**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**MASTER STATISTICĂ APLICATĂ ȘI DATA SCIENCE**



**Aplicarea unor tehnici de NLP pentru detectarea phisingului într-un email**

Profesori coordonatori:

Prof. univ. dr. COSTEA Adrian

Lector univ. dr. OȚOIU Adrian

Student:

Marin Ramona-Alexandra

BUCUREȘTI

**Cuprins**

[**INTRODUCERE 3**](#_Toc154312740)

[**LITERATURA DE SPECIALITATE 4**](#_Toc154312741)

[**METODOLOGIE 5**](#_Toc154312742)

[**SETUL DE DATE 5**](#_Toc154312743)

[**REZULTATE ȘI INTERPRETĂRI 5**](#_Toc154312744)

[**Analiza exploratorie a datelor 5**](#_Toc154312745)

[**Modelul Multinomial Naive Bayes 8**](#_Toc154312746)

[**Aplicarea modelului LSTM (Long Short-Term Memory) 9**](#_Toc154312747)

[**CONCLUZII 10**](#_Toc154312748)

[**BIBLIOGRAFIE 11**](#_Toc154312749)

INTRODUCERE

Phishingul este un atac care încearcă să vă fure banii sau identitatea, determinând persoanele să dezvăluie informații personale, cum ar fi numere de card de credit, informații bancare sau parole pe site-uri web care se prefac a fi legitime. De obicei, criminalii cibernetici pretind că sunt companii, prieteni sau cunoștințe de renume într-un mesaj fals, care conține un link către un site web de phishing.

Phishingul este o formă populară de criminalitate cibernetică datorită cât de eficient este. Infractorii cibernetici au avut succes folosind e-mailuri, mesaje text și mesaje directe pe rețelele sociale sau în jocuri video, pentru a îi determina pe oameni să răspundă cu informațiile lor personale.

Astfel, unele dintre semnele care pot dezvălui faptul că un email poate fi un atac de phishing sunt următoarele:

* erori evidente de ortografie sau gramaticale;
* domenii de e-mail nepotrivite cu greșeli de ortografie foarte subtile ale numelui de domeniu legitim. (exemplu: micros0ft.com, rnicrosoft.com);
* linkuri suspecte sau atașamente neașteptate, majoritatea companiilor nu trimit un mail prin care să solicite parole, informații despre cardul de credit și nici link-uri care cer autentificarea;
* adresarea cu salutări generice precum „Stimate membru stimat”, „Stimate titular de cont” sau „Stimate client”.

În cazul în care o persoană a fost victima unui astfel de atac prezentat anterior, aceasta trebuie să își schimbe parola imediat pentru toate conturile care pot fi afectate, să aibă autentificarea multifactorială (cunoscută și ca verificare în doi pași). Dacă acest atac a afectat conturi de la serviciu sau de la școală, ar trebui să fie notificat personalul de IT despre posibilul atac. Dacă s-au distribuit informații despre cardurile de credit sau conturi bancare, este recomandat să se contacteze companiile pentru a le avertiza cu privire la o posibilă fraudă.

Lumea modernă se confruntă cu mai multe amenințări, inclusiv cea semnificativă a e-mailurilor de phishing, care provoacă uriașe pierderi financiare. Astfel, în această lucrare, mi-am propus să realizez un model de învățare prin tehnii NLP pe textul unor email-uri, având ca obiectiv detectarea email-urilor de phishing pentru reducerea numărului de victim ale tipului acesta de atac.

Motivul alegerii temei îl reprezintă amploarea fenomenului de phishing ca fiind cel mai frecvent atac cybernetic. Într-o companie, angajații se ocupă de email-uri zi de zi, fapt pentru care riscul de a fi victima unei înșelătorii de tip phishing este foarte ridicat.

LITERATURA DE SPECIALITATE

În ultimii ani au fost publicate mai multe articole pentru a oferi diferite cunoștințe esențiale pentru înțelegerea fenomenului de phishing și modul în care este utilizat cu diverse abordări. AI s-a dezvoltat mult, iar acum detectarea e-mailurilor de phishing a adoptat și ML și NLP, contribuind considerabil la combaterea e-mailurilor de tip phishing.

