**Detectarea spam-ului în SMS**

**Reprezentarea Cunoștințelor,**

**ISI, an 1**

**Cicic Marian Denis**

**Cuprins**

[**1.** **Introducere** 3](#_Toc158395570)

[**2.** **Setul de date folosit** 4](#_Toc158395571)

[**3.** **Implementarea soluțiilor** 5](#_Toc158395572)

[**3.1 Prelucrarea și vizualizarea datelor** 5](#_Toc158395573)

[**3.2** **Antrenarea modelelor** 5](#_Toc158395574)

[**4.** **Rezultate obținute** 6](#_Toc158395575)

[**5.** **Studiu de piață** 6](#_Toc158395576)

[**6.** **Concluzii și direcții viitoare** 7](#_Toc158395577)

[**7.** **Bibliografie** 8](#_Toc158395578)

1. **Introducere**

Contextul tehnologic actual este marcat semnificativ de varietatea căilor de comunicare digitale, foarte prezente în viața de zi cu zi a oamenilor. Comunicarea prin intermediul mesajelor SMS a fost și încă este una dintre modalitățile principale de comunicare, încă de la începutul apariției telefoanelor mobile moderne. Din păcate, sistemul de comunicare prin SMS este amenințat de apariția multor feluri de mesaje spam, conținând oferte sau invitații din partea serviciilor reale, în vederea unei strategii de marketing, ori ajungând să fie folosite în scopuri rău intenționate, pentru a înșela victimele sau a le fura datele.

În general, mesajele spam apar în diferite forme, în text, imagini sau alte formate multimedia care sunt trimise în aplicațiile tradiționale de SMS oferite de majoritatea telefoanelor, dar și în platforme de socializare. O altă modalitate prin care mesajele spam sunt trimise și care are legătură strânsă cu lucrarea de față este sistemul de mail. Acest proiect se va concentra însă pe mesajele spam trimise prin mesaje text scurte, fiind cele mai întâlnite în viața de zi cu zi.

Potrivit Robokiller, o inițiativă care încearcă combaterea spam-ului prin analiza predictivă și amprenta audio, în decembrie 2023, aproximativ 19.2 milioane de mesaje text spam au fost primite de către locuitori, fiind un punct culminant comparativ cu restul anului.

O imagine care conține text, Interval, captură de ecran, diagramă

Descriere generată automat

Fig 1 – Mesaje spam trimise în SUA în 2023 [1]

Același site susține că aproximativ 7,431 mesaje spam au fost trimise per secundă în luna ianuarie a anului 2024, în aceeași țară.

Acest proiect vizează cercetarea algoritmilor de inteligență artificială care pot detecta mesajul spam, pentru a determina care se poate dovedi a fi mai eficient în îndeplinirea acestei sarcini și care algoritm ar fi cel mai potrivit pentru a fi extins în detectarea mesajelor spam din surse mai diversificate.

1. **Setul de date folosit**

După testarea cu seturi de date diverse, seturile de date finale care au fost folosite pentru stabilirea rezultatelor au fost două seturi existente pe platforma Kaggle, și anume:

1. „SMS Spam Dataset” [2] – un set de date cu 5574 intrări, din care 475 de mesaje spam fiind extrase manual din forumul britanic Grumbletext (în prezent închis), 3,375 mesaje non-spam au fost extrase din Secțiunea SMS a Universității Naționale din Singapore, iar alte ~1600 de mesaje au fost încorporate din alte două surse. Mai multe detalii pot fi citite pe arhiva Universității din California [3].
2. „SMS Spam” [4] – având 5156 intrări.

Ambele seturi de date conțin doar două coloane, una menționând felul în care intrarea este clasificată iar cealaltă, textul în sine.

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

Fig 2 – o serie de mesaje din setul de date „SMS Spam”

În ceea ce privește raportul dintre mesajele „Ham” și cele „Spam” – luând în considerare ambele seturi de date, acestea conțin aproximativ 13% mesaje spam, din totalul de aproximativ 10.000 de mesaje.

O imagine care conține text, diagramă, captură de ecran, Interval

Descriere generată automat

Fig 3 – răspândirea lungimii mesajelor, în funcție de tip și frecvență

1. **Implementarea soluțiilor**

## **3.1 Prelucrarea și vizualizarea datelor**

Pentru implementarea algoritmilor a fost folosită aplicația Jupyter Notebook, care reprezintă o modalitate ușoară și rapidă de a scrie cod Python și de a-l rula.

În ceea ce privește abordarea tehnică, pentru integrarea algoritmilor a fost folosită librăria Sci-kit learn.

Pentru manipularea seturilor de date, în vederea încărcării, preprocesării și concatenării lor, a fost folosită librăria Pandas.

Pentru pregătirea textului pentru antrenarea modelului, a fost folosită librăria NLTK pentru tokenizare și eliminarea textului inutil. De asemenea, a fost aplicat un vectorizator TD-IDF pentru convertirea textului într-o matrice de caracteristici.

