**Analiza Sentimentelor Curselor cu Uberul**

Șoroagă Simina

**Link proiect github:**

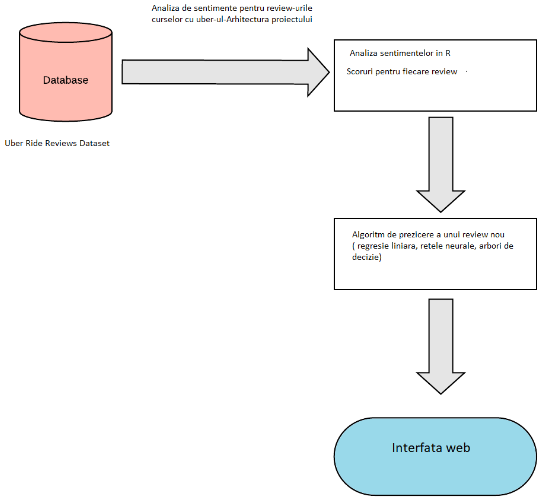
**Scurtă descriere a proiectului**

În acest proiect mi-am propus să analizez sentimentele, opiniile și emoțiile recenziilor oferite în legatură cu traseele cu uberul. Pentru aceasta analiză am utilizat setul de date *Uber\_Ride\_Reviews* ce conține 760 de recenzii și două variabile: ride\_review(string) și ride\_rating(numeric).

Pentru a crea corpusul ce va fi utilizat la analiza sentimentelor am adus cateva modificari variabilei ride\_review astfel încât să analizăm din recenzie doar cuvintele care au semnificație și sunt importante în analiză. Astfel, am eliminat mai întăi toate semnele de punctuație, am eliminat cuvintele de legătura, am lematizat cuvintele, am eliminat cifrele și alte caractere nesemnificative, etc. În final, folosind funcția **sentiment\_by** disponibilă in pachetul **pcman** din R am obținut pentru fiecare recenzie cate un scor.

Pentru modulul de învătare automata am utilizat Retelele Neuronale și Regresia Liniara și am obținut o acuratețe de 96% a modelului.

**Arhitectura proiectului:**

****

**Detaliile proiectului:**

*Scopul proiectului*

Scopul proiectului a fost acela de a identifica sentimentele persoanelor care au lasat recenziile curselor cu uberul, să centrazilizez toate recenziile asfel încât urmatorii clienți să iși poată crea o imagine asupra companiei de uber și să poată lua o decizie, fără să fie nevoiți sa citească miile de recenzii primate deja.

*Provocarea proiectului:* Provocarea a constat în crearea unei interfețe interactive și atrăgătoare astfel încât clienții să accese site-ul și să ofere recenzii.

**Literatura de specialitate:**

Fiind un domeniu vast, analiza sentimentelor a fost studiată de mai mulți cercetatori din întreaga lume care au încercat să construiască metode de clasificare a recenziilor produselor sau serviciilor care să aibă o acuratețe cât mai mare.