În ciuda actualizării constante a metodelor de evitare a unor astfel de atacuri cibernetice, rezultatul final este în prezent inadecvat. Pe pe de altă parte, e-mailurile de tip phishing au crescut exponențial în ultimii ani, ceea ce sugerează necesitatea unor e-mailuri mai eficiente și mai avansate. Au fost stabilite numeroase metode pentru a filtra e-mailurile de tip phishing, dar problema încă mai are nevoie de a solutie completa.

Articolul *„Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey”* [1], publicat în anul 2021, oferă o analiză a numeroaselor strategii NLP de ultimă generație utilizate în prezent pentru a identifica e-mailurile de phishing în diferite etape ale atacului, cu accent pe ML. Lucrarea prezintă o analiză critică a cercetării despre tehnicile de detectare a e-mailurilor de tip phishing.

Clasificarea e-mailurilor folosind SVM a avut o acuratețe de 87%, iar rețeaua neuronală a avut acuratețea de 90,5%, în schimb, metoda hibridă propusă între cele două metode a obținut o performanță cu o acuratețe de 98%. De-a lungul timpului, pentru detectarea phishingului în email, s-a mai folosit o combinație de analiză a componentelor principale (PCA) și analiză semantică latentă (LSA). Astfel se produc seturi reduse de caracteristici care combinate cu algoritmii XGBoost, duc la 100% rată de succes.

De asemenea, cercetarea a aplicat modele ML distinctive precum „XGBoost, LightGBM și Bernoulli Naive Bayes”, care sunt extrem de rapide și sunt, de asemenea, marcate de imprevizibilitate în timp mai mică. Principalele constatări ale acestui studiu indică faptul că „Bernoulli Naive Bayesurmat de LightGBM cu matricea TF-IDF a produs cea mai mare precizie de 96,5% în 0,157 secunde și 95,4% în 1,708 secunde respectiv”. Pentru a crește performanța, modelele ar fi mai stabile atunci când studiul utilizează modele ML care pot fi antrenate utilizând seturi de date din diverse resurse, precum și seturi de date cu o uriașă număr de documente.

THEMIS este un nou model DL pentru detectarea e-mailurilor de tip phishing. Modelul rulează pe un neural convoluțional mai bun rețea (RCNN) care include vectori pe mai multe niveluri ca și mecanism de atenție, care permite concurența modelarea unui e-mail la nivel de antet, corp, caracter și cuvânt. THEMIS are o precizie de 99,848% conform rezultatelor studiului [1]. Singurul defect al modelului este că nu poate detecta phishing în e-mailurile cu un corp e-mail, dar fără antet e-mail

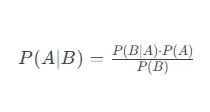
Conform cercetărilor, NLP se caracterizează cu o precizie mai mare în comparație cu alte tehnici atunci când vine vorba de identificarea e-mailurilor de tip phishing; cu toate acestea, NLP nu este încercat până acum pe mai mari seturi de date.

# **METODOLOGIE**

Procesarea limbajului natural (NLP) se referă la ramura informaticii, mai precis, ramura inteligenței artificiale (AI), care oferă computerelor capacitatea de a înțelege textul și cuvintele rostite în același mod în care poate o ființă umană. NLP combină lingvistica computațională, modelarea bazată pe reguli a limbajului uman, cu modele statistice, de învățare automată și de învățare profundă (exemplu sisteme GPS operate de voce, asistenți digitali, software de dictare a vorbirii în text).

Pentru aplicarea tehnicilor de NLP, am ales să folosesc limbajul de programare python în aplicația Jupyter Nootebook din Anaconda Navigator. Limbajul python oferă o gamă largă de instrumente și biblioteci precum biblioteca NLTK folosită pentru analizarea propozițiilor, segmentarea cuvintelor, derivarea și lematizarea (metode de tăiere a cuvintelor până la rădăcini) și tokenizare (pentru ruperea frazelor, propozițiilor, paragrafelor).

În literatura de specialitate, cei mai folosiți algoritmi de clasificare a textelor sunt: algoritmi Naïve Bayes, mașini virtuale de suport (SVM) și învăatrea profundă (deep learning). În această lucrare, am aplicat algoritmul Multinomial Naive Bayes care se bazează pe teorema lui Bayes, formulă matematică care calculează probabilitatea unui eveniment pe baza cunoștințelor anterioare a condițiilor care ar putea fi legate de eveniment.



