UNIVERSITATEA ”ALEXANDRU-IOAN CUZA” DIN IAȘI

FACULTATEA DE INFORMATICĂ

LUCRARE DE LICENȚĂ

Antrenarea unui model pentru detectarea efectelor adverse ale medicamentelor in mediul online

propusă de

Dragoș-Constantin Ciocan

Sesiunea: iulie, 2020

Coordonator științific

Conf. Dr. Mădălina Răschip

Avizat,
Îndrumător lucrare de licență,
Conf. Dr. Mădălina Răschip.
Data: ..... Semnătura: .....

Declarație privind originalitatea conținutului lucrării de licență

Subsemnatul **Ciocan Dragoș-Constantin** domiciliat în **România, jud. Vaslui, orașul Negrești, strada Casa Apelor, nr. 15A**, născut la data de **11 octombrie 1998**, identificat prin CNP **1981011375024**, absolvent al Facultății de informatică, **Facultatea de informatică** specializarea **informatică**, promoția 2020, declar pe propria răspundere cunoscând consecințele falsului în declarații în sensul art. 326 din Noul Cod Penal și dispozițiile Legii Educației Naționale nr. 1/2011 art. 143 al. 4 și 5 referitoare la plagiat, că lucrarea de licență cu titlul **Antrenarea unui model pentru detectarea efectelor adverse ale medicamentelor in mediul online** elaborată sub îndrumarea domnului **Conf. Dr. Mădălina Răschip**, pe care urmează să o susțin în fața comisiei este originală, îmi aparține și îmi asum conținutul său în întregime.

De asemenea, declar că sunt de acord ca lucrarea mea de licență să fie verificată prin orice modalitate legală pentru confirmarea originalității, consimțind inclusiv la introducerea conținutului ei într-o bază de date în acest scop.

Am luat la cunoștință despre faptul că este interzisă comercializarea de lucrări științifice în vederea facilitării falsificării de către cumpărător a calității de autor al unei lucrări de licență, de diplomă sau de disertație și în acest sens, declar pe proprie răspundere că lucrarea de față nu a fost copiată ci reprezintă rodul cercetării pe care am întreprins-o.

Data: ..... Semnătura: .....

Declarație de consimțământ

Prin prezenta declar că sunt de acord ca lucrarea de licență cu titlul **Antrenarea unui model pentru detectarea efectelor adverse ale medicamentelor in mediul online**, codul sursă al programelor și celelalte conținuturi (grafice, multimedia, date de test, etc.) care însoțesc această lucrare să fie utilizate în cadrul Facultății de informatică.

De asemenea, sunt de acord ca Facultatea de informatică de la Universitatea "Alexandru-Ioan Cuza" din Iași, să utilizeze, modifice, reproducă și să distribuie în scopuri necomerciale programele-calculator, format executabil și sursă, realizate de mine în cadrul prezentei lucrări de licență.

Absolvent **Dragoș-Constantin Ciocan**

Data: ..... Semnătura: .....

Cuprins

Motivație

În anul 2017, în cadrul Workshop-ului AMIA-2017 despre Exploatarea Rețelelor de Socializare pentru Aplicații în Sănătate, a fost propusă, printre altele, următoarea temă: depistarea efectelor adverse ale medicamentelor lansate pe piață, din postările consumatorilor acestora pe rețele de socializare. M-am decis ca aceasta este și tema mea de licență și voi încerca să antrenez un model care să îndeplinească această sarcină. Motivul pentru care

Introducere

Problema reacțiilor adverse ale medicamentelor este una de mare notorietate în lumea medicinală. Cei care au cel mai mult de suferit și de pierdut de pe urma acestei probleme sunt consumatorii acestora, aceștia putându-și pune viața în pericol în anumite cazuri (prescripții greșite, nerespectarea prescripției), însă afectate sunt și companiile, producătorii și toți cei care lucrează în sfera farmaceuticelor.

