UNIVERSITATEA DIN BUCUREȘTI

Facultatea de Matematică și Informatică

Departamentul de Informatică

Contribuții la Sistemele de Recomandare

(Rezumat)

Conducător Științific:

Prof. Dr. Denis Enăchescu

Student Doctorand:

Andreea Salinca

București, Septembrie 2017

1. Introducere

Numărul mare de informații disponibile pe World Wide Web precum și numărul utilizatorilor de Internet au avut o creștere explozivă în ultimii zece ani. În plus, an de an este înregistrată o creștere semnificativă a acestora. Astfel, accesarea informațiilor relevante pe Internet a devenit o provocare, deoarece capacitatea umană de a filtra informații este depășită. Sistemele de redare a informațiilor (*information retrieval systems*) au rezolvat parțial această problemă, dar fără a răspunde nevoii de personalizare și ordonare în funcție de preferințele utilizatorilor. Cercetarea privind sistemele de recomandare își are rădăcini în domeniul științelor cognitive *(cognitive science*), sistemelor de redare a informațiilor, teoria aproximării și modelarea de marketing (*marketing modeling*) (Adomavicius & Tuzhilin, 2005).

În ultimii ani, sistemele de recomandare au schimbat viața oamenilor vis-a-vis de modul în care găsesc informații și produse sau comunică cu alte persoane. Sistemele de recomandare reprezintă un set de tehnici avansate care au capacitatea de a anticipa preferințele unui anumit utilizator. Ele sunt utilizate pentru a sugera elemente relevante într-un proces decizional precum: ce produs să cumpere, ce film să vizioneze, ce muzică să asculte, ce cărți sau ce știri să citească, ce restaurant să aleagă sau sugestii legate de călătorii. Aceste sisteme sunt proiectate ținând cont de preferințele, istoricul, profilul, trăsăturile individuale sau informațiile demografice ale utilizatorului (Ricci, Rokach, & Shapira, 2011).

În primul rând, în luarea unei decizii alegem să cerem opinia familiei sau a prietenilor noștri. Deoarece opinia persoanelor apropiate nouă este de cele mai multe ori subiectivă, considerăm că putem fi influențați în a obține sugestii utile și eficiente. Astfel, căutăm sa aflăm și alte opinii. Sistemele de recomandare sunt destinate persoanelor care nu dispun de suficiente informații, cunoștințe sau competențe în evaluarea opțiunilor alternative (Resnick & Varian, 1997).

În comunitatea științifică problema accesului la informațiile relevante este abordată prin dezvoltarea de tehnici și algoritmi noi care să ofere recomandări personalizate și de înaltă calitate. A devenit din ce în ce mai important ca oamenii să aibă acces la informații relevante care țin cont funcție de preferințele și gusturile acestora, în special în contextul unor seturi mari de date. Același principiu este valabil și pentru companii, acestea trebuie să furnizeze utilizatorilor informații clare și relevante pentru a-și maximiza veniturile și profitul.

Cu toate că în decursul ultimii ani o serie de abordări au fost propuse pentru dezvoltarea sistemele de recomandare, interesul în domeniu rămâne în continuare ridicat datorită creșterii rapide a cantității de informație care este produsă în zilele noastre.

2. Motivație

Au fost depuse eforturi de cercetare semnificative pentru a furniza informații utile și relevante prin dezvoltarea de noi algoritmi și abordări în construirea sistemelor eficiente de recomandare.

De la începutul anilor '90, când a fost introdus conceptul de filtrare colaborativă, algoritmii au evoluat împreună cu înțelegerea provocărilor și a particularităților sistemelor de recomandare dar și introducerea acestora într-o gamă largă de domenii și aplicații. În 2006, Netflix, un serviciu online de închiriere de DVD-uri și de streaming video, a anunțat o competiție de 1M $ pentru îmbunătățirea corectitudinii sistemului acestora de recomandare cu 10%. Competiția a crescut interesul pentru acest subiect de cercetare (Netflix, 2006). Abordarea câștigătoare a concursului cuprinde un ansamblu de algoritmi de filtrare colaborativă precum și combinarea mai multor predictori care a depășit cu 10,06% acuratețea algoritmului de estimare a scorurilor proprietar Netflix - Cinematch (Töscher, Jahrer, & Bell, 2009). În același an, MyStrands a organizat o școală de vară destinată studiul sistemelor de recomandare: Recommenders06.

