Contents

[**Prezentarea setului de date** 2](#_Toc95758950)

[**Împărțirea setului de date** 3](#_Toc95758951)

[**Distribuția claselor in seturile de date** 5](#_Toc95758952)

[Setul de antrenament 5](#_Toc95758953)

[Setul de validare 6](#_Toc95758954)

[Setul de testare 6](#_Toc95758955)

[**Preprocesarea datelor** 7](#_Toc95758956)

[Eliminarea semnelor de punctuație 7](#_Toc95758957)

[Eliminarea stopwords 8](#_Toc95758958)

[Trunchierea cuvintelor 12](#_Toc95758959)

[Eliminarea cuvintelor rar întâlnite 12](#_Toc95758960)

[Eliminarea recenziilor de tip „outlier” 13](#_Toc95758961)

[Concluzie 13](#_Toc95758962)

# **Prezentarea setului de date**

**Surse:**

* [SNAP: Web data: Amazon reviews (stanford.edu)](http://snap.stanford.edu/data/web-Amazon.html)
* Xiang Zhang, Junbo Zhao, Yann LeCun. [Character-level Convolutional Networks for Text Classification](https://arxiv.org/abs/1509.01626). Advances in Neural Information Processing Systems 28 (NIPS 2015).

**Prezentare generala:**

Setul de date constă in recenzii de pe Amazon. Datele se întind pe o perioadă de 18 ani intre 1995 si 2013, incluzând aproximativ 35 de milioane de entități. In total setul înglobează aproximativ 2,84 miliarde de cuvinte cu o medie de 82 cuvinte pe recenzie.

**Descriere:**

Setul de date conține 34,686,770 de recenzii de pe Amazon de la 6,643,669 de utilizatori pentru 2,441,053 de produse. Recenziile includ informații despre produse și utilizatori, evaluări (valori de la 1 la 5) și o recenzie în text.

Subsetul de date utilizat a fost construit luând scorurile 1 și 2 ca fiind negative iar scorurile 4 și 5 ca pozitive. Recenziile cu scorul 3 sunt ignorate. În setul de date, clasa 0 reprezintă recenziile negative iar clasa 2 cele pozitive. Fiecare clasă are 20.000.000 de mostre împărțite in 16.000.000 pentru subsetul de antrenament 2.000.000 pentru subsetul de validare si 2.000.000 pentru subsetul de testare însumând 40.000.000 de recenzii.

# **Împărțirea setului de date**

Pentru a putea evalua corect performanța modelului dezvoltat**,** trebuie săîmpărțim setul de date in trei parți distincte, respectiv seturile de antrenare, testare si validare, fiecare fiind utilizatîntr-o etapă diferită a procesului de dezvoltare.

**Setul de antrenante** este constituit din 32.000.000 de recenzii cu o lungime maxima de 255 cuvinte si un număr mediu de 74.16 cuvinte pe recenzie. Acesta este folosit pentru a antrena și a face modelul să învețe. În fiecare epocă datele de antrenament sunt transmise rețelei neuronale în mod repetat, iar modelul continuă să învețe caracteristicile acestora.

**Setul de validare** este utilizat pentru a verifica performanța modelului în timpul antrenamentului. La fiecare epocă modelul este antrenat pe setul de antrenament si evaluat cu setul de validare, acest proces oferă informații care ne ajută să reglam parametrii modelului. Acesta este constituit din 4.000.000 de recenzii cu o lungime maxima de 219 cuvinte si un număr mediu de 74.17 cuvinte pe recenzie. Fata de setul de antrenament lungimea maxima a scăzut cu 14.12% însă cea medie a rămas constanta, indicând posibilitatea prezentei unor recenzii “outliers” in setul de antrenare.

**Setul de testare** este utilizat pentru a testa modelul după finalizarea antrenării. Acesta trebuie sa urmează aceeași distribuție a claselor ca și setul de date de antrenament pentru a putea oferii o imagine de încredere a performantelor modelului. Acesta conține 4.000.000 de recenzii cu o lungime maxima de 217 cuvinte si un număr mediu de 74.10 cuvinte pe recenzie. Lungimea maxima este in scădere cu 14.91% fata de setul de antrenament, având o valoare aproximativ egala cu setul de validare, confirmând prezenta unor “outliers” in setul de antrenare.

# **Distribuția claselor in seturile de date**

Deoarece avem o problema de clasificare iar cele doua clase(pozitiv si negativ) conțin un nivel egal de informație este important sa folosim un set de date echilibrat pentru a nu introduce in model prejudecați(biases).