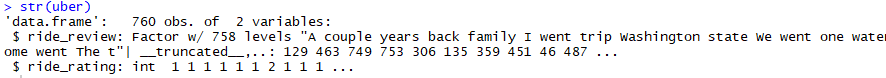
În lucrarea “*Classification* *of Sentimental Reviews Using Machine Learning Techniques*” scrisă de Abinash Tripathy, Ankit Agrawal și Santanu Kumar Rath este prezentata o comparație a rezultatelor obținute prin aplicarea algoritmului de clasificare Naïve Bayes(NB) și Suport Vector Machine (SVM). Setul de date folosit pentru instuirea și testarea modelelor în aceasta lucrare a cuprins recenziile filmelor și s-a facut o comparație a rezultatelor obținute cu cele existente deja în literaturade specialitate. Rezultatul studiului a aratat faptul că metoda SVM depășește orice alt clasificator în prezicerea sentimentului unei recenzii. O altă abordare a clasificarii recenziilor este propusa de Deepali Virmani, Preeti Arora și Pradnya Satish Kulkarni în articolul “ *Cross domain analyzer to acquire review proficiency in big data*”. Aceștia propun o clasificare transversală a domeniilor, un analizator de domenii incrucișate(CDA) ce extrage cuvintele pozitive din recenziile unui domeniu și le inlocuiește cu sinonimele sale, astfel încât să reușească să analizeze sentimentele din recenziile unui alt domeniu.Datele utilizate în acest experiment au fost culese de pe Amazon și s-a obtinut o precizie de 92%, algoritmul creând o rafinare de 5% față de tehnicile deja existente. În literatura de specialitate s-a studiat și calitatea produselor software. Deoarece modelele de masurare a calitatii software-ul în uz (QinU) sunt ineficiente și multe dintre ele sunt subiective, Issa Atoum propune în lucrarea intitulată *“ A novel framework for measuring software quality-in-use based on semantic similarity and sentiment analysis of software reviews*” un cadru nou de lucru numit QinUF pentru a masura calitatea software-ului în utilizare folosind recenziile obținute de aceste produse. Acest nou cadru este alcătuit din trei componente: predicția QinU, clasificarea polarității și notarea QinU. În viața de zi cu zi, QinUF permite utilizatorilor să compare și să ofere recenzii produselor software. Experimentele efectuate pe parcursul acestui studiu au demonstrat faptul că QinUF a obținut cea mai înaltă performanță, în comparație cu metodele cunoscute. Tot pentru analiza sentimentelor din spatele recenziilor Shahana P.H și Bini Omman folosesc în studiul lor intitulat “ *Evaluation of Features on Sentimental Analysis*” tehnici precum Mutual Information, Chi-Square, Information gain și TF-idf pentru a selecta caracteristicile din baza de date. Aceste modele sunt evaluate pentru un set de date ce conține 2000 de recenzii ale celor mai cunoscute filme și se investighează care este cea mai buna metoda de analiză a sentimentelor din recenzii. Autorii au constatat faptul că cea mai buna metoda de extragere a sentimentelor din recenzii este unigramul care oferă o precizie de 83% în ceea ce privește clasificarea sentimentelor.

Deoarece pentru un singur produs sau serviciu pot exista sute sau chiar mii de recenzii și este dificil să citim și să analizam toate recenziile pentru a lua o decizie cu privire la achiziționarea acestora,  [Aarti Potdar,](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!)  [Pranav Patil,](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) [Raunak Bagla,](https://www.sciencedirect.com/science/article/pii/S1877050916000983" \l "!) [Rohitashwa Pandey](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) și Prof. [Nagesh Jadhav](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) ne propun in lucrarea “*SAMIKSHA - Sentiment Based Product Review Analysis System*” un bot de recenzie care generează o rezumare faptică a tuturor recenziilor utilizatorilor. Software-ul propus produce o evaluare numerică medie a tuturor caracteristicilor specific produsului pentru a ajuta cumparatorul să obțină o imagine de ansamblu detaliată a produsului. Astfel, “Samiksha” s-a dovedit a fi un mediu convenabil de analiza a tuturor recenziilor și este benefic pentru cumpărătorii de pe site-urile de comerț electronic.

Deoarece exista numerice studii pe acestă temă, am încercat să aplic o parte din algoritmii și metodele de analiză a sentimentelor din recenzii descrise în aceste articole.

**Descrierea proiectului**

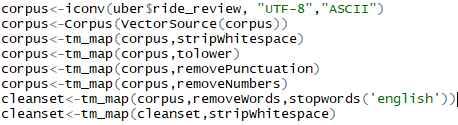
Setul de date utilizat se numește Uber\_Ride\_Review și are următoarea componență



Pentru analiza sentimentelor am folosit programul R.

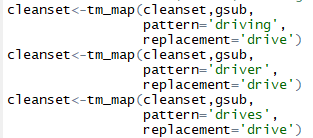
* **Modulul 1- Analiza sentimentelor**

Înainte de a analiza sentimentele recenziilor am făcut câteva modificări preliminare corpusului, în special variabilei *ride\_review*, care conține recenziile.

