text dále využít, je nutné jej upravit do takové podoby, aby mohl být zpracován algoritmem strojového učení. Text je tedy reprezentován za pomoci vektorového modelu, což je zajištěno prostřednictvím výpočtu hodnoty TF-IDF [3]. Jde o metodiku pro hodnocení relevance jednotlivých slov v textu. Umožňuje slovům přiřazovat váhu na základě převrácené četnosti slova v dokumentech, pokud je tedy velký počet dokumentů ze sbírky, ve kterých se dané slovo vyskytuje, jeho důležitost se snižuje. Pro určení váhy jednotlivých složek vektoru je nejprve vypočítána četnost výskytu tokenu v dokumentu, následně je vypočtena důležitost tokenu v rámci celého korpusu, načež vynásobením těchto dvou hodnot dostáváme celkovou výslednou váhu tokenu.

V oblasti zpracování textu je zřejmě nejpoužívanější technikou algoritmus podpůrných vektorů (SVM) založený na konceptu hledání rozhodovacích rovin, který je pro klasifikátor využit. Jak již bylo zmíněno dříve, systém je postaven na projektu Spark, jež tvoří sjednocený systém určený ke zpracování dat v různých podobách. K dispozici je několik knihoven Sparku, z nichž pro tuto práci je významnou knihovna MLlib určená pro strojové učení, která obsahuje SVM algoritmus zde využitý pro vytvoření modelu klasifikátoru. Jedná se o SVM s lineárním jádrem, jehož parametry byly nastaveny podle předešle zjištěné konfigurace vhodné pro texty. Volání funkce pro vytvoření modelu obsahuje parametr počtu iterací, označující kolikrát má proces učení proběhnout. Počet iterací je podstatným prvkem se značným vlivem na úspěšnost klasifikace, jak bude dále ukázáno v prezentaci výsledků.

4 VÝSLEDKY

Úspěšnost natrénovaného modelu pro klasifikaci zobrazená v Tab. 1 je zhodnocena ověřením správnosti rozhodování modelu za pomoci testovacích dat. Jedná se o poměr počtu správně zařazených dokumentů do dané kategorie vzhledem k počtu veškerých dokumentů zařazených do dané kategorie (včetně nesprávně zařazených). Zvažované počty iterací v každém jazyku jsou v rozsahu 1 až 10 000.

Tab. 1 zobrazuje dosaženou přesnost natrénovaného modelu pro každý z jazyků při různých počtech iterací, kde úspěšnost je vyjádřena procentuálně. Lze vyčíst, že adekvátní počet iterací se pohybuje

	jazyk		
iterace	čeština	němčina	angličtina
1	67,13	67,00	73,00
5	75,92	70,33	79,00
10	78,00	78,28	91,05
100	78,05	79,73	91,88
1000	77,02	79,05	91,45
10000	75,47	77,78	88,05

Tabulka 1: Úspěšnost modelu pro jednotlivé jazyky a počty iterací

kolem hodnoty 100 iterací, a sice pro všechny 3 jazyky. Celkově nejnižší přesnosti bylo dosaženo u českého jazyka. Možným důvodem je fakt, že čeština je odlišnější od zbylých dvou jazyků více, než jsou tyto jazyky odlišné navzájem od sebe, navíc jde o bohatější jazyk se složitější větnou skladbou. Dále mohla k nižší úspěšnosti přispět skutečnost, že vzorky sice pochází z české databáze, zlomek z nich však byl vytvořen ve slovenštině. Nejvyšší přesnosti je při stejných počtech iterací jako u ostatních jazyků dosaženo pro anglický jazyk, a sice více než 90%. Přesnost se pak pro český a německý jazyk pohybovala na společné hodnotě, konkrétně téměř u hranice 80%.

5 ZÁVĚR

Článek popisuje vytvořený systém pro detekci pozitivní či negativní emoce v textu v českém, německém a anglickém jazyce. Texty použité pro trénovací množinu pocházejí z internetových diskuzí a obsahují označení převažující emoce. Z prezentace výsledků je zřejmé, že natrénovaný model byl při adekvátním počtu iterací schopen správně zařadit přibližně 3 ze 4 vzorků v českém jazyce, podobně jako u němčiny. Pro stejný počet iterací dovede model správně klasifikovat dokonce přibližně 9 z 10 předložených vzorků v anglickém jazyce. Přesnost lze považovat za relativně uspokojivou, přihlédneme-li k faktu, že na vstupním textu byla provedena pouze tokenizace, podoba slov tedy byla zachována v jejich původním tvaru, ve které byla autorem příspěvku vytvořena.

REFERENCE

[1] POWER, D.J. Using 'Big Data' for analytics and decision support. *Journal of Decision Systems, 23:2, s. 222-228, 2014*. DOI: 10.1080/12460125.2014.888848.

[2] CHUANG, Z.J., WU, C.H. Multi-modal emotion recognition from speech and text. In *The Association for Computational Linguistics and Chinese Language Processing*, vol. 9, s. 45-62, 2004.

[3] DE SILVA, J., HADDELA, P.S. A term weighting method for identifying emotions from text content. *Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on Industrial and Information systems*. ISBN 978-1-4799-0908-7.

[4] TRIPATHY, A., AGRAWAL, A., RATH, S.K. Classification of Sentimental Reviews Using Machine Learning Techniques In *Procedia Computer Science*, vol. 57, 2015, pp. 821–829.

[5] LIMA, A.C.E.S, DE CASTRO, L.N., CORCHADO, J.M. A polarity analysis framework for Twitter messages, In *Applied Mathematics and Computation*, vol. 270, 2015, pp. 756–767.

[6] MAŠEK, J., BURGET, R., KARÁSEK, J., UHER, V., DUTTA, M.K. Multi-GPU Implementation of k-Nearest Neighbor Algorithm. In *2014 37th International Conference on Telecommunications and Signal Processing (TSP)*. Berlin, Germany: 2014. s. 652-655. ISBN: 978-80-2144983-1.