Setul de antrenament

Intre cele 2 clase prezente setul de antrenament exista o diferența de 8.3% intre lungimea medie a recenziilor pozitive si lungimea medie a recenziilor negative.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Număr de cuvinte | Lungime maxima | Lungime medie |
| Subset pozitiv | 113,897,346 | 214 | 71.18 |
| Subset negativ | 123,441,444 | 255 | 77.15 |

Setul de validare

Distribuția dinte cele 2 clase a rămas relative identica, subsetul negativ având o lungime medie cu 8.21% mai mare decât subsetul pozitiv.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Număr de cuvinte | Lungime maxima | Lungime medie |
| Subset pozitiv | 14,249,477 | 219 | 71.24 |
| Subset negativ | 15,419,213 | 214 | 77.09 |

## Setul de testare

Distribuția dinte cele 2 clase este consistenta cu cele 2 seturi de date prezentate anterior, subsetul negativ având o lungime medie cu 8.16% mai mare decât subsetul pozitiv.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Număr de cuvinte | Lungime maxima | Lungime medie |
| Subset pozitiv | 14,240,026 | 216 | 71.20 |
| Subset negativ | 15,402,099 | 217 | 77.01 |

# **Preprocesarea datelor**

Preprocesarea reprezintă totalitatea metodelor prin care textul este curățat si pregătit pentru a fii introdus in model. Acest pas este efectuat atât înainte de procesele de antrenare si evaluarea a modelului cat si înainte de clasificarea unui text.

In acest proces se urmărește păstrarea unui nivel de informații cat mai mare in textul dat, obținând însă o dimensiune cat mai mica a datelor pentru a optimiza procesele de antrenare, evaluare si clasificare.

Un prin pas in preprocesarea datelor reprezintă transformarea tuturor cuvintelor in litera mica pentru a da o oarecare uniformitate a datelor, necesara pentru ulterioarele metode de preprocesarea aplicate.

## Eliminarea semnelor de punctuație

Deoarece utilizarea semnelor de punctuație în recenziile seturilor de date existente și modul în care acestea sunt utilizate în știri diferă semnificativ, am decis să le eliminam pentru a asigura o mai bună generalizare a modelului dezvoltat.

Cele trei seturi de date utilizate prezinta o medie cuprinsa intre 6.88 si 6.93 cuvinte la fiecare semn de punctuație utilizat asigurând astfel menținerea unei uniformități a datelor după procesul de curățare.

In contextul dat vom definii un semn de punctuație, conform cu modulul standard „string” aferent limbajului de programare Python 3.0, ca orice caracter prezent in următorul sir : **!"#$%&'()\*+,-./:;<=>?@[\]^\_`{|}~** .

**Numărul de semne de punctuație din diferitele seturi**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Antrenament | Validare | Testare |
| Total | 34,366,892 | 4,295,238 | 4,292,727 |
| Subset pozitiv | 16,434,458 | 2,056,272 | 2,054,143 |
| Subset negativ | 17,932,434 | 2,238,966 | 2,238,584 |
|  |  |  |  |

In urma eliminării am redus lungimea medie a frazelor din setul de antrenament cu AICI%, obținând o lungime medie de AICI cuvinte, scăzând astfel timpul si resursele necesare pentru antrenarea modelului .

## Eliminarea stopwords

“Stopwords” sunt cuvintele care nu oferă nicio informație utilă pentru a determina în ce categorie ar trebui să se încadreze textul. Aceasta se poate datora faptului că nu au sens (prepoziții, conjuncții etc.), sau pentru că sunt prea frecvente în contextul clasificării.

Prin extragerea acestor cuvinte, eliminam informațiile de nivel scăzut din text pentru a ne concentra mai mult pe informațiile importante. Cu alte cuvinte, putem spune că ștergerea acestor cuvinte nu va avea niciun impact negativ asupra modelului pe care îl antrenăm.

In procesul de preprocesare vom folosii lista de **stopwords** din limba englezaoferita de către modulul **Natural Language Toolkit (NLTK)** pentrulimbajului de programare Python 3.0.

Datorita rolului specific pe care trebuie sa îl îndeplinească modelul dezvoltat am ales separarea listei oferite de modulul **NLTK** in doua liste distincte in funcție de conotațiile negative ale cuvintelor (“Stopwords neutre” si “Stopwords negative”).

Acestea vor fii analizate in mod individual si eliminate/menținute in funcție de rezultatele obținute.

