Aplikace pro odhad volebních preferencí osob na základě veřejných dat

Marek Zábran[1], Adam Mištera[2]

1 Úvod

V České Republice je možné používat internet již třicet let . Za tuto dobu se výrazně změnil způsob, jakým je internet vnímaný a zvláště pak vztah lidí k němu. Z původně pro běžného člověka téměř neznámého konceptu se stala každodenní nutnost ke komunikaci a získávání informací ze zbytku světa - tato skutečnost je zvláště jasné zvláště v rámci posledního roku za krize pandemie *Covid-19*, kdy internet často představuje jediné spojení i s naším blízkým okolím.

Největší rozdíl v chování na internetu je patrně v naší ochotě sdílet informace. Zatímco internetové generaci bylo vštěpováno, že nemají na internetu sdílet své osobní informace, dnes většina osob [ČSÚ (2020)] ochotně sdílí pomocí sociálních sítí všechny různé údaje, včetně svých životních příběhů.

Shromažďování a sdílení cizích osobních informací bylo sice výrazně omezeno z iniciativy EU v rámci GDPR, neznamená to ovšem, že by se stalo nemožným. Například informace o trvalém bydlišti cizích osob již není možné za běžných okolností dohledat, ani si vyžádat od příslušného úřadu, je ovšem stále možné vyhledat vlastněné nemovitosti (v ČÚZK (2021)) a z nich odhadnout trvalé bydliště.

Vzhledem k objemu osobních informací, které jsou na internetu veřejně dostupné, si můžeme klást otázku, co všechno je možné zjistit o průměrné osobě a jestli nehrozí zneužití těchto informací, pokud by se je někomu podařilo sjednotit na naklíčovat na jednotlivé osoby. Systém, který právě toto do určité míry dokázal, patrně vyvinula například *Cambridge Analytica*, s tím rozdílem, že využila špatně ošetřených pomocných systémů sociální sítě *Facebook* a dostala se tak i k soukromým (neveřejným) informacím uživatelů *Facebook*. Tato data byla použita k ovlivnění voleb v USA v roce 2017 (dokládá Jungherr (2020)) a možná i v rámci jiných politických událostí.

Můžeme se domnívat, že velké množství organizací a osob by mělo o podobný systém zájem a je tedy vhodné vyzkoušet, jak přesný takový systém může být (pro potřeby bezpečnosti). Tento projekt se zabývá právě tímto problémem: V rámci projektu byla vytvořena aplikace, která dokáže stáhnout veřejná data o hledané osobě a odhadnout její volební preference. Aplikace tak funguje jako *proof of concept* možnosti vytvoření dobrého odhadu o volebních preferencí osoby na základě základních veřejně dostupných informací[1].

[1] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, specializace Softwarový inženýr, e-mail: zabran@students.zcu.cz

[2] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, specializace Softwarový inženýr, e-mail: amistera@students.zcu.cz

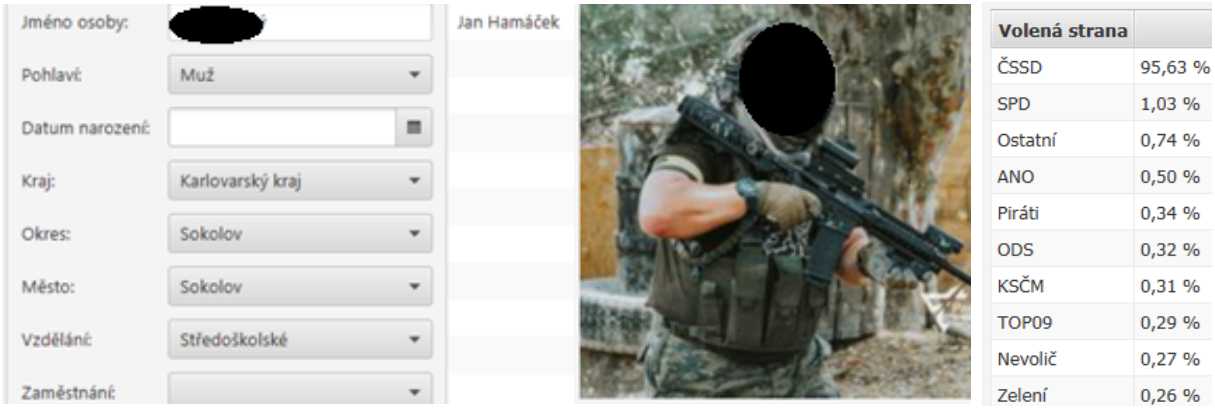
[1]Zatím pouze za předpokladu, že tyto informace jsou k dispozici

2 Aplikace

Aplikace doposud vytvořená v rámci tohoto projektu dokáže stáhnout pomocí crawleru [2] veřejná data dané osoby ze sociální sítě *Facebook* a některé statistické údaje z *Českého statistického úřadu* a následně z nich odhadnout volební preference. K tomuto odhadu se používá mírně upravený *Naivní Bayesův klasifikátor* (např. Webb (2011)) s následujícím vzorcem:

$$P_{Posterior}(A) = e^{(1-n)*log(P(A)+c)+\sum_{i=1}^{n}(log(P(Attribute_i)+c)+log(P(A|Attribute_i)+c))}, \quad (1)$$

Kde: *A* je politická strana. *n* je počet vlastností dané osoby. *c* je bezpečnostní konstanta (= 1e-6). *Attribute* je množina vlastností dané osoby.



Obrázek 1: Obrázek osobních údajů a vygenerované volební preferenci k nim.

3 Závěr

Aplikace doposud nebyla testována na větší množině dobrovolníků, při větším části dodaných informací, zvláště pak tzv. *lajcích* u facebookových stránek politických stran, dodává ovšem aplikace velmi přesvědčivé výsledky, zvláště pak u osob do třiceti let. Přestože pro přesné zhodnocení je stále nutné zvětšit testovací vzorek, aplikace již nyní dobře uvozuje problém možnosti odhadu neznámých osobních informací na základě informací a dat veřejně dostupných. Tato skutečnost je varující zvláště proto, že aplikace používá jen velmi malé množství údajů a triviální mechanizmus pro odhad nových skutečností.

Literatura

Český úřad zeměměřičský a katastrální (2021). *Nahlížení do katastru nemovitostí*. Available from: `https://nahlizenidokn.cuzk.cz/` [Accessed 3rd June 2021].

Český statistický úřad (2020). *Využívání ICT v domácnostech a mezi jednotlivci - 2020*. Available from: `https://tinyurl.com/4rzmpj3r` [Accessed 3rd June 2021].

Jungherr, A., Rivero, G., Gayo-Avello, D. (2020). *Retooling Politics: How Digital Media Are Shaping Democracy*. Cambridge, Cambridge University Press.

Webb G.I. (2011) Naïve Bayes. *Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning*. Springer, Boston, MA.

[2] *Facebook* umožňuje aplikacím přístup k uživatelským datům pouze v případě, že to uživatel aplikaci povolí, proto je crawler nutný. A i vůči crawleru se chová poměrně nepřátelsky.