Odată cu creșterea rapidă a comercializării a sistemelor de recomandare pe internet, au apărut provocări legate de scalabilitate și performanță. Față de cercetarea anterioară, aceste sisteme trebuie să gestioneze în timp real tranzacții cu sute de mii de cereri pe secundă privind milioane de utilizatori și articole. Cercetătorii au înțeles problema și au abordat-o ținând cont de provocările tehnologice. Au fost elaborați o serie de noi algoritmi, care includ abordări bazate pe corelația elementelor, de reducere a dimensiunilor datelor precum și evaluarea unei liste top-n a elementelor de recomandare.

Începând cu anul 2007, ACM are o conferință dedicată adresată acestui subiect: ACM Recommender Systems Conference. Alte conferințe au subiecte sau chiar workshop-uri dedicate sistemelor de recomandare: IJCAI International Joint Conference on Artificial Intelligence, AAAI Conferences on Artificial Intelligence, ACM KDD Knowledge Discovery and Data Mining, European Conference on Information Retrieval, IEEE Conference on Data Mining, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. În plus, cercetarea în domeniul sistemelor de recomandare a avut o creștere semnificativă și influențe din diferite domenii: inteligența artificială, data mining, redarea informațiilor, securitate, confidențialitate și cercetare de marketing (Jannach, Zanker, Felfernig, & Friedrich, 2010).

Popularitatea sistemelor de recomandare a crescut în ultimii zece ani; atât mediul academic, cât și industria au dezvoltat instrumente și tehnici de reducere a complexității în regăsirea de informații utile și relevante. Sistemele de recomandare s-au dovedit a fi o soluție valoroasă și un instrument puternic în comerțul electronic și în alte servicii. În plus, unii furnizori, cum ar fi Google, Amazon, Netflix, LinkedIn, IMDb, Yahoo, YouTube și Yelp, au dezvoltat servicii de includ sisteme de recomandare în platformele lor (Ricci, Rokach, & Shapira, 2011).

Interesul în domeniul cercetării sistemelor de recomandare rămâne ridicat datorită nevoii utilizatorilor de a face față creșterii volumului de informație și a diversității datelor, precum și nevoia de a dezvolta aplicații practice care să ofere recomandări și conținut personalizat.

În această teză, ne vom concentra pe îmbunătățirea abordărilor anterioare, propuse pentru construirea sistemelor de recomandare, introducând o nouă serie de metode și algoritmi eficienți în proiectarea sistemelor de recomandare.

Pentru a obține recomandări eficiente, ne concentrăm asupra cercetării privind proiectarea și elaborarea modelelor predictive utilizate în construirea sistemelor de recomandare. Vom folosi mai multe baze de date de mari dimensiuni, furnizate de Yelp într-o competiție având ca scop cercetarea. Yelp este o companie cu peste 100 de milioane de utilizatori la nivel mondial, care oferă recomandări pentru restaurante de top, magazine, divertisment și diverse alte servicii.

Vom continua direcția de cercetare în domeniul analizei sentimentelor (*sentiment analysis*), opinion mining, sentiment mining și extracției sentimentelor (*sentiment extraction*) dezvoltând un algoritm pentru clasificarea opiniilor acordate serviciilor.

În teza de față vom studia algoritmi de deep learning, ce prezintă o scalabilitate semnificativă și vom introduce acești algoritmi folosind seturi de date de mari dimensiuni.