## 

**Stopwords neutre** :

In contextul dat vom definii un stopword neutru ca orice cuvânt prezent in următorul sir:

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y']

**Procentul de Stopwords neutre din totalitate cuvintelor din diferitele seturi**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Antrenament | Validare | Testare |
| Set | 46.05% | 46.06% | 46.02% |
| Subset pozitiv | 46.14% | 46.14% | 46.12% |
| Subset negativ | 45.97% | 45.97% | 45.93% |

Deoarece cuvintele din grupul examinat reprezintă o proporțiile aproximativ identica din cuvintele aflate in seturile si subseturile analizate acestea pot fii extrase fără a influenta substanțial rezultatele modelului dezvoltat.

Eliminarea cuvintelor aparținând categoriei “Stopwords neutre” definite mai sus va reduce dimensiunea setului de antrenament cu aproximativ 46.05%, reducând astfel timpul de antrenament datorita numărului mai mic de cuvinte implicate in procesul de antrenare.

**Stopwords negative :**

In contextul dat vom definii un stopword neutru ca orice cuvânt prezent in următorul sir:

['no', 'nor', 'not', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't’ , 'don', ‘don't’]

**Procentul de Stopwords negative din totalitate cuvintelor din diferitele seturi**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Antrenament | Validare | Testare |
| Set | 1.26% | 1.26% | 1.26% |
| Subset pozitiv | 0.87% | 0.87% | 0.87% |
| Subset negativ | 1.62% | 1.62% | 1.62% |

In urma analizei se poate observa o diferența semnificativa intre proporția cuvintelor din categoria stopwords negative din subseturile pozitive si negative , acestea devenind astfel un nou posibil input pentru antrenarea modelului.

Am decis păstrarea cuvintelor aparținând acestei categorii deoarece in ciuda frecventei lor crescute ele conțin un nivel mare de informație necesara procesului de antrenare a modelului .

## Trunchierea cuvintelor

Trunchierea cuvintelor reprezintă procesul de eliminare a sufixelor si prefixelor cuvintelor cu scopul de a le aduce la forma lor de baza.

Pentru preprocesarea seturilor de date am ales folosirea implementarea algoritmului de trunchiere Porter din modulul **Natural Language Toolkit (NLTK)** pentrulimbajului de programare Python 3.0.

Algoritmul lui Porter a fost dezvoltat și prezentat de Martin Porter în: Porter, M. "An algorithm for suffix stripping." Program 14.3 (1980): 130-137. Avantajele acestuia față de alți algoritmi, cum ar fi algoritmul Lancaster, sunt timpul redus de rulare și simplitatea.

In urma aplicării metodei de trunchiere a cuvintelor asupra setului de date de antrenare am redus numărul de cuvinte unice din acesta de la AICI de cuvinte la AICI , reducând astfel dimensiunea dicționarului necesar pentru transformarea datelor in intrări pentru modelul dezvoltat.

## Eliminarea cuvintelor rar întâlnite

Deoarece setul nostru de date este format din recenzii provenite direct de pe Amazon, fără a fii corectate sau preprocesate in prealabil am decis eliminarea cuvintelor ce se găsesc de mai puțin de 1.000 de ori in cele 40.000.000 de recenzii ale setului de antrenament.

Prin acest proces am scăzut numărul cuvintelor unice la AICI si am redus lungimea medie a unei recenzii la AICI prin eliminarea unor componente ce dăunătoarea sau ce dețineau un nivel scăzut de informație pentru modelul dezvoltat.

## Eliminarea recenziilor de tip „outlier”

In urma aplicării metodelor anterioare de preprocesare au rezultat 5 recenzii cu o lungime mai mare de 128 de cuvinte . Acestea au dimensiunile de AICI, AICI, AICI, AICI si respectiv AICI cuvinte si sunt formate din utilizări repetate ale stopwords-urilor negative „no” sau „don't” .

Prin eliminarea acestora am limitat dimensiunea maxima a recenziilor la 128 de cuvinte, reducând astfel timpul si resursele necesare pentru antrenarea modelului.

## Concluzie

In urma preprocesării datelor am redus substanțial dimensiunea, numărul recenziilor si informațiile dăunătoare modelului precum cuvintele greșite gramatical sau repetările de cuvinte.

**Statistici seturi de date înainte de preprocesare**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cuvinte unice | Lungime maxima | Lungime medie |  |  |  |  |
| Set de antrenare | AICI | AICI | AICI |  |  |  |  |
| Set de testare | AICI | AICI | AICI |  |  |  |  |
| Set de validare | AICI | AICI | AICI |  |  |  |  |

**Statistici seturi de date după preprocesare**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cuvinte unice | Lungime maxima | Lungime medie |
| Set de antrenare | AICI | AICI | AICI |
| Set de testare | AICI | AICI | AICI |
| Set de validare | AICI | AICI | AICI |