IDENTIFIKÁCIA SPRÁV TYPU „FAKE NEWS“ Z TEXTOV V SLOVENSKOM JAZYKU

Benedikt Pecuch Školiteľ: doc. PaedDr. Jozef Kapusta, PhD.

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

Cieľ:

Vytvoriť návrh metódy identifikácie falošných správ. V práci budú extrahované základné charakteristiky správ a budú vybrané vhodné metódy z dostupných metód strojového učenia. Práca bude zameraná na identifikáciu falošných správ v slovenských textoch, pričom bude potrebné vyriešiť niekoľko problémov typických pre flektívne typy jazykov, do skupiny ktorých patrí aj slovenčina.

Dataset:

Rozhodli sme sa zozbierať dataset článkov, ktoré sa zaoberajú témou COVID-19. Pri rozhodovaní, či sa jedná o falošné alebo pravdivé články nám pomohlo najmä hodnotenie zdroja článku od občianskeho združenia *Košpirátori.sk*. Rozhodli sme sa používať slovné druhy a triedy pre trénovanie rozhodovacích modelov. Na obrázku môžeme vidieť relatívnu početnosť niektorých slovných druhov v článkoch, ako napríklad A = Adjektívum (prídavné meno).

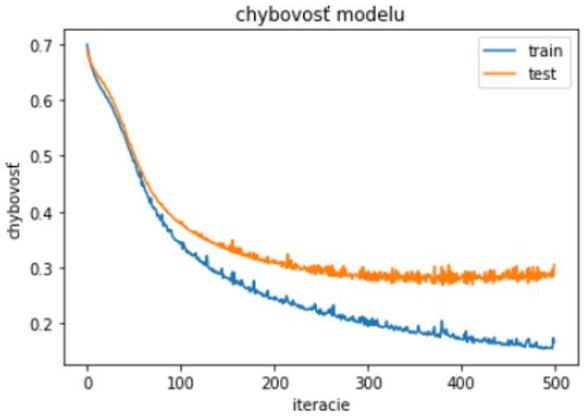
row	fake_index	R	J	A	O	E
145	0	0.027855	0.0	0.144847	0.052925	0.114206
146	0	0.004878	0.0	0.092683	0.039024	0.131707
147	0	0.019417	0.0	0.095700	0.063800	0.130374
148	0	0.013699	0.0	0.091324	0.041096	0.086758
149	0	0.017544	0.0	0.098246	0.049123	0.140351

Rozhodovacie stromy:

Pri použití rozhodovacích stromov naša detekcia dosiahla úspešnosť o výške približne 80 %, čo považujeme sa veľmi dobrý výsledok. Jeden z dôvodov, prečo sa nám podarilo dosiahnuť lepšie výsledky môže byť aj fakt, že v bežnej praxi sa pri predspracovaní datasetu odstraňujú stop slová. K týmto slovám patria predložky, spojky a neplnovýznamové slová, ktoré sme ale v našej práci neodstraňovali a práve rozhodovacie stromy niektorým týmto slovám pridelili vysokú dôležitosť. Predpokladáme, že vymazaním niektorých stop slov a to najmä zámen (ty, ja, vy, ten, tento, ktorý), by sme mohli vymazať aj typické črty, ktoré sú charakteristické pre falošné správy.

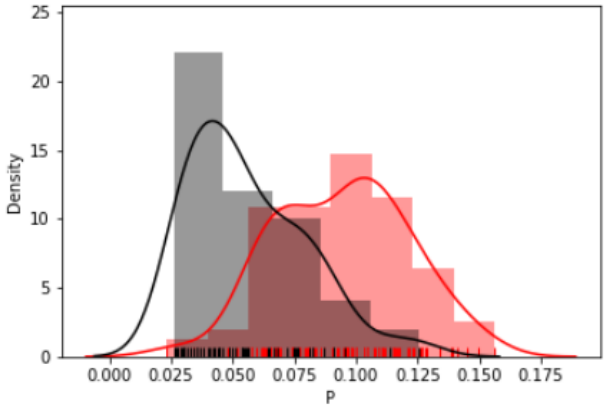
Neurónové siete:

S použitím druhého klasifikátora - neurónových sietí, sme očakávali vyššiu úspešnosť detekcie falošných správ. Získané výsledky našej implementácie to potvrdzujú. Zatiaľ čo úspešnosť za použitia rozhodovacích stromov sa pohybovala okolo 80 %, pri neurónových sieťach sa nám podarilo dosiahnuť úspešnosť výsledkov vyše 90 %. Na druhej strane je ale tiež dôležité pripomenúť, že neurónová sieť má aj svoju nevýhodu, ktorou je značne vyššia výpočtová náročnosť. Pri našej implementácií sme to pocítili hlavne pri tom, ako sa učenie neurónovej siete iterovalo päťsto krát.



Závery práce:

Pre falošné správy je typickejší vysoký počet zámen, čo sa napríklad dá pozorovať aj na grafe početnosti zámen. Falošné správy sme vyznačili červenou farbou a čiernou farbou sú vyznačené správy pravdivé. Na tomto grafe môžeme pozorovať, že hodnoty početnosti zámen falošných správ sa pohybujú v priemere okolo hodnoty 0,11. V pravdivých správach sa tieto hodnoty pohybujú okolo čísla 0,04 .To nám vysvetľuje, prečo v našich výsledkoch boli zámená v rozhodovacích stromoch najdôležitejším faktorom pri detekcii falošných správ. Preto pri porovnaní súboru falošných správ zo súborom pravdivých správ, sú tieto súbory od seba zreteľne rozlíšiteľné počtom zámen, tak ako to vidíme na grafe.



Na druhej strane keď sa pozrieme na graf početnosti prísloviek rozdiel medzi hodnotami falošných a pravdivých správ nie je značne viditeľný. Fakt, že tieto hodnoty sú si dosť podobné vysvetľuje, prečo rozhodovací strom pridelil nulovú dôležitosť početnosti prísloviek.