3. Structura tezei

Teza este organizată în șapte capitole după cum urmează:

1. În Capitolul 1 introducem și motivăm problema abordată în această teză. Acest capitol este împărțit în 4 secțiuni. Oferim o prezentare generală și descriem principalele contribuții ale acestei teze.
2. Capitolul 2 oferă o trecere în revistă a celor mai recente tehnologii în sistemele de recomandare și o clasificare a principalelor tipuri de abordări precum și un studiu al literaturii privind evaluarea performanței sistemelor de recomandare. Acest capitol este împărțit în 3 secțiuni:
    a) În prima secțiune, definim în mod formal problema de recomandare.
    b) În cea de-a doua secțiune, prezentăm principalele tehnici de recomandare în dezvoltarea sistemelor de recomandare care se împart în trei categorii principale: abordări bazate pe conținut, filtrarea colaborativă și tehnici hibride de recomandare. Descriem limitările specifice fiecărei tehnici de recomandare și prezentăm mai multe metode hibride care depășesc aceste probleme.
    c) În cea de-a treia secțiune, discutăm evaluarea performanței sistemelor de recomandare și descriem principale abordări utilizate: offline și online. Prezentăm cele mai populare metrici pentru evaluarea performanțelor sistemelor de recomandare utilizate în literatură (Shani & Gunawardana, 2011).
3. În Capitolul 3, dezvoltăm un sistem de recomandări bazat pe conținut. Propunem o tehnică hibridă de extragere a caracteristicile din setul de date (Yelp, Yelp 2013 Dataset Challenge, 2013) și folosim modele bazate pe conținut pentru a identifica preferințele utilizatorilor. Evaluăm performanța sistemului nostru și demonstrăm că acesta este capabil să facă previziuni bune asupra scorurilor acordate serviciilor. Acest capitol este împărțit în 5 secțiuni:
    a) În prima secțiune, ilustrăm abordarea noastră privind dezvoltarea unui sistem de recomandări bazat pe conținut prin construirea unui model pentru a prezice scorul pe care un utilizator îl va oferi unei serviciu.
    b) În cea de-a doua secțiune, prezentăm o trecere în revistă a lucrărilor anterioare care adresează sistemele de recomandare bazate pe conținut.

c) În a treia secțiune, vom descrie setul de date Yelp și introducem modelele predictive pentru recomandarea serviciilor. Descriem abordarea noastră în dezvoltarea sistemului de recomandări ce cuprinde a metodă de extragere a caracteristicilor bazată pe clustering precum și două modele predictive.

d) În următoarea secțiune, prezentăm rezultatele evaluării sistemului de recomandări bazate pe conținut propus.

4. În ultima secțiune, prezentăm concluzii și remarci finale, comparații cu alte abordări și idei privind direcțiile viitoare de cercetare.

5. Capitolul 4 prezintă cercetarea în dezvoltarea unui algoritm de procesare a textului comentariilor introduse de către utilizatori, utilizând tehnici de redare a informațiilor (*information retrieval*) și tehnici de deep learning. Propunem un model de clasificare a serviciilor utilizând Restricted Boltzmann Machines (RBMs) pentru a extrage setul de caracteristici binare. Dezvoltăm o serie algoritmi de învățare pentru clasificarea recenziilor text pentru servicii, pentru a prezice dacă un serviciu este bun pentru copii. Rezultatele experimentale obținute folosind modelarea cu RBMs depășesc rezultatele obținute folosind clasificatori tradiționali. Acest capitol este structurat după cum urmează:

   a) În prima secțiune, descriem abordarea noastră în crearea unui model pentru clasificarea serviciilor folosind Restricted Boltzmann Machines pentru a extrage setul de caracteristici binare.

   b) În cea de-a doua secțiune, prezentăm stadiul actual al tehnicii asupra dezvoltării sistemelor de recomandare folosind Restricted Boltzmann Machines.

