UNIVERSITATEA POLITEHNICA BUCUREȘTI

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

DEPARTAMENTUL CALCULATOARE  
  
INGINERIA SISTEMELOR INTERNET

**SISTEME INTERNET INTELIGENTE**

**Detectarea de email-uri spam**

Nechita Ioana-Valeria  
Tudorescu Ana-Maria  
Ciufu Victor Andrei

2024

1. **Soluții existente – State of the Art**

Detecția spamului reprezintă un domeniu vast și esențial în securitatea informațiilor, cu aplicații care se extind dincolo de simpla filtrare a emailurilor. În mod tradițional, detecția spamului a fost esențială pentru platformele de email, dar acum este utilizată pe scară largă în social media, aplicații de mesagerie și platforme de e-commerce pentru a preveni spamul și mesajele nesolicitate. De exemplu, Facebook, Instagram, Twitter și alte rețele sociale integrează detecția spamului pentru a proteja utilizatorii de mesaje de tip „phishing”, fraude și reclame invazive. Aplicațiile de mesagerie, cum ar fi WhatsApp și Messenger, utilizează și ele filtre de spam pentru a bloca încercările de promovare agresivă și distribuirea de linkuri periculoase.

În trecut, majoritatea filtrelor de spam erau bazate pe tehnici de clasificare binară, cum ar fi Naive Bayes și regresia logistică, care se antrenau pe seturi de date etichetate pentru a distinge între mesaje de tip spam și ham (non-spam). Aceste metode, deși utile, au limitări în fața evoluției spamului modern, care include texte mai subtile și mai sofisticate. Algoritmi precum Support Vector Machines (SVM) și Random Forest au adus o îmbunătățire în detecția spamului datorită capacității lor de a analiza caracteristicile complexe ale textului. [1]

Odată cu dezvoltarea rețelelor neuronale profunde, abordările de tip RNN și CNN au câștigat teren. Rețelele neuronale convoluționale (CNN) sunt utilizate frecvent pentru clasificarea imaginilor și sunt acum adaptate pentru a analiza structura textului și pentru a identifica tipare de spam ascunse. RNN, în special LSTM (Long Short-Term Memory), sunt capabile să capteze relațiile temporale și contextuale între cuvinte, fiind extrem de utile în analiza emailurilor mai lungi sau a mesajelor cu secvențe logice complexe. [2]

În prezent, modelele NLP avansate, cum ar fi BERT (Bidirectional Encoder Representations from Transformers) și GPT (Generative Pre-trained Transformer), sunt aplicate în detectarea spamului pentru a înțelege mai bine contextul fiecărui cuvânt și frază din mesaj. Aceste modele sunt antrenate pe milioane de texte și sunt capabile să recunoască subtilitățile limbajului uman, identificând astfel spamul chiar și în situații unde este scris într-un limbaj foarte sofisticat.

Cea mai eficientă soluție modernă de detecție a spamului implică o abordare hibridă, combinând algoritmi de învățare automată tradițională cu rețele neuronale profunde și modele NLP avansate. De exemplu, se poate utiliza un model bazat pe Random Forest pentru a realiza o clasificare preliminară și un model bazat pe BERT pentru o analiză detaliată a conținutului. Această combinație de metode permite o clasificare extrem de precisă și o reducere semnificativă a falsurilor pozitive, fiind eficientă chiar și împotriva spamului sofisticat. [3]

În literatură, detecția spamului este un subiect intens cercetat, iar dataseturi precum Enron Email Dataset, SpamAssassin și Ling-Spam sunt frecvent utilizate pentru antrenarea și testarea modelelor. Enron Email Dataset, de exemplu, conține mii de emailuri dintr-o corporație reală, iar SpamAssassin este un alt dataset utilizat frecvent, cu emailuri marcate explicit ca spam sau ham. Aceste seturi de date sunt citate în numeroase articole științifice, iar ele reprezintă o bază solidă pentru testarea modelelor de clasificare. [4]

Un alt aspect menționat în literatură este provocarea reprezentată de spamul sofisticat, cum ar fi spamul „phishing” și mesaje care încearcă să pară autentice pentru a păcăli utilizatorii. Studiile recente arată că un model bazat pe rețele neuronale poate fi suplimentat de un model NLP pentru a îmbunătăți detecția acestor mesaje înșelătoare. [5]

De asemenea, tehnici precum analiza multivariată și metode de clustering pot fi utilizate pentru a identifica mesaje de spam care au caracteristici similare între ele, ajutând la detecția grupurilor de mesaje false care altfel ar putea fi clasificate greșit. [6]