Clasificatorul atribuie o etichetă de clasă unui document pe baza probabilităților calculate folosind această teoremă. Astfel, orice vector care reprezintă un text va trebui să conțină informații despre probabilitățile de apariție a anumitor cuvinte în textele unei anumite categorii, pentru a calcula probabilitatea ca acel text să aparțină categoriei respective. Naive Bayes este mai potrivit pentru variabilele de intrare categoriale, funcționează rapid și poate economisi mult timp.

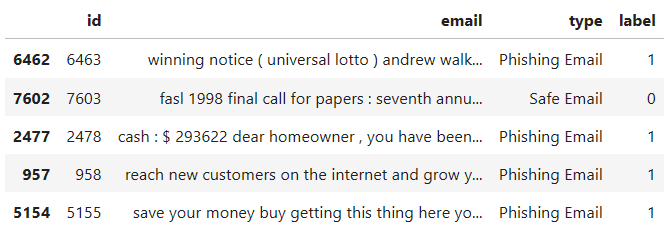
Vectorizarea textului se realizează cu ajutorul lui Count Vectorizer pentru a converti o colecție de documente text într-un vector de numărătoare de termeni/jetoane. În plus, s-a utilizat și algoritmul deep learning GloVe pentru a obține reprezentări vectoriale pentru cuvinte și reușește să îmbunătățească acuratețea clasificatorilor antrenați cu algoritmi tradiționali de învățare automată.

LSTM (Long Short-Term Memory) este o arhitectură de rețea neuronală recurentă (RNN) utilizată pe scară largă în Deep Learning. Excelează în capturarea dependențelor pe termen lung, ceea ce îl face ideal pentru sarcinile de predicție a secvenței.

Avantajele modelelor LSTM constau în: existența unei celule de memorie care este capabilă să stocheze informații pe termen lung, capacitatea modelului să capteze și să rețină contextul important, chiar și atunci când există un interval de timp semnificativ între evenimentele relevante din secvență. Din păcate, antrenamentul LSTM-urilor necesită adesea mai multe date și timpi mai lungi de antrenament pentru a obține performanțe ridicate.

SETUL DE DATE

Pentru a îmi îndeplini obiectivul propus, am uitilzat un set de date preluat de pe site-ul Kaggle.



*Figura 1. Output set de date*

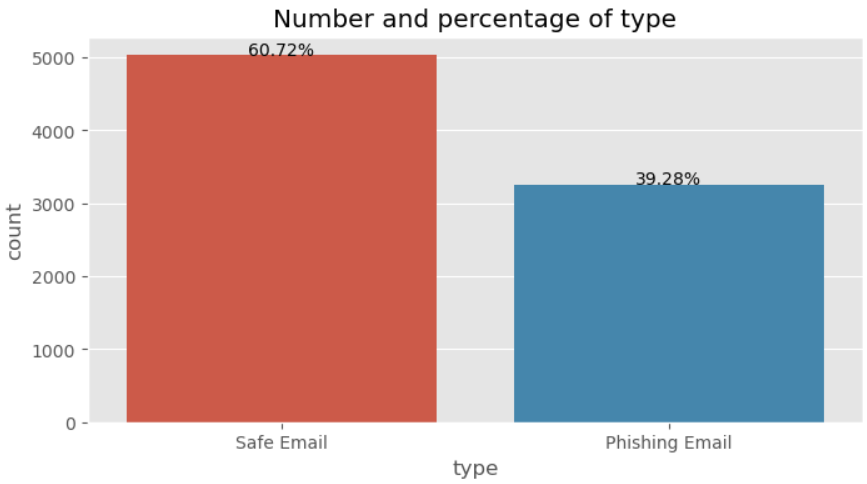
Setul de date conține următoarele variabile:

* id
* email – textul din email
* type – variabilă binară cu variantele: Phishing Email și Safe Email
* label – variabila type a fost codată astfel 1- dacă emailul este un atac de phishing și 0 – dacă emailul nu prezintă un pericol.

Baza de date inițială a fost verificată și curățată de valori nule și valori eronate. Din totalul de 10439 de observații, au fost alocate 8284 de observații setului de date de antrenare și 2155 de observații pentru setul de date de testare.