   c) În cea de-a treia secțiune, introducem un algoritm nou utilizând o versiune binară a Restricted Boltzmann Machines pentru a extrage setul de caracteristici folosit în modelarea recenziilor textului acordat serviciilor. Recenziile textului acordat serviciilor sunt reprezentate folosind modelul sac-de-cuvinte (*Bag-of-words*). Setul de caracteristici binare extrase este utilizat într-o problemă de clasificare pentru a prezice dacă un serviciu este bun pentru copii.

   d) În continuare, prezentăm rezultatele clasificării textului recenziilor utilizând modelul propus, care cuprinde procesarea textului folosind tf-idf (*Term Frequency-Inverse Document Frequency*), extragerea caracteristicilor folosind Restricted Boltzmann Machines și regresie logistică. Comparăm

rezultatele noastre cu rezultatele obținute utilizând clasificatorii tradiționali, precum Stochastic Gradient Descent Classifier.

e) În ultima secțiune, oferim posibilele direcții de urmat în cercetările viitoare.

6. Capitolul 5 descrie noi abordări capabile să detecteze polaritatea sentimentelor unui comentariu acordat serviciilor. Continuăm direcția de cercetare în domeniul analizei sentimentelor prin introducerea de noi algoritmi în care aplicăm tehnici de prelucrare a limbajului natural în pașii de pre-procesare pentru a clasifica sentimentele utilizatorilor ținând cont de scorurile acordate. Prezentăm un studiu comparativ privind eficiența metodelor de tip ansamblu (*ensemble methods*) în clasificarea sentimentelor. Acest capitol este organizat în 5 secțiuni:
   a) În prima secțiune, descriem pe scurt metoda propusă de clasificare a polarității sentimentelor unui număr mare de recenzii acordate de utilizatori pentru restaurante, servicii și furnizori de servicii ținând cont de scorurile acestora.
   b) Apoi, prezentăm lucrări realizate în domeniul analizei sentimentelor și a clasificării recenziilor textului utilizând o serie de tehnici de învățare automată.
   c) În secțiunea a treia, dezvoltăm doi algoritmi pentru extragerea caracteristicilor. Prima abordare propune crearea unui dicționar propriu folosind setul de date de antrenament. A doua abordare propune efectuarea unei analize lexicale a textului recenziilor acordate serviciilor. Aplicăm următorii clasificatori: Naive Bayes, Linear Support Vector Classification (SVC), Logistic Regression and Stochastic Gradient Descent (SGD) Classifier.
   d) În următoarea secțiune, evaluăm acuratețea metodelor propuse. Abordarea noastră explorează o serie metode de extragere a caracteristicilor precum și metode de învățare pentru clasificarea recenziilor textului utilizând un set de date de mari dimensiuni: *Yelp challenge dataset* care cuprinde peste 1,6 milioane de recenzii acordate serviciilor (Yelp, Yelp 2015 Dataset Challenge, 2015).
   e) În ultima secțiune discutăm idei extrase din această lucrare privind direcții noi de cercetare.

7. Capitolul 6 continuă cercetarea în domeniul analizei sentimentelor și deep learning precum și a activității descrise în capitolul precedent prin introducerea unei abordări

specifice pentru clasificarea comentariilor acordate serviciilor, abordare care utilizează rețele neuronale de tip Convolutional Neural Networks bazate pe cuvinte. Propunem două abordări ale modelelor de rețele neuronale care au dimensiuni arhitecturale diferite și reprezentări vectoriale ale cuvintelor: încorporări de cuvinte pre-instruite și reprezentări vectoriale end-to-end. Efectuăm un studiu empiric asupra efectului hiper-parametrilor în performanța clasificării sentimentelor din textul comentariilor. De asemenea, implementăm mai multe experimente pe un set de date de mari dimensiuni pentru a capta relația semantică din texul comentariilor. Arătam că rezultatele noastre sunt comparabile cu rezultatele din abordările folosind metode tradiționale. Acest capitol este structurat după cum urmează:

a) În prima secțiune, prezentăm rețele neuronale de tip Convolutional Neural Networks și descriem abordarea noastră pentru clasificarea textului comentariilor acordate serviciilor.
b) Apoi, discutăm rezultatele anterioare în domeniul analizei sentimentelor și clasificării textului utilizând Convolutional Neural Networks pe seturi de date de mari dimensiuni.
c) În secțiunea a treia introducem două abordări ale modelelor de rețele neuronale cu dimensiuni arhitecturale diferite și o serie de reprezentări vectoriale ale cuvintelor de diferite dimensiuni: încorporări de cuvinte pre-instruite dar și reprezentări vectoriale end-to-end.
d) Apoi, efectuăm o serie de experimente pentru a explora efectul componentelor din arhitectură asupra performanței modelului: hyperparameters tuning, dimensiunea regiunii filtrului, numărul de caracteristici - feature maps și parametrii de regularizare pentru rețelele de tip Convolutional Neural Networks propuse. Discutăm deciziile de proiectare pentru clasificarea sentimentelor pe setul de date *Yelp 2017 dataset* (Yelp, Yelp 2017 Dataset Challenge, 2017) care cuprinde 4,1M de recenzii. Oferim o comparație între aceste modele și raportăm acuratețea obținută. Mai mult decât atât, deoarece la momentul acestei scrieri nu există rezultate raportate pe setul de date *Yelp 2017* privind clasificare a sentimentelor, extindem experimentele noastre pe setul de date *Yelp 2015* (*Yelp 2015 polarity dataset*) pentru a oferi o comparație cu rezultate raportate în (Zhang, Zhao, & LeCun, 2015).

e) În secțiunea a cincea vom concluziona rezultatele prezentate în acest capitol și vom oferi posibile direcții de cercetare viitoare pentru clasificarea sentimentelor folosind textul comentariilor acordate de utilizatorii Yelp.

8. În ultimul capitol, prezentăm concluziile acestei teze și principalele contribuții. Discutăm o serie de problemele deschise și posibile direcții viitoare pentru îmbunătățirea acestei cercetări în domeniul sistemelor de recomandare.

4. Contribuții

Am discutat motivația privind cercetarea noastră în domeniul sistemelor de recomandare. În continuare, descriem contribuțiile științifice prezentate în această teză. Cele mai importante rezultate care au fost publicate în (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014), (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014), (Salinca, Business reviews classification using sentiment analysis, 2015) și (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

Obiectivul principal al tezei este de a dezvolta abordări eficiente în construirea sistemelor de recomandare folosind diferite tehnici de învățare automată cu accent pe algoritmi de deep learning, ce prezintă o scalabilitate semnificativă și vom introduce acești algoritmi folosind seturi de date de mari dimensiuni.

Pentru a obține recomandări eficiente, ne concentrăm asupra cercetării privind proiectarea și elaborarea modelelor predictive utilizate în construirea sistemelor de recomandare. Vom folosi mai multe baze de date de mari dimensiuni, furnizate de Yelp într-o competiție având ca scop cercetarea (Yelp, Yelp 2013 Dataset Challenge, 2013), (Yelp, Yelp 2015 Dataset Challenge, 2015), (Yelp, Yelp 2017 Dataset Challenge, 2017).

De asemenea, am participat într-o competiție organizată de Yelp: *Yelp Dataset Challenge* (Yelp, Yelp 2013 Dataset Challenge, 2013). Yelp este o companie cu peste 100 de milioane de utilizatori la nivel mondial, care oferă recomandări pentru restaurante de top, magazine, divertisment și diverse alte servicii. Prima contribuție este discutată în Capitolul 3. Introducem o abordare de tip hibrid pentru extragerea caracteristicilor pentru dezvoltarea unui sistem de recomandări utilizând un set de date de mari dimensiuni furnizat de Yelp, unul dintre cele mai populare site-uri de recomandări. Pentru a identifica preferințele utilizatorilor Yelp, extragem caracteristicile din setul de date folosind o tehnică hibridă și construim modele bazate pe conținut. Evaluăm performanțele sistemului de recomandare propus folosind eroarea rădăcinii medie pătrată - *Root Metrics Mean Squared Error* (RMSE) (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014).

