Detector de spam folosind inteligenţa artificială

Mihai Petrescu-Miron

Rezumat

e-mail-ul (electronic mail) este o formă de comunicare prin care se trimit mesaje de tip text sau multimedia între utilizatorii unor diferite dispozitive electronice.

Deşi scopul inițial a fost de a trimite mesaje între doi utilizatori de dispozitive electronice conectate la Internet, e-mail-urile pot fi folosite şi ca metodă prin care se pot transmite reclame pentru anumite produse, fişiere infectate cu viruşi care pot compromite date confidențiale şi/sau personale sau furt de date trimise unui număr mare de utilizatori. Acest tip de e-mail se numeşte **spam**.

Această lucrare urmăreşte clasificarea e-mail-urilor în două categorii: *ham* şi *spam*. Pentru a determina cărei clase aparține un e-mail sunt folosite reguli de clasificare (pentru antetul acestuia) şi un algoritm de clasificare (pentru corpul e-mail-ului).

Pentru a determina dacă textul din corpul e-mail-ului conține sau nu caracteristici (en. „*features”*) de tip *spam*, asupra acestuia se efectuează prelucrări pentru a:

- elimina cuvintele care nu au nicio relevanță în determinarea clasei (en. *„stop words”*);
- transforma cuvintele într-o formă cât mai apropiată de cea de dicționar (algoritm de „*stemming*”);
- transforma textul într-o formă vectorială astfel încât să se poată determina pe baza algoritmului de clasificare dacă e-mail-ul este *spam* sau *ham*.

Algoritmul de clasificare utilizat este **regresia logistică**. S-a ales acest algoritm deoarece prin aplicare se obține pentru un anumit e-mail o valoare din intervalul [0, 1]. Astfel, dacă valoarea rezultată este mai mică decât 0,5, atunci e-mail-ul este de tip *ham*, dacă este mai mare, atunci e-mail-ul este de tip *spam*.

Pentru a putea folosi regresia logistică trebuie determinat într-o primă etapă (faza de antrenare) coeficienții funcției logistice. Apoi se observă în următoarea etapă (faza de testare) dacă aceşti coeficienți determină rezultatele aşteptate.

Având toate aceste date, pentru un e-mail nou se verifică regulile pentru antet şi se aplică algoritmul de clasificare cu coeficienții obținuți în faza de antrenare pentru a determina dacă acesta este *spam* sau *ham*.