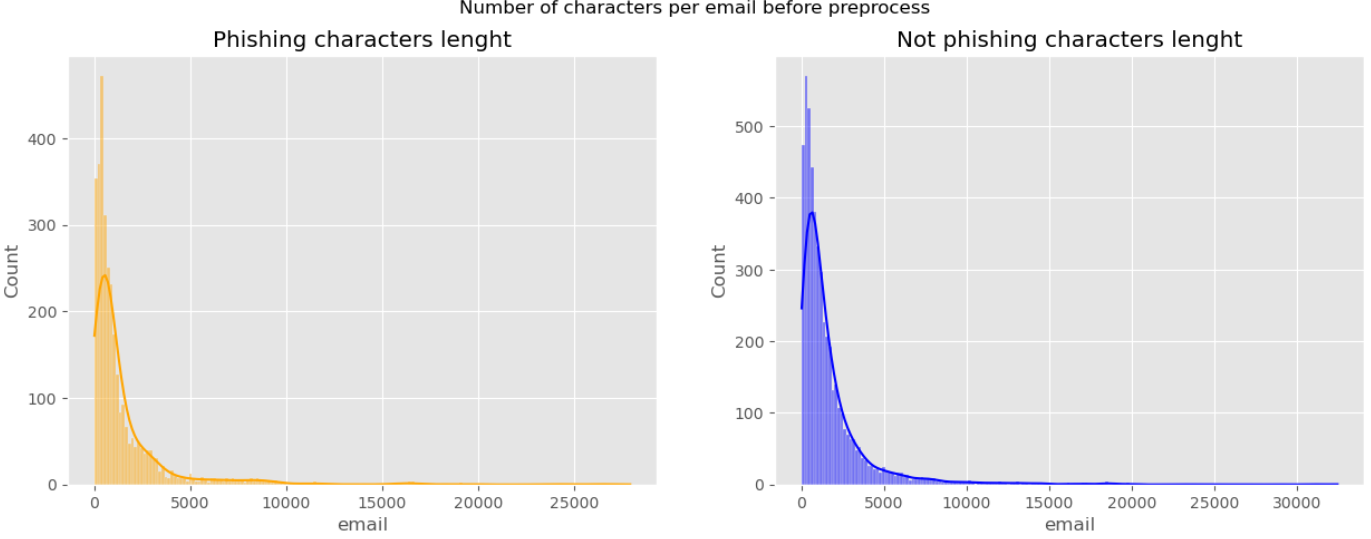
REZULTATE ȘI INTERPRETĂRI

Analiza exploratorie a datelor

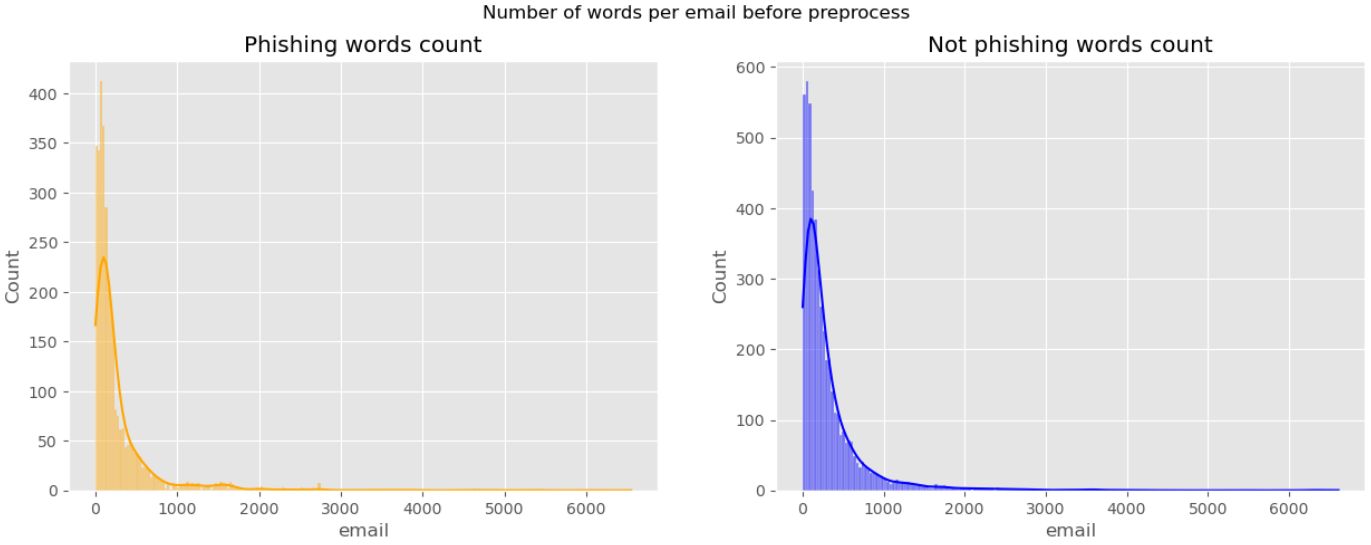


*Figura 2. Distribuția celor două categorii de email*

În figura 2, observăm că din totalul de 8284 de observații, ponderea mai mare, de 60,72%, este reprezentată de categoria email-urilor sigur(Safe email), iar 39,28% aparțin categoriei de email-uri care conțin un atac de phishing.

