Klasifikácia YouTube komentárovpomocou neurónovej siete

Maroš Kopec

Masarykova Univerzita
487595@muni.cz

30. júna 2019

I. Úvod

Klasifikácia spamu je problémom, ktorému sa venuje množstvo odborných článkov a prác. Za spam sa považuje nevyžiadaná správa rozosielaná veľkému počtu adresátov alebo rozosielaná na mnoho miest, zväčša za účelom reklamy. Pre množstvo internetových stránok je práve spam veľkým problémom, pretože ich používatelia sú zahltený obsahom, ktorý im znepríjemňuje skúsenosť s ich obsahom. Konkrétne YouTube je terčom veľkého množstva spamu od používateľov, ktorí sa týmto spôsobom snažia zviditelniť vlastné video či kanál.

II. Súvisiace práce

Táto práca je inšpirovaná prácou TubeSpam[1] z Federálnej Univerzity Sao Carlos. V tejto práci akademici porovnávali niekoľko metód klasifikácie spamu, menovite rozhodovacie stromy, K-najbližších susedov, logistickú regresiu, Bernoulliho naivný Bayes, Gaussov naivný Bayes, Multinomiálny naivný Bayes, náhodné lesy, Support vector machines s lineárnym kernelom, Support vector machines s polynomiálnym kernelom a Support vector machines s Gaussovým kernelom. Práca dopĺňa vyššie zmienenú štúdiu o modely postavené na rekurentných neurónových sieťach.

V pôvodnej práci autori klasifikovali spam oddelene pre každé video. Neurónová sieť potrebuje veľké množstvo záznamov, aby bola schopná naučiť sa rozpoznávať spam. Dataset však obsahuje len obmedzený počet záznamov. Preto sa dataset spojil a všetky experimenty sa vykonávali nad zmiešanými záznamami.

Tabuľka 1: *Kompozícia datasetu*

Dataset	YouTube ID	# Spam	# Ham	Total
Psy	9bZkp7q19f0	175	175	350
KatyPerry	CevxZvSJLk8	175	175	350
LMFAO	KQ6zr6kCPj8	236	202	438
Eminem	uelHwf8o7_U	245	203	448
Shakira	pRpeEdMmmQ0	174	196	370

III. Dataset

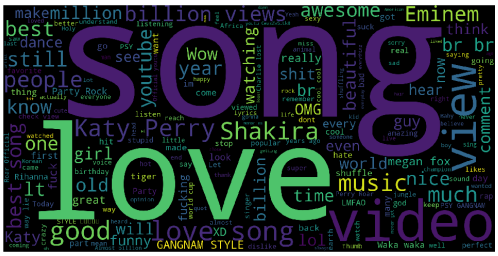
i. Dáta

Pre trénovanie a testovanie modelov sú použité rovnaké dáta ako boli použité v práci TubeSpam [1]. Dataset obsahuje 1956 záznamov z 5 najpozeranejších YouTube videí vo formáte csv. Kompozícia datasetu je znázornená v tabuľke 1. Pre potreby experimentu bol dataset rozšírený o popis videa.

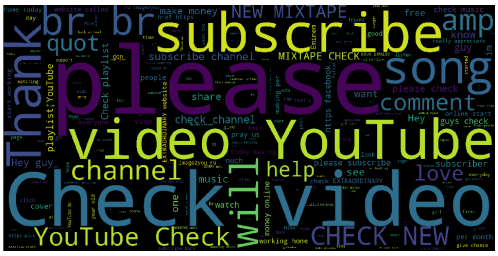
Pre predstavu obsahu záznamov boli vygenerované slovné mapy, ktoré sú zobrazené na obrázkoch 1 a 2. Nad záznamami bol vykonaný experiment, pri ktorom bolo náhodne vybraných 100 záznamov z videa od autora Eminem, ktoré boli znovu označované. Z týchto

[1] https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

sto záznamov bolo 8.9% označených odlišne. Z tohto malého experimentu je možné vyhodnotiť, že označenie 100% spamu je nepravdepodobné, keďže sa to nepodarilo ani človeku.



Obr. 1: *Slovná mapa hamu*



Obr. 2: *Slovná mapa spamu*

ii. Vstupy

Vstupné dáta do rekurentnej siete sú komentáre s binárnym označením spamu a popis videa. Spracovanie datasetu spočíva v načítaní dát z CSV súboru a uloženie v dátovom rámci pandas[2]. Následne sú oddelené datasety pre trénovanie a testovanie s pomerom 9 : 1. Čo znamená, že na 90% sa bude sieť trénovať a na zvyšných 10% testovať. Z datasetu sa vyberú len podstatné stĺpce, a to Content - samotný komentár, Class - označenie, či ide o spam alebo nie, Description - popis videa. Spravocanie komentárov a popisu videa je ďalej rovnaký. Popis a komentáre sa vektorizujú vytvorením vnútorého slovíku pre popis aj pre komentáre reprezentujúce výskyt každého slova. Túto reprezentáciu prevedieme z interného slovníka do jednorozmerného vektoru, teda na sekvenciu integerov. Tie sú potom zarovnané podľa najdlhšieho koemntára do dvojrozmernej matice s rozmermi (n, max_n), kde n je počet vstupov (v tomto prípade komentárov), max_n je dĺžka najdlšieho komentára.