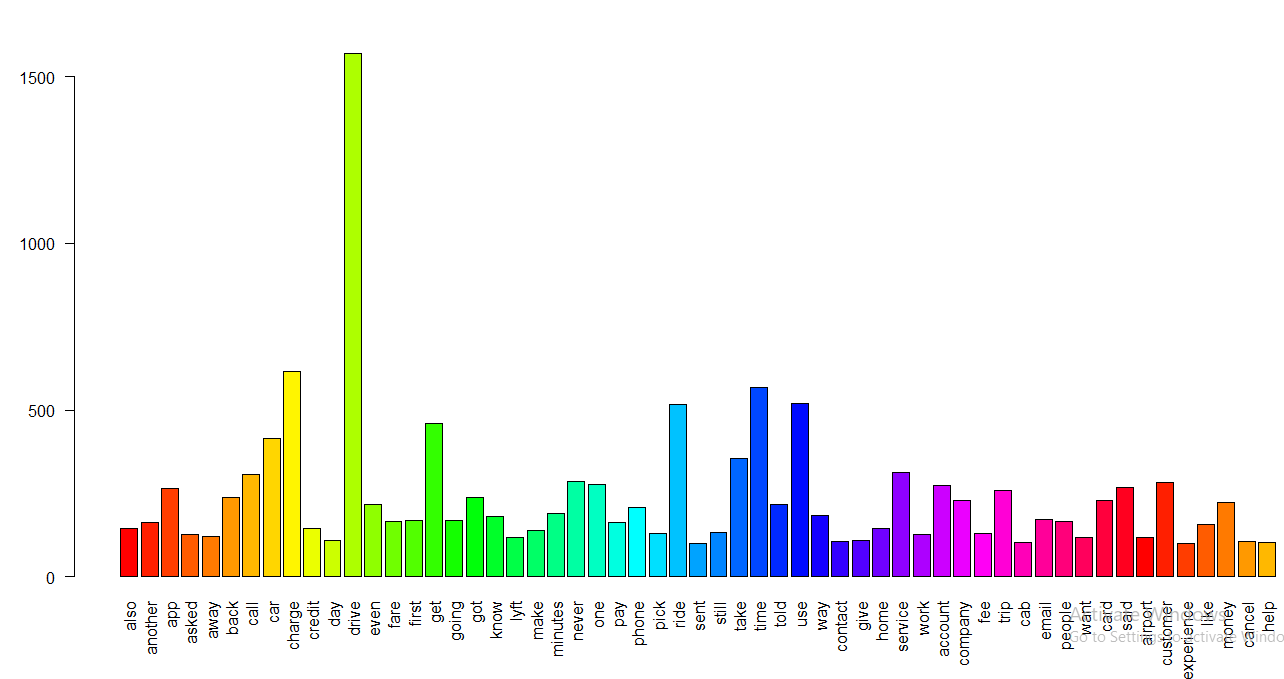


Am eliminat semnele de punctuație, am transformat tot textul in litere mici, am eliminat numerele, spațiile albe și cuvintele de legătură.

Apoi am lematizat cuvintele



și am reprezentat frecvența apariției cuvintelor intr-o histogramă

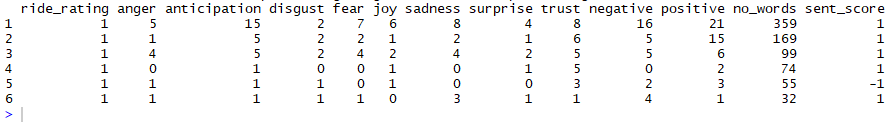


După ce am curățat corpusul, am aplicat algoritmul de analiză a sentimentelor, am folosit funcția sentiment\_by și am obțiunut un scor pentru fiecare recenzie



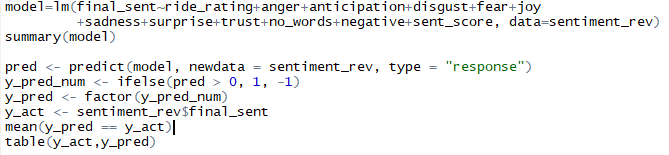
De asemenea, pe langă scorurile rezultate, am introdus și o nouă variabilă *no\_words* în care am memorat numărul de cuvinte pe care îl are fiecare recenzie.