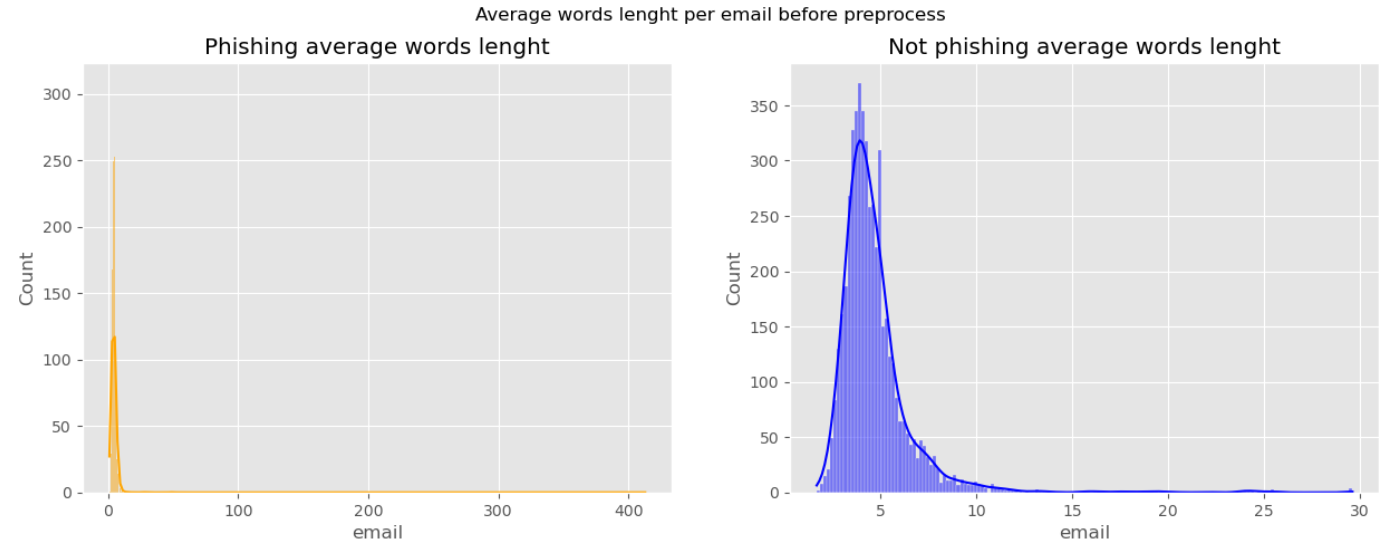


*Figura 3. Distribuția numărului de caractere per email*



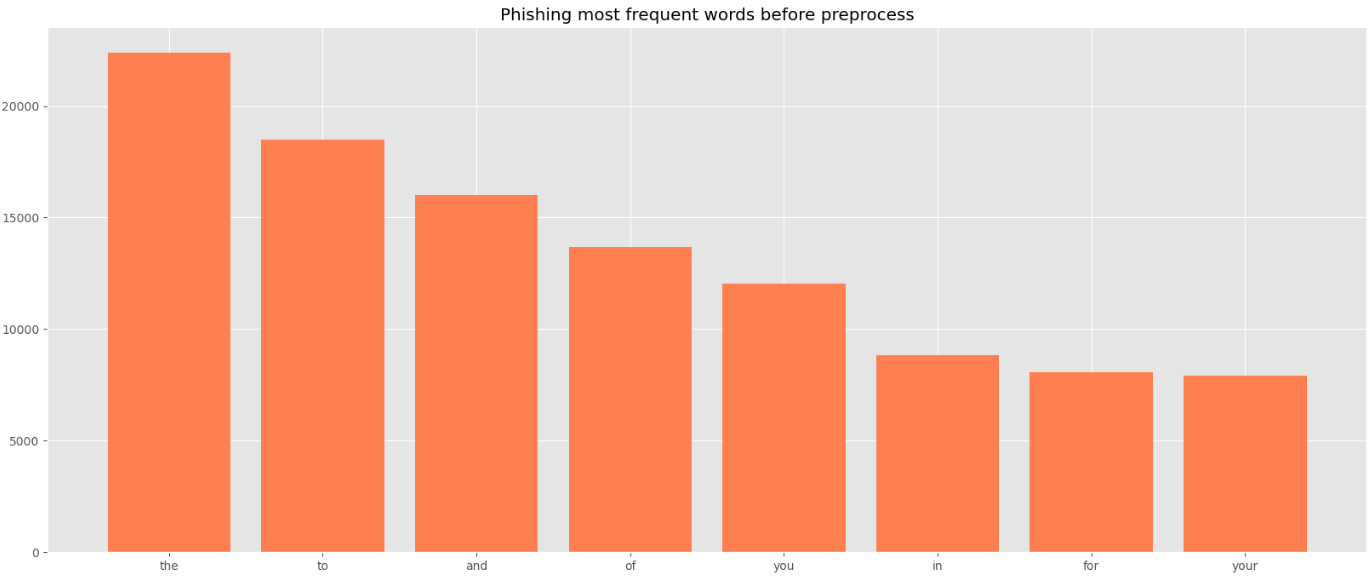
*Figura 4. Distribuția numărului de cuvinte per email*

În figurile 3 și 4, se poate observa asemănarea distribuțiilor dintre cele două categorii de email (phishing și non-phishing) atat din punct de vedere al numărului de caractere per email, cât și al numărului de cuvinte per email.