Reacțiile adverse ale medicamentelor înseamna orice simptome neprevăzute, malicioase pe care pacientul le simte după un anumit interval de timp de la ingerarea medicamentului. Pentru a se preveni această problemă, se efectuează seturi de teste, pe pacienți care se înscriu voluntar la această acțiune, însă, există cazuri în care aceste teste nu sunt suficiente, iar la momentul lansării medicamentului, oamenii să experimenteze reacții adverse. Acest fapt este bine cunoscut de toată lumea, de aceea se colectează mereu date despre efectele medicamentelor și după lansare. Colectarea se dovedește a fi destul de minuțioasă, deoarece nu toți pacienții vorbesc cu doctorii lor despre aceste lucruri, dar și din alte motive. Ideea este ca nu există o metodă standardizată pentru colectarea datelor referitoare la efectele acestor medicamente.

S-a observat că unii pacienți preferă să-și exprime nemulțumirea, alături de efectele adverse ale medicamentelor, în mediul online. Dar, desigur că și de aici este anevoios de extras anumite informații, din cauza unor exprimări precare, utilizarea argourilor, etc. La Workshop-ul AMIA-2017, prin intermediul unui concurs, s-a încercat rezolvarea acestei ramuri a problemei. La concurs au putut participa echipe din toată lumea, iar în total au fost desemnate trei probleme. Eu în această lucrare mă voi axa doar pe una, cea descrisă mai sus.

Pentru rezolvarea acestei probleme, am încercat antrenarea unui model care să aibă în final un scor cât mai mare la testare, inspirându-mă din lucrarea de aici. Am încercat prin două metode de transformare a textului într-un format ușor de înțeles de către calculator (transformând în vectori de numere), acestea fiind n-grame-ele, care

s-au dovedit a nu da un scor foarte bun, și transformarea word2vec, combinată cu un algoritm SVM din biblioteca sci-kit learn, care a dat și rezultatele cele mai bune.

Contribuții

Chapter 1

Setul de date

Setul de date pentru această sarcină a fost preluat din cadrul unui proiect de detectare a efectelor adverse ale medicamentelor de către DIEGO Lab, din cadrul universtății "Arizona State University". Acestea au fost preluate făcându-se căutări pe baza numelor de medicamente, după care niște experți în domeniul farmaceutic le-au clasificat în cele două categorii (conțin sau nu conțin reacții adverse ale medicamentelor).

Tweet-urile au fost furnizate pe bază de ID-uri, ele trebuind descărcate utilizând un script în python 2.*. Lucrând în python 3.* a trebuit să-mi creez eu acest script (pe baza celui dat). Descărcarea acestora a durat aproximativ o oră, iar din această cauză le-am salvat local, nemaifiind nevoie să-l rulez din nou. Din cele 15667 de ID-uri primite, doar 9257 au rămas valabile. Din acestea 9257, am folosit 70% din acestea pentru setul de antrenament, iar restul de 30% pentrul setul de test.

O mare problemă cu acest set de date este faptul că este o mare diferență dintre numarul de date de clasă 0 si numărul de date de clasă 1 (datele sunt nebalansate).

Chapter 2

Preprocesarea limbajului

Primul pas pe care l-am folosit pentru antrenarea modelului a fost preprocesarea textului. Deoarece textele noastre, ca date de antrenare și de testare, sunt postări de pe Twitter, ele vor avea un limbaj informal, cuvinte prescurtate, argouri, ceea ce face antrenarea modelului nostru dificilă. Ca urmare a acestui fapt, am preprocesat textul pentru a-l aduce într-o formă relativ mai ușoară de înțeles pentru calculator.

Primul pas, alături de cel de-al doilea, au fost inspirați din lucrarea aceasta, aceștia constând în înlocuirea unor termeni specifici (medicamentele și reacțiile adverse) cu un simbol, deoarece acestea nu sunt relevante în clasificarea tweet-urilor. Am folosit o listă de medicamente și de reacții adverse de pe situl celor de la Diego Lab. Am înlocuit medicamentele cu simbolul ”MED”, iar reacțiile adverse cu ”ADR”.

Următorul pas, cel de-al treilea, a fost înlocuirea cuvintelor prescurtate, de tipul ”i’m”, ”you’re”, în cuvinte întregi, pentru a restrânge numărul cuvintelor totale, dar și pentru a le da un sens mai puternic acestora.