Modelele de detecție a spamului sunt de obicei evaluate prin metrici precum acuratețea, precizia, „recall” și măsura F1. Evaluarea este esențială pentru a asigura că modelul nu doar identifică corect spamul, dar minimizează și falsurile pozitive, asigurând o experiență bună a utilizatorului. Un model eficient de detecție a spamului ar trebui să mențină o acuratețe ridicată în identificarea spamului, dar să aibă o rată minimă de clasificări greșite ale mesajelor legitime drept spam. [7]

1. **Seturile de date și metode folosite**

Dataset-ul de training este luat din Kaggle datasets (https://www.kaggle.com/datasets/team-ai/spam-text-message-classification). Acest fisier contine doua coloane, o coloana de „label” care este „ham” sau „spam” reprezentand clasificarea email-ului si o coloana de text care este subiectul email-ului. De asemenea, vom folosi un dataset si pentru testare pentru a vedea daca modelul nostru este indeajuns de bun.

Acest dataset trebuie intai formatat: categoriile de „ham” si „spam” trebuie convertite intr-o valoare numerica de 0 sau 1.

Avand in Vedere ca aplicatia noastra implica procesare de text, iar datele nu trebuie analizate rapid, in timp real, putem folosi o metoda mai înceată, dar mai precisa, pentru rezultate cat mai bune.

Ulterior, putem extinde algoritmul pentru a verifica atat textul, cat si adresa/numele celui care a trimis email-ul, de exemplu: daca un sender are observate cateva email-uri de spam, iar rata de spam este de procentaj, putem creste procentajul oricarui email trimis de el. De asemenea, putem verifica timpul la care a trimis email-ul, impreuna cu daca email-ul a fost trimis de mai multe ori.

Variante disponibile de modele: random forest, decision tree, naive bayes (gaussian), neural network with tensorflow sau pytorch.

Vom incerca mai multe modele de invatare supervizata, de la cele mai simple precum Naive Bayes, si putem extinde pentru modele de decision tree sau SVM in functie de performantele obtinute. Putem ulterior sa incercam variante de clustering sau invatare nesupervizata pentru a grupa email-urile in diverse categorii (work, general, spam, junk). [8]

**Bibliografie:**

[1] K.Varun Kumar, M. Ramamoorthy, “Machine Learning-based spam detection using Naïve Bayes Classifier in comparison with Logistic Regression for improving accuracy”, 2022. [Online]. Available:  
https://www.pnrjournal.com/index.php/home/article/download/945/744/1130?\_\_cf\_chl\_tk=nE24WhqTsua8e\_Qj1T\_3m3\_Xvrj.oykR83TrIBngLws-1729866118-1.0.1.1-wzLN5iiO\_NJ.5K5Fo1zYlHIVkY.dbbhoMDOV4dbkDAc.

[2] Pumrapee Poomka, Wattana Pongsena, Nittaya Kerdprasop, Kittisak Kerdprasop, “SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit”, 2019. [Online]. Available:   
https://www.ijfcc.org/vol8/532-CE024.pdf.

[3] Tida Vijay Srinivas, Hsu Sonya, “Universal Spam Detection using Transfer Learning of BERT Model”, 2022. [Online]. Available:  
https://www.researchgate.net/publication/358458191\_Universal\_Spam\_Detection\_using\_Transfer\_Learning\_of\_BERT\_Model.

[4] Maxime Labonne, Sean Moran, “Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection”, 2023. [Online]. Available:  
https://arxiv.org/pdf/2304.01238.

[5] Kutub Thakur, Md Liakat Ali, Muath A. Obaidat, Abu Kamruzzaman, “A Systematic Review on Deep-Learning-Based Phishing Email Detection”, 2023. [Online]. Available:  
https://www.mdpi.com/2079-9292/12/21/4545.

[6] Karim Asif, Azam Sami, Shanmugam Bharanidharan, Kannoorpatti Krishnan, “Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework”, 2020. [Online]. Available:  
https://www.researchgate.net/publication/343709430\_Efficient\_Clustering\_of\_Emails\_Into\_Spam\_and\_Ham\_The\_Foundational\_Study\_of\_a\_Comprehensive\_Unsupervised\_Framework.

[7] Wani Mudasir, Elaffendi Mohammed, Shakil Kashish, “AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing”, 2024. [Online]. Available:  
https://www.researchgate.net/publication/384920638\_AI-Generated\_Spam\_Review\_Detection\_Framework\_with\_Deep\_Learning\_Algorithms\_and\_Natural\_Language\_Processing.

[8] https://www.youtube.com/watch?v=E0Hmnixke2g