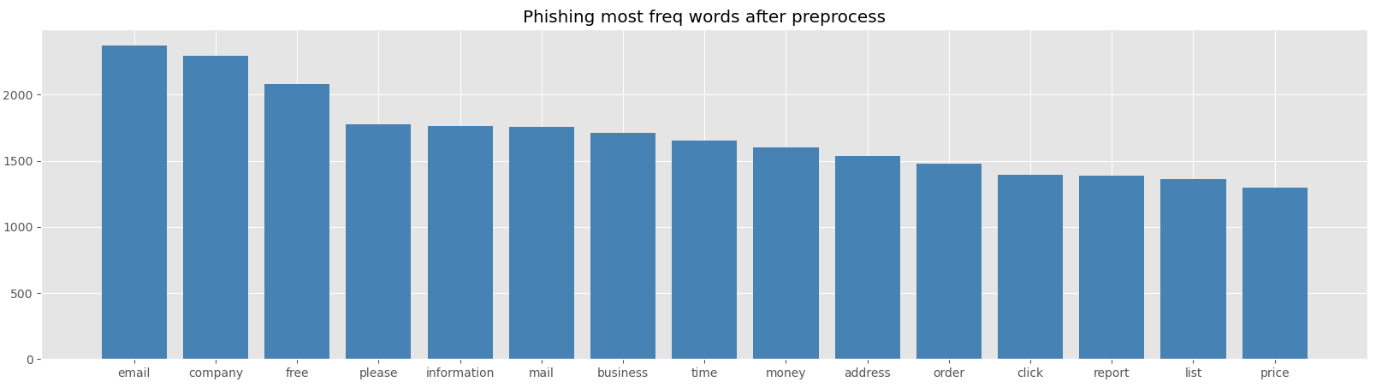
*Figura 5. Distribuția lungimii medie a cuvintelor per email*

În ceea ce privește distribuția lungimii medie a cuvintelor per email, prezentată în figura 5, email-urile de tip phishing tind să conține cuvinte mai scurte decât email-urile sigure. Acest aspect poate fi considerat un semn de detectare al atacului.