Ukážky vstupnych dát sa nachádzajú v súbore *data_representation.txt*.

iii. Výstupy

Výstupom rekurentnej siete je pravdepodobnosť, $p \in (0, 1)$, že daný komentár je spam.

IV. Modely

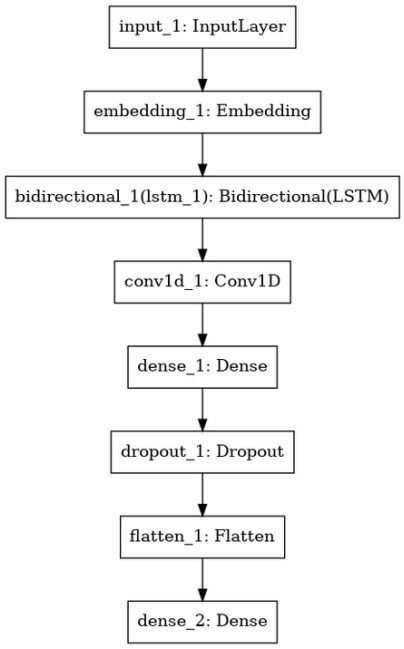
Rozmery jednotlivých vrstiev sú v súbore *layers_shape.txt*.

i. Model č.1

Prvý implementovaný model, zobrazený na obrázku 3, číta na vstupe komentáre, ktoré transformuje na sekvencie čísel reprezentujúce jednotlivé slová. Ako bolo vysvetlené v podsekcii ii Vstupy, vstupom je dvojrozmerná matica. Následne je každému slovu priradená váha na základe predpočítaného embedding glove.6B zo Standfordskej Univerzity [3]. Obojsmerná vrstva LSTM hľadá vzdialené súvislosti, v ktorých následne konvolučná vrstva hľadá súvislosti blízko seba. Dense vrstva spája každý uzol konvolučnej vrsty s dropout vrstvou. Posledná zmienená vrstva zabezpečuje, že sa model učí stále na trochu iných dátach. Flatten vrstva upraví tvar pre poslednú dense vrstvu, ktorej výstup je buď 0 alebo 1, čo interpretujeme ako ham a spam v rovnakom poradí.

[2] https://pandas.pydata.org/

[3] https://nlp.stanford.edu/projects/glove/



Obr. 3: *Model č.1*

ii. Model č.2

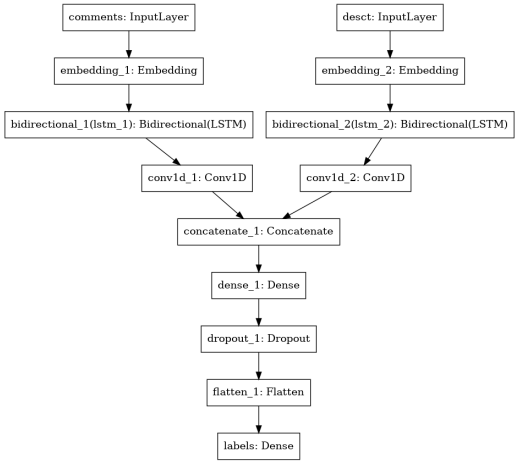
Pre nasledujúci experiment bol k modelu pridaný ďalší vstup - popis videa. Analogicky podľa podsekcie ii Vstupy, vstupom sú dve dvojrozmerné matice. Grafické znázornenie modelu je vyobrazené na obrázku 4. Nový model sa od prvého líši len rozdvojením časti siete. Popis videa je spracovávaný rovnako ako aj komentáre a reprezentácie oboch vstupov z konvolučnej siete sú konkatované do dense vrstvy.

iii. Model č.3

Ide o model č.2 obohatený o vrstvu LSTM po konkatovaní výsledkov zo sietí spracuvajúcich komentáre a popis videa. Je znázornený na pobrázku č.5.

V. Výsledky

Výsledky k prvému navrhnutému modelu 3, so vstupom len pre komentáre, sa kvôli technickej chybe nepodarilo zachovať. K dispozícií sú len dáta z oplimalizácie hyperparametrov. Tieto



Obr. 4: *Model č.2*

výsledky boli zohľadnené pri experimentoch s druhým modelom 4.