Baza de date rezultată este următoarea:

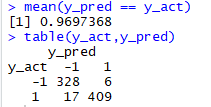


* **Modulul 2-Învățarea automată**

Invățarea automată am realizat-o utilizând Regresia Liniară și Rețele Neuronale.



Folosind Regresia Liniară am obțiunut cel mai bun model, un model cu o acuratețe de prezicere de 96%.

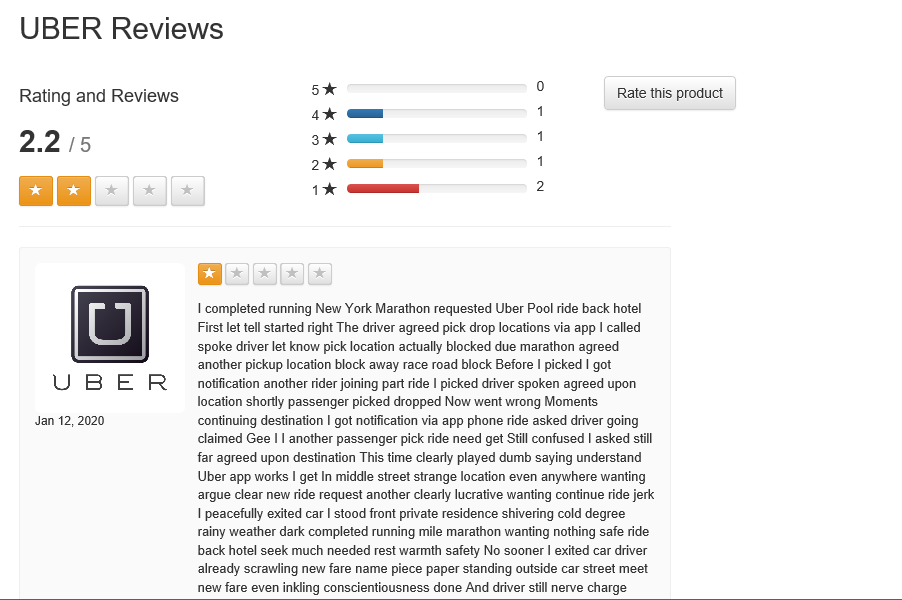


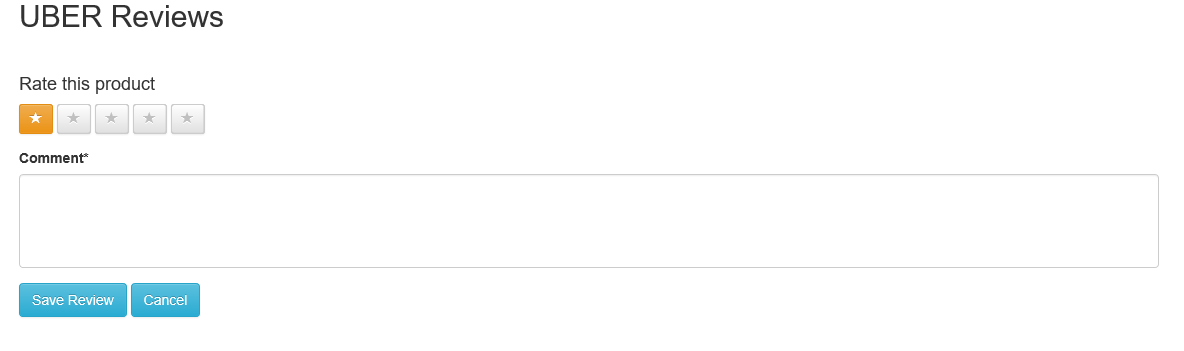
* **Modulul 3-Interfața Web**

Am realizat o interfata web care să permită utilizatorilor să lase recenzii companieide uber și să vadă recenziile lăsate de ceilalți utilizatori. În plus, aceasta oferă și o rezumare a recenziilor primate, clienții pot verifica ușor care este numărul de recenzii lăsate, si care este media acestora.