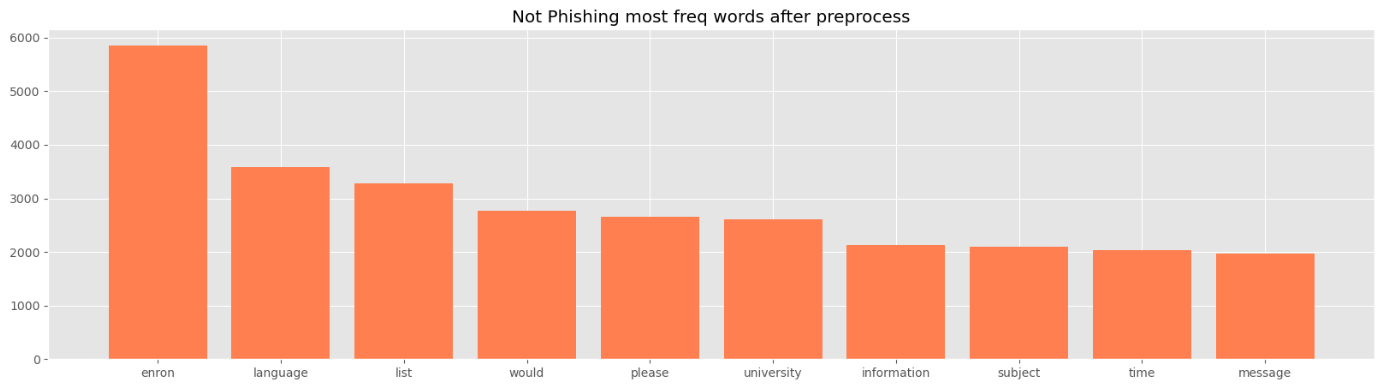
*Figura 6. Cele mai frecvente cuvinte din categoria Phishing Email înainte de preprocesare*

Cuvintele care au cea mai mare frecvență în email-urile de tip Phishing, înainte de preprocesare, sunt cuvinte de legătură precum “the”, “to”,”and”,”of”, “in”. Cel mai utilizat cuvânt este “the” cu o frecvență peste 20 mii de cuvinte.



*Figura 7. Cele mai frecvente 15 cuvinte din categoria Phishing Email după preprocesare*

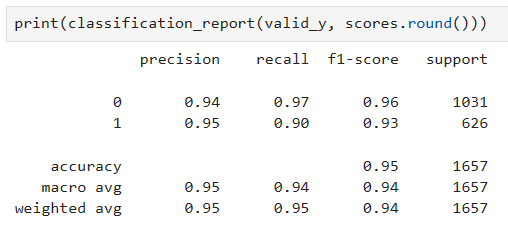
În figura 7, observăm că după procesarea textului, cel mai frecvent cuvânt în email-urile de tip phishing este ”email”, ”company”,“free” cu o frecvență de cel puțin 2000 de apariții. Alte cuvinte frecvente au fost:”please”,”business”,”money”,”address”, ”order”, ”click” și ”price”.

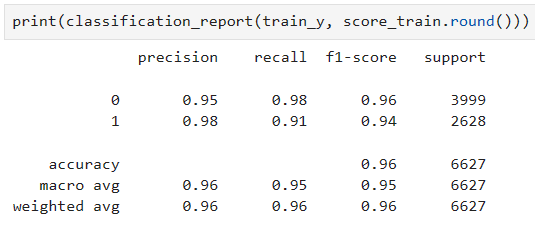


*Figura 8. Cele mai frecvente 10 cuvinte din categoria Safe Email după preprocesare*

După procesarea textului, cel mai utilizat cuvânt pentru email-urile sigure este “enron” cu o frecvență de aproximativ 6000. Restul cuvintelor cu un număr mare de apariții au fost: “language”, ”list”,”would”,“please”,”univeristy”,”information”,”subject”,”time” și ”message”.

Modelul Multinomial Naive Bayes



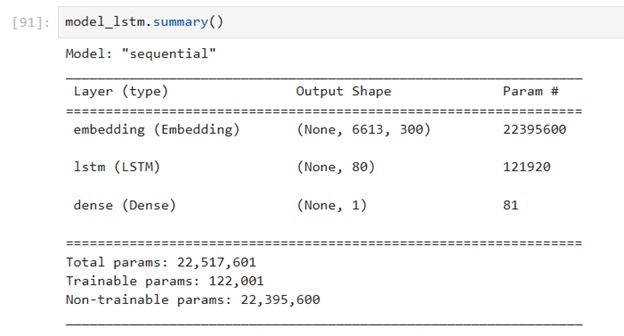


*Figura 9. Matrici - Multinomial Naive Bayes*

Modelul Naïve Bayes, aplicat asupra setului de date de antrenare, prezice corect 94% dintre email-urile care sunt sigure și 95% dintre cele care sunt un atac de tip phishing. Acuratețea modelului este de 95%.