O imagine care conține text, Font, linie, captură de ecran

Descriere generată automat

Fig 4 – prelucrarea datelor

Pentru a oferi reprezentări vizuale ale datelor au fost Matplotlib și Seaborn, pentru a reprezenta grafic distribuția cuvintelor, lungimea textelor etc.

* 1. **Antrenarea modelelor**

Pentru antrenare au fost folosite mai multe modele de clasificare, precum: Naive Bayes, Random Forest, Linear SVC, Logistic Regression și XGBoost.

Încercări suplimentare de îmbunătățire a rezultatelor prin creșterea complexității algoritmilor au fost încercate doar în cazul XGBoost, unde s-a încercat folosirea Grid Search, însă făcând încercări minime la nivelul optimizării parametrilor, deoarece alți algoritmi s-au dovedit deja a fi suficient de eficienți pentru îndeplinirea așteptărilor.

O imagine care conține text, Font, captură de ecran

Descriere generată automat

Fig 5 – parametrii folosiți pentru Grid Search

1. **Rezultate obținute**

Rezultatele obținute au fost, în mare măsură, cele așteptate. Random Forest s-a clasificat pe primul loc, cu o acuratețe de 0.9955. Cele mai slabe rezultate le-a avut Logistic Regression, deși acestea îndeplinesc așteptările inițiale.

Punctual, clasificarea algoritmilor arată astfel:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algoritm** | **RF** | **L SVC** | **XGB** | **NB** | **LR** |
| **Acuratețe** | 0.9955 | 0.9897 | 0.9892 | 0.9780 | 0.9767 |

O imagine care conține text, captură de ecran, diagramă, Font

Descriere generată automat

Fig 5 – matricea de confuzie dată de XGBoost

1. **Studiu de piață**

Pentru a putea stabili o comparație cu alte proiecte similare, au fost făcut un studiu de piață în vederea găsirii și altor încercări de a implementa algoritmi pentru detectarea mesajelor SMS spam.

Un studiu selectat ca exemplu este cel făcut de Rachel B., intitulat „SMS Spam Detection”, în cadrul Institutului de Tehnologie și Științe din cadrul Universității Karunya din India [5].

Setul de date folosit conținea doar ~5000 de intrări, iar doi dintre algoritmii folosiți în studiul nostru, a fost folosit în studiul în cauză, pe lângă K-NN, SVM și Decision Tree.

Rezultate puțin mai slabe, însă similare au fost obținute, cu SVM clasificându-se pe primul loc, urmat de Naive Bayes și Random Forest.

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

Fig 6 – Rezultatele obținute de Rachel B. [5]

1. **Concluzii și direcții viitoare**

Pentru a concluziona, studiul de față reușește să motiveze necesitatea folosirii inteligenței artificiale în sarcini specifice erei digitale, precum cazul nostru, al detectării SMS-urilor spam.

Rezultatele obținute îndeplinesc așteptările inițiale ale proiectului, însă rămâne de explorat integrarea algoritmilor folosind și alte seturi de date, pentru a fi siguri de funcționarea lor corespunzătoare, pentru a putea exclude potențiale probleme la nivelul datelor, precum existența prea puținor exemple de mesaje spam.

În viitor, proiectul poate fi extins spre detectarea mail-urilor spam, ori a textelor mai lungi. Algoritmii de față s-au dovedit a fi foarte potriviți pentru această sarcină, în special Random Forest, care este cunoscut a avea performanțe crescute și în sarcini de acest gen.

În concluzie, abordarea problemei spam-ului prin machine learning nu numai că reușește să îmbunătățească capacitatea de detectare a mesajelor spam, rău intenționate, dar și deschide calea către dezvoltarea sistemelor de comunicare mai sigure și eficiente, adaptate nevoilor și provocărilor contextului digital actual.

1. **Bibliografie**

|  |  |
| --- | --- |
| [1] | Robokiller, „2023 United States robotext trends,” [Interactiv]. Available: https://www.robokiller.com/spam-text-insights. [Accesat 09 februarie 2024]. |
| [2] | M. Edward, „SMS Spam Dataset,” [Interactiv]. Available: https://www.kaggle.com/datasets/marslinoedward/sms-spam-dataset. [Accesat 02 februarie 2024]. |
| [3] | A. Tiago și H. Jos, „SMS Spam Collection,” [Interactiv]. Available: https://archive.ics.uci.edu/dataset/228/sms+spam+collection. [Accesat 09 februarie 2024]. |
| [4] | S. Kumar, „Sms spam,” [Interactiv]. Available: https://www.kaggle.com/datasets/shravan3273/sms-spam. [Accesat 09 februarie 2024]. |
| [5] | B. Rachel, „SMS Spam Detection,” martie 2021. [Interactiv]. Available: https://www.karunya.edu/aqar/2020-21/QIF/Criteria\_1/1.3.4/568.pdf. [Accesat 09 februarie 2024]. |