Ca parte a cercetării noastre, dezvoltăm un sistem de recomandări bazat pe conținut și propunem o metodă de tip hibrid de extragere a caracteristicilor. Extragerea caracteristicilor are un rol major în construirea de sisteme de recomandare eficiente. Extragerea caracteristicilor este esențială în reducerea costurilor computaționale dar și în reducerea dimensiunilor datelor. Identificăm mai multe variabile referitoare la caracteristicile extrase din modelarea serviciilor și a utilizatorilor și evaluăm impactul acestora asupra acurateței modelului. Efectuăm o evaluare de tip offline asupra scorului acordat serviciilor, ce cuprinde 200 000 de recomandări acordate de 40 000 de utilizatori. Sistemul de recomandări bazat pe conținut propus a obținut un scor RMSE de 1,24 atunci când am utilizat Decision Regression Trees ca model predictiv. În competiția ACM Yelp RecSys challenge (Kaggle, 2013), cel mai bun scor RMSE obținut a fost 1,21251 folosind un sistem hibrid de recomandări care combină mai multe abordări precum filtrarea colaborativă și filtrarea pe bază de conținut (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014).

În continuare, contribuția din Capitolul 4 prezintă cercetarea în dezvoltarea celui de-al doilea sistem de recomandări care se concentrează pe modelarea textului comentariilor introduse de către utilizatori folosind setul de date Yelp pentru a prezice dacă o afacere este bună pentru copii. În dezvoltarea acestui sistem de recomandări vom folosi doar textul comentariilor acordat serviciilor.

În metoda propusă, extragem caracteristici non-liniare din textul comentariilor introduse de către utilizatori folosind un model probabilistic: Restricted Boltzmann Machines (RBMs), și creăm un model pentru clasificarea serviciilor. Combinăm tehnici de redare a informațiilor (*information retrieval*) pentru procesarea textului comentariilor acordate serviciilor, împreună cu tehnici de deep learning (RBMs) pentru a extrage setul caracteristicile binare și propunem o serie algoritmi de învățare pentru clasificarea serviciilor folosind diferiți algoritmi de învățare (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014).

Rezultatele experimentale obținute folosind procesarea textului comentariilor folosind tf-idf (*Term Frequency-Inverse Document Frequency*), modelarea cu RBMs pentru extragerea caracteristicilor și regresie logistică, depășesc rezultatele obținute folosind clasificatori tradiționali precum Stochastic Gradient Descent și metoda tradiționala de regresie logistică (*Logistic Regression*) (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014).

În Capitolul 5, dezvoltăm noi algoritmi în domeniul analizei sentimentelor, opinion mining, sentiment mining și sentiment extraction capabil să detecteze polaritatea sentimentulelor unui comentariu acordat serviciilor utilizând un nou set de date de mari dimensiuni furnizat de Yelp: Yelp Challenge dataset (care cuprinde mai mult de 1,6 milioane de comentarii acordate de utilizatori). Construim mai multe metode de învățare de tip ansamblu (*ensemble methods*) pentru clasificarea sentimentelor comentariilor acordate serviciilor folosind două metode de extragere a caracteristicilor și patru modele de învățare automată. În plus, introducem noi algoritmi în care aplicăm tehnici de prelucrare a limbajului natural în pașii de pre-procesare pentru a obține o acuratețe ridicată în clasificarea sentimentelor. Utilizând doar textul comentariilor construim abordări capabile sa detecteze sentimentele utilizatorilor ca fie pozitive sau negative, ținând cont de scorurile acordate pentru restaurante, servicii și furnizori de servicii (Salinca, Business reviews classification using sentiment analysis, 2015).