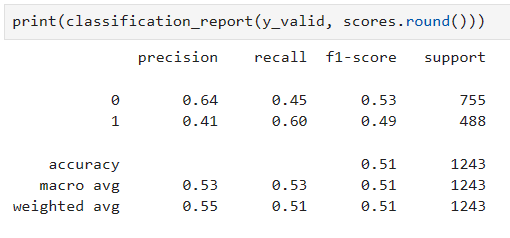
În ceea ce privește modelul Naïve Bayes, aplicat asupra setului de date de testare, prezintă o acuratețe de 96%. Modelul prezice corect 95% dintre email-urile de tip “Safe Email” și în proporție de 98% pe cele de tip “Phishing email.”

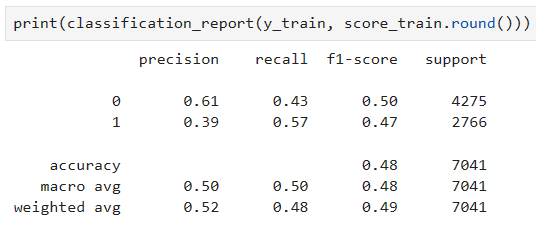
Aplicarea modelului LSTM (Long Short-Term Memory)



*Figura 10. Output – Model LSTM*

Modelul LSTM aplicat are trei straturi: cel de embedding (vectorizarea cuvintelor prin GloVe), modelul în sine și ultimu strat dat de funcția sigmoid cu scopul de activare a neuronilor din rețea care poate lua valori între 0 și 1.





*Figura 11. Matrici – Model LSTM*

În figura 11, observăm că modelul LSTM aplicat asupra setului de date de antrenare prezice corect 64% dintre email-urile sigure și în proporție de 41% prezice corect email-urile care reprezintă un atac phishing. Acuratetea modelului este 51%.

În schimb, aplicând modelul LSTM pe setul de date de testare, obținem o acuratețe de 48%. Modelul prezice corect 61% dintre email-urile de tip “Safe Email” și 39% dintre email-urile de tip “Phishing Email”.

CONCLUZII

În concluzie, rezultatul lucrării nu a fost cel așteptat ca modelul LSTM să prezinte o acuratețe mai mare decât modelului Multinomial Naive Bayes. Per ansamblu, cel din urmă a obținut o acuratețe de 95%-96%, iar modelul LSTM a avut o acuratețe de 48%-51% desi este mai performant. Acest lucru se poate datora faptului că modelul LSTM necesită un set de date cu mai multe observații.

Pentru cercetările viitoare, ar trebui să se urmărească antrenarea unui model LSTM cu o baza de date mult mai mare, desi rețelele LSTM sunt mai scumpe din punct de vedere computațional. De asemenea, conform literaturii de specialitate, o metoda hibridă între două metode poate obține o acuratețe mai ridicată.

BIBLIOGRAFIE

1. K. Agarwal, T.Kumar (2018) - „Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization”
2. Said Salloum, Tarek Gaber, Sunil Vadera, Khaled Shaalan (2021) - “Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey”
3. <https://support.microsoft.com/en-us/windows/protect-yourself-from-phishing-0c7ea947-ba98-3bd9-7184-430e1f860a44>
4. <https://www.ibm.com/topics/natural-language-processing>
5. <https://www.todaysoftmag.ro/article/2463/clasificarea-automata-a-textelor-folosind-tehnici-de-procesare-a-limbajului-natural>
6. <https://www.kaggle.com/datasets/sadikaljarif/global-vectorglove>
7. <https://www.kaggle.com/datasets/subhajournal/phishingemails?select=Phishing_Email.csv>
8. <https://medium.com/@evertongomede/understanding-multinomial-naive-bayes-classifier-fdbd41b405bf>
9. <https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c>
10. <https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/>