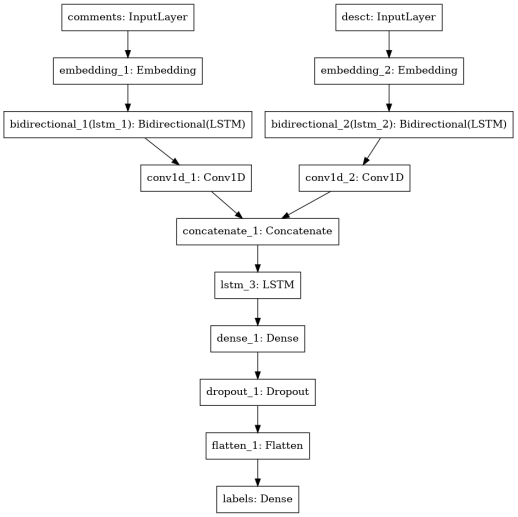
i. Metódy evaluácie

Pre zhodnotenie modelu boli použité štatistické metriky *presnosť*, *chytený spam*, *blokovaný ham*, *F-measure*, *Matthews korrelačný koeficient*, ktoré boli použité aj pri evaluácii vo vyššie spomínanej práci TubeSpam[1]. Presnosť vyjadruje koľko percent spamu klasifikátor úspešne označil za spam. Chytený spam a blokovaný ham udávajú percentuálny podiel označeného spamu k neoznačenému a ham komentáre označené za spam. Najlepšie výsledky zobrazené v tabuľke 2 sa podarilo dosiahnuť s parametrami vymenovanými nižšie.

- Dropout: 0.3
- Activačná funkcia: elu
- Počet neurónov Dense vrsty: 15
- Kernel inicializácia: normal
- Posledná aktivačná funkcia: sigmoid
- Optimizér: Nadam
- Počet epoch: 12

Vysvetlivky pre tabuľky 2,3 a 4: acc - presnosť [%], sc - chytený spam [%], hb - blokovaný ham [%], F1 - F-measure, MCC - Matthews korrelačný koeficient.

Zatiaľ čo presnosť je pomerne dostačujúca, počet chybne označeného hamu za spam je



Obr. 5: *Model č.3*

Tabuľka 2: *Najlepšie výsledky modelu č.2*

acc	sc	hb	F1	MCC
69.9%	84.7%	49.4%	0.82	0.008

alarmujúco vysoký, až takmer 50%. Keďže žiaden z experimentov nedosiahol oveľa lepšie výsledky môžeme z toho vyvodiť, že žiaden z experimentov vhodne nerieši problém klasifikácie spamu.

Naproti tomu práca TubeSpam[1] dosiahla oveľa lepšie výsledky, zobrazené v tubuľke 4. Táto skutočnosť môže byť spôsobená nedostatkom vstupných dát pre učenie rekurentnej neurónovej siete. Model sa nebol schopný naučiť rozoznávať spam na takmer 2000 záznamoch. Predpoklám, že ak by sa vstupné dáta zvýšili 100-násobne (t.j. aspoň na 200000), boli by výsledky oveľa lepšie. Ďalším faktorom, ktorý mohol negatívne ovplyvniť výsledky môže byť vstup krátkeho charakteru. Komentáre sú často len krátke pár slovné heslá. Pre túto úlohu by mohlo byť vhodnejšie spracuvávať vstup nie po slovách ale po znakoch. To má však opäť rovnaké limitácie vo forme malého datasetu.

Tabuľka 3: *Najlepšie výsledky modelu č.š*

acc	sc	hb	F1	MCC
62.2%	80.85%	54.9%	0.69	-0.077

Tabuľka 4: *Najlepšie výsledky práce TubeSpam*

acc	sc	hb	F1	MCC
97.73%	95.77%	0.00%	0.978	0.955

VI. Príloha

i. Návod na spustenie

1. Je potrebné stiahnuť Dataset [4], Embedding [5].
2. Nainštalujte si nástroj:

   ```
   pipenv
   ```

3. Spustite príkaz:

   ```
   # pipenv install
   ```

4. Pre spustenie trénovania modelu použite príkaz:

   ```
   # python classificator.py
   ```

   Kde Number je číslo experimentu a meno zložky výstupov.

Literatúra

[1] Alberto, T.C., Lochter J.V., Almeida, T.A. *TubeSpam: Comment Spam Filtering on YouTube*. Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15), 1-6, Miami, FL, USA, December, 2015. (preprint)

[4] https://archive.ics.uci.edu/ml/machine-learning-databases/00380/

[5] https://nlp.stanford.edu/projects/glove/