Aceasta a fost prima abordare a clasificării sentimentelor unui comentariu acordat serviciilor care utilizează cele mai recente comentarii din setul de date Yelp challenge dataset (Yelp, Yelp 2015 Dataset Challenge, 2015). Demonstrăm că algoritmul nostru de analiză a sentimentelor construit folosind algoritmi de învățare automată, precum Naive Bayes și Linear Support Vector Classification (SVC), au o acuratețe de peste 90%. Cele mai bune rezultate au fost obținute folosind clasificatorii Linear SVC și Stochastic Gradient Descent Classifier (SGD) obținând o acuratețe de 94,4%. Cu toate acestea, în studiul lui Wilson et al. analiza umană în exprimarea sentimentelor, a trăirilor, a emoțiilor și a opiniilor reprezintă o sarcină dificilă. A fost măsurat un procent de acord de 82% în identificarea automată a expresiei subiective între două persoane, dovedind că algoritmul pe care l-am propus este eficient în clasificarea sentimentelor (Salinca, Business reviews classification using sentiment analysis, 2015).

Rețelele neuronale de tip Convolutional Neural Networks (CNNs) au dovedit rezultate remarcabile în clasificarea textului și în domeniul analizei sentimentelor. În Capitolul 6 continuăm cercetările noastre privind analiza sentimentelor și introducem o abordare specifică pentru clasificarea comentariilor acordate serviciilor, abordare care utilizează rețele neuronale de tip Convolutional Neural Networks bazate pe cuvinte și reprezentări vectoriale de cuvinte (*word embeddings*). Comparăm mai multe abordări folosind rețele neuronale de tip Convolutional Neural Networks bazate pe cuvinte, care au reprezentări vectoriale diferite: încorporări de cuvinte pre-instruite și reprezentări vectoriale end-to-end.

Implementăm mai multe experimente pe un set de date de mari dimensiuni *Yelp 2017 challenge dataset* furnizat de Yelp (Yelp, Yelp 2017 Dataset Challenge, 2017), pentru a capta

relația semantică din texul comentariilor. Utilizăm tehnici de deep learning și arătam că rezultatele noastre sunt comparabile cu rezultatele din abordările folosind metode tradiționale. Analizăm mai multe modele de încorporări de cuvinte pre-instruite construite folosind algoritmi de învățare nesupervizați pentru a obține reprezentări vectoriale ale cuvintelor: GloVe (Karpathy & Fei-Fei, 2015), word2vec împreună cu vectori pre-instruiți folosind 100 miliarde de cuvinte din setul de date Google News. Modelele conțin vectori de dimensiune 100 pentru trei milioane de cuvinte și fraze (Mikolov, Chen, Corrado, & Dean, 2013). De asemenea, folosim în experimente FastText - vectori de cuvinte pre-instruiți pentru limba engleză, care reprezintă o extensie a word2vec. Acești vectori cu dimensiunea 300 au fost obținuți folosind setul de date Wikipedia și utilizând modelul de skip-gram descris în (Bojanowski, Grave, Joulin, & Mikolov, 2017) cu parametrii impliciți. Mai mult, vom folosi în stratul embedding layer din arhitectura ambelor modele CNNs propuse o reprezentare vectorială ce cuprinde vectori de încorporare word2vec (*embedding vectors*) de dimensiune 100 pe care i-am antrenat folosind text comentariilor acordate serviciilor din setul de date de antrenament.

