ВИЯВЛЕННЯ СПАМУ МЕТОДАМИ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ДАНИХ

д-р техн. наук, проф. А.І. Поворознюк, магістр І.І. Литвин, НТУ "ХПІ", м. Харків

На сьогоднішній день, електронні листи вважаються важливим засобом комунікації у світі. Тому вони є вразливими до загроз у вигляді небажаних повідомлень. Спамерам не складає труднощів розсилати безліч повідомлень мільйонам користувачів. Звідси випливає необхідність розпізнавання спаму [1 – 3]. Однією із стратегій виявлення спаму є використання методів інтелектуального аналізу даних.

Процес виявлення спаму зазвичай включає в себе етапи обробки та аналізу текстової інформації з електронного листа. На початку виконується попередня обробка вмісту листа – це очищення тексту, видалення закінчень слів, заміна слів на їх початкову форму. Далі застосовується токенізація - розбиття змісту листа на окремі слова і перетворення його в послідовність за кількістю повторень кожного унікального слова. Наступним кроком є побудова моделі за допомоги алгоритму класифікації. Для покращення результатів набір даних розбивається на дві окремі вибірки: навчальну та тестову. Модель тренуємо на навчальному наборі. Після чого оцінюємо її на тестовому наборі даних з точки зору точності та повноти.

В даній роботі було проведено експеримент з розпізнаванням спаму з використанням запропонованого підходу. Для цього використано набір електронних листів Spam Email Dataset [1], робота виконувалась за допомоги мови програмування Python, для токенізації використовувалась бібліотека TfidfVectorizer, а для класифікації листів використовувався алгоритм логістичної регресії.

Результатом застосування запропонованого підходу стало успішне створення класифікаційної моделі для виявлення спаму та результати її застосування з точністю (0.96) та повнотою (0.94).

Список літератури: **1.** Spam Email Dataset [Електронний ресурс] // Режим доступу www URL: https://www.kaggle.com/datasets/mfaisalqureshi/spam-email. **2.** *N. Kumar, S. Sonowal and Nishant*, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113. **3.** *Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa*, Machine learning for email spam filtering: review, approaches and open research problems, Heliyon, Volume 5, Issue 6, 2019, e01802, ISSN 2405-8440.