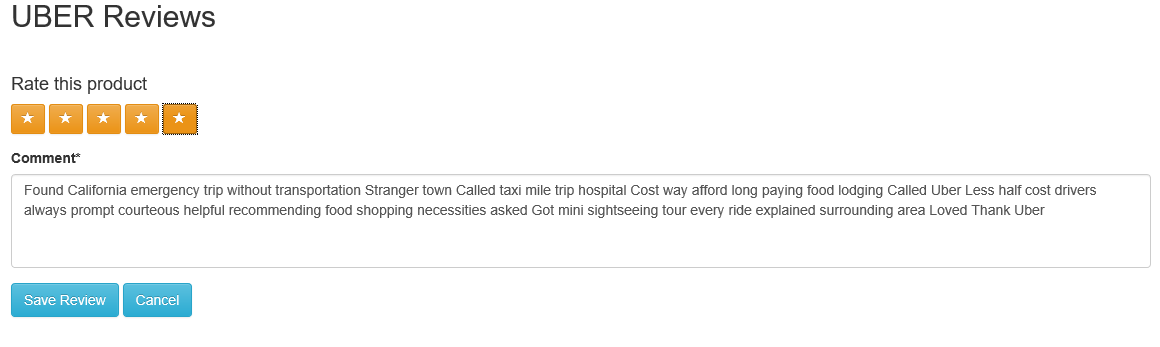
Pe prima pagina a site-ului putem citi recenziile pe care le-a primit compania de uber si putem vedea usor care este media numarului de stelute pe care le-a primit.

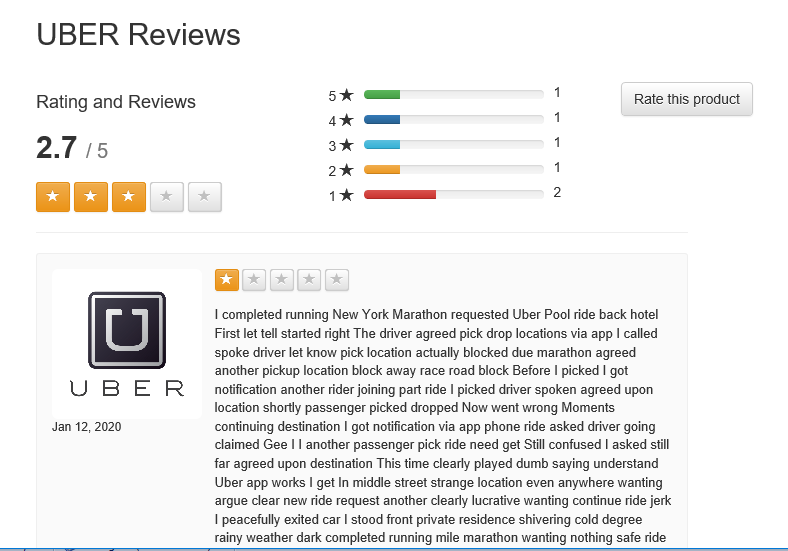
Iar dacă apăsăm butonul *Rate this product* suntem trimiși catre pagina a doua, unde putem lasa o recenzie a unei curse cu uberul.





In momentul de fata nu este nicio recenzie cu 5 stelute, vom introduce noi o recenzie si vom vedea cum se modifică media stelutelor și numărul de recenzii.





**Managementul Proiectului**

Termenele limita pentru fiecare modul al proiectelor au fost următoarele:

-08.12- am ales tema proiectului și am citit literature de specialitate pentru a întelege mai bine ce presupune analiza sentimentelor și care sunt algoritmii pe care îi pot utilize

-15.12- am construit arhitectura proiectului si primul modul, analiza sentimentelor

-22.12-am contruit modulul 2

-5.01-varianta beta a proiectului si interfata

-12.01-finalizare proiect, modificare interfata și descrierea proiectului.