Raportăm un scor de acuratețe de 95,6% utilizând 3-fold cross validation, reprezentări vectoriale pre-instruite de cuvinte FastText și CNNs având trei regiuni de filtru și 128 de harți de trăsături (*feature maps*) (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

Mai mult decât atât, deoarece în momentul în acestei scrieri, nu s-au raportat rezultate folosind setul de date *Yelp 2017 challenge dataset*, evaluăm modelul propus pe setul de date de polaritate *Yelp 2015* pentru a compara rezultatele noastre cu cele obținute de alte lucrări. Raportăm un scor de acuratețe de 94,27% folosind CNNs bazate pe cuvinte și reprezentări vectoriale word2vec pre-instruite de dimensiune 300. În (Zhang, Zhao, & LeCun, 2015), autorii propun o arhitectură de rețele CNNs la nivel de caractere care obține un scor de acuratețe de 94,11% pentru arhitectura cu caracteristici de mari dimensiuni și 93,47% pentru arhitectura cu caracteristici mici dimensiuni pe setul de date de polaritate *Yelp 2015*. Autorii prezintă, de asemenea, două implementări pentru CNNs bazate pe cuvinte cu reprezentări vectoriale pre-instruite word2vec și obțin un scor de 95,40% pentru o arhitectură cu caracteristici de mari dimensiuni și 94,44% pentru o arhitectură cu caracteristici de mici dimensiuni pe setul de date de polaritate *Yelp 2015* (Zhang, Zhao, & LeCun, 2015). Astfel, arătăm că arhitectura noastră folosind CNNs obține rezultate comparabile cu rezultatele abordărilor din literatură utilizate în clasificarea textului comentariilor acordate serviciilor (Zhang, Zhao, & LeCun, 2015), (Tang, Qin, & Liu, 2015).

În această teză, vom oferi o descriere detaliată a provocărilor care apar în contextul proiectării sistemelor de recomandare folosind tehnici de învățare automată și vom descrie abordări, idei, concepte și tehnici noi.

Am prezentat principalele rezultate din această teză la conferințe internaționale de top din domeniul inteligenței artificiale, precum International Joint Conference on Artificial Intelligence (IJCAI) (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

Sperăm că abordările noastre inovatoare vor ajuta în cercetarea domeniului sistemelor de recomandare pentru a descoperi noi orizonturi.

5. Bibliografie

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering, 17*(6), 734-749.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5, 2*, 135-146.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction.* Cambridge University Press.

Kaggle. (2013). *ACM Yelp RecSys challenge*. Retrieved September 2014, from Kaggle: https://www.kaggle.com/c/yelp-recsys-2013

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3128–3137).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Netflix. (2006). *The Netflix Prize*. Retrieved from Netflix: http://www.netflixprize.com/

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook.* Springer US.

Salinca, A. (2014). A hybrid approach of feature extraction for content-based recommender system. *4th International Conference on Hybrid Intelligent Systems, HIS* (pp. 54-57). Trends in Innovative Computing.

Salinca, A. (2014). Restricted Boltzmann Machines for modeling businesses. *International Conference on Artificial Intelligence and Pattern Recognition, AIPR* (pp. 24-28). SDIWC.

Salinca, A. (2015). Business reviews classification using sentiment analysis. *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015 17th International Symposium on.* Timisoara: IEEE.

Salinca, A. (2017). Convolutional Neural Networks for Sentiment Classification on Business Reviews. *IJCAI Workshop on Semantic Machine Learning.* Melbourne: CEUR.

Shani, G., & Gunawardana, A. (2011). *Evaluating Recommendation Systems.* Boston, MA: Springer US.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. *EMNLP*, (pp. 1422-1432).

Töscher, A., Jahrer, M., & Bell, R. M. (2009). The bigchaos solution to the Netflix grand prize. *Netflix prize documentation*, 1-52.

Yelp. (2013). *Yelp 2013 Dataset Challenge*. Retrieved from Yelp Dataset Challenge: http://www.yelp.com/dataset_challenge

Yelp. (2015). *Yelp 2015 Dataset Challenge*. Retrieved from Yelp 2015 Dataset Challenge: https://www.yelp.com/dataset_challenge

Yelp. (2017). *Yelp 2017 Dataset Challenge*. Retrieved from Yelp Dataset Challenge: https://www.yelp.com/dataset_challenge

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, (pp. 649-657).