Al patrulea pas, unul care nu este important de unul singur, dar prinde valoare datorită următorilor pași, este cel de înlocuirea url-urilor cu un simbol specific. Url-urile având caractere speciale (”:”, ”/”) care trebuie și ele eliminate din text, se vor crea cuvinte nefolositoare, de exemplu ”http”. Simbolul folosit pentru acestea este ”LINK”.

Al cincilea pas constă în înlocuirea referințelor către alte persoane, de forma ”@cineva”, într-un simbol specific (”REF”), deoarece acestea conțin diferite nume, care nu au nicio importanță.

Ultimul pas cuprinde eliminarea din text a majorității caracterelor non-alfanumerice pentru a mai aerisi textul și a despărți unele cuvinte de acestea, existând situații când aceste caractere, fiind lipite de cuvinte, construiesc altele noi.

Chapter 3

Procesarea limbajului

Diam sit amet nisl suscipit adipiscing bibendum. Aliquet lectus proin nibh nisl condimentum id. Urna duis convallis convallis tellus id interdum velit laoreet. Amet tellus cras adipiscing enim eu turpis egestas pretium aenean. Tortor condimentum lacinia quis vel eros donec ac odio tempor. Volutpat ac tincidunt vitae semper. Urna cursus eget nunc scelerisque viverra mauris in aliquam. Aliquam id diam maecenas ultricies. Molestie a iaculis at erat. Tincidunt nunc pulvinar sapien et ligula ullamcorper malesuada proin. Consequat interdum varius sit amet. Eget est lorem ipsum dolor sit amet consectetur adipiscing. Pharetra diam sit amet nisl suscipit adipiscing bibendum. Maecenas sed enim ut sem viverra aliquet eget sit. Enim blandit volutpat maecenas volutpat blandit aliquam etiam erat velit.

3.1 Unigrame

Id donec ultrices tincidunt arcu non sodales neque. Integer eget aliquet nibh praesent. Euismod in pellentesque massa placerat duis ultricies lacus sed. Mauris ultrices eros in cursus turpis massa. Integer quis auctor elit sed vulputate mi. Nibh ipsum consequat nisl vel pretium lectus quam id leo. Vel elit scelerisque mauris pellentesque pulvinar pellentesque. Suscipit tellus mauris a diam maecenas. Ultrices eros in cursus turpis massa tincidunt. Tristique senectus et netus et malesuada fames ac turpis egestas. Suspendisse interdum consectetur libero id faucibus nisl tincidunt eget. Sed risus pretium quam vulputate dignissim suspendisse in. Donec adipiscing tristique risus nec feugiat in fermentum posuere. A lacus vestibulum sed arcu non odio euismod lacinia at.

3.2 Bigrame

3.3 Word2Vec

Pellentesque pulvinar pellentesque habitant morbi tristique senectus et. Ornare suspendisse sed nisi lacus sed viverra tellus in hac. Non sodales neque sodales ut etiam sit. In hendrerit gravida rutrum quisque non. Diam quam nulla porttitor massa id neque aliquam. Diam sit amet nisl suscipit adipiscing bibendum est ultricies integer. Cras fermentum odio eu feugiat pretium nibh ipsum. Egestas integer eget aliquet nibh praesent tristique magna. Porttitor eget dolor morbi non arcu risus quis varius quam. Gravida rutrum quisque non tellus orci. Diam volutpat commodo sed egestas egestas.

Concluzii

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Nunc mattis enim ut tellus elementum sagittis vitae et. Placerat in egestas erat imperdiet sed euismod. Urna id volutpat lacus laoreet non curabitur gravida. Blandit turpis cursus in hac habitasse platea. Eget nunc lobortis mattis aliquam faucibus. Est pellentesque elit ullamcorper dignissim cras tincidunt lobortis feugiat. Viverra maecenas accumsan lacus vel facilisis volutpat est. Non odio euismod lacinia at quis risus sed vulputate odio. Consequat ac felis donec et odio pellentesque diam volutpat commodo. Etiam sit amet nisl purus in. Tortor condimentum lacinia quis vel eros donec. Phasellus egestas tellus rutrum tellus pellentesque eu tincidunt. Aliquam id diam maecenas ultricies mi eget mauris pharetra. Enim eu turpis egestas pretium.

Bibliografie

- Author1, *Book1*, 2018
- Author2, *Boook2*, 2017