УДК: 004.8:004.94

М. В. Білецька
С. О. Жуков

ВИЯВЛЕННЯ СПАМУ В ПОВІДОМЛЕННЯХ ЗА ДОПОМОГОЮ НАЇВНОГО БАЄСІВСЬКОГО КЛАСИФІКАТОРА

Вінницький національний технічний університет

Анотація

Робота присвячена дослідженню та застосуванню наївного баєсівського класифікатора для виявлення спаму в текстових повідомленнях. Проведено аналіз ефективності алгоритму на наборі даних з повідомленнями Telegram, визначено найвпливовіші маркери спаму та досліджено точність класифікації. Розроблена модель демонструє високу ефективність у ідентифікації небажаних повідомлень.

Ключові слова: наївний баєсівський класифікатор, спам, машинне навчання, обробка тексту, класифікація повідомлень, аналіз даних.

Abstract

The paper is devoted to the research and application of the naive Bayesian classifier for spam detection in text messages. The effectiveness of the algorithm was analyzed on a dataset of Telegram messages, the most influential spam markers were identified, and classification accuracy was investigated. The developed model demonstrates high efficiency in identifying unwanted messages.

Keywords: naive Bayesian classifier, spam, machine learning, text processing, message classification, data analysis.

Вступ

Проблема виявлення спаму залишається актуальною в сучасному інформаційному суспільстві. Масові рекламні та шахрайські повідомлення створюють істотне навантаження на комунікаційні канали та знижують ефективність електронної комунікації. Наївний баєсівський класифікатор представляє собою простий, але потужний алгоритм машинного навчання, що базується на теоремі Баєса та застосовується для вирішення задач класифікації текстових повідомлень.

Унікальність підходу полягає в тому, що класифікатор аналізує частоту появи окремих слів у повідомленнях різних типів і на основі цього визначає ймовірність належності нового повідомлення до категорії спаму. Незважаючи на "наївне" припущення про незалежність ознак, цей метод демонструє високу ефективність у задачах фільтрації небажаних повідомлень.

Дане дослідження спрямоване на розробку та оцінку моделі наївного баєсівського класифікатора для виявлення спаму на основі набору даних повідомлень Telegram, а також аналіз ключових характеристик спам-повідомлень.

Методологія дослідження

Для проведення дослідження використано набір даних "Telegram Spam or Ham" з платформи Kaggle, що опублікована користувачем Mexwell [1]. Він містить 20348 повідомлень, класифікованих як спам (spam) або звичайні повідомлення (ham). Набір даних включає два основні атрибути: тип повідомлення (text_type) та текст повідомлення (text).

У роботі використано наївний баєсівський класифікатор для класифікації спаму. Реалізацію виконано в Kaggle з використанням бібліотек scikit-learn, NLTK, Pandas і Matplotlib [2–4]. Повна версія ноутбука доступна за посиланням [5].

Попередня обробка текстових даних включала приведення тексту до нижнього регістру, токенізацію, видалення неалфавітних символів та стоп-слів, лематизацію та видалення дублікатів слів у межах одного повідомлення. Ці кроки дозволяють зменшити варіативність тексту та сфокусуватись на змістовних елементах повідомлень.

Для навчання та тестування моделі набір даних було розділено у співвідношенні 80% на 20%. Навчальна вибірка містила 11424 звичайних повідомлень та 4854 спам-повідомлень, а тестова - загалом 4052 повідомлення. Такий поділ забезпечує достатню кількість даних для навчання моделі, зберігаючи при цьому репрезентативну частину для оцінки її ефективності.

Реалізація наївного баєсівського класифікатора базувалася на розрахунку апріорних ймовірностей для класів спаму та звичайних повідомлень, підрахунку частоти входження кожного слова в повідомлення різних класів та застосуванні згладжування Лапласа для уникнення нульових ймовірностей. Функція Bayes() була розроблена для класифікації нових повідомлень, використовуючи логарифмічний масштаб для запобігання чисельній нестабільності при роботі з дуже малими ймовірностями. Цей підхід дозволяє ефективно визначати належність повідомлень до категорії спаму чи звичайних повідомлень на основі статистичних закономірностей у тексті.

Аналіз набору даних показав, що 70,5% повідомлень класифіковано як звичайні (ham), а 29,5% - як спам. Така незбалансованість класів характерна для задач виявлення спаму і була врахована при розробці моделі. На рисунку 1 представлено розподіл повідомлень за класами: Ham (не спам) і Spam (спам). Видно, що кількість повідомлень Ham значно перевищує кількість повідомлень Spam, що свідчить про незбалансованість набору даних. Така візуалізація необхідна для попереднього аналізу, щоб оцінити, чи може дисбаланс класів вплинути на роботу класифікатора. Модель може виявляти упередженість у бік численнішого класу при значній різниці в кількості зразків між класами. У таких випадках доцільно застосувати техніки балансування даних для забезпечення більш справедливого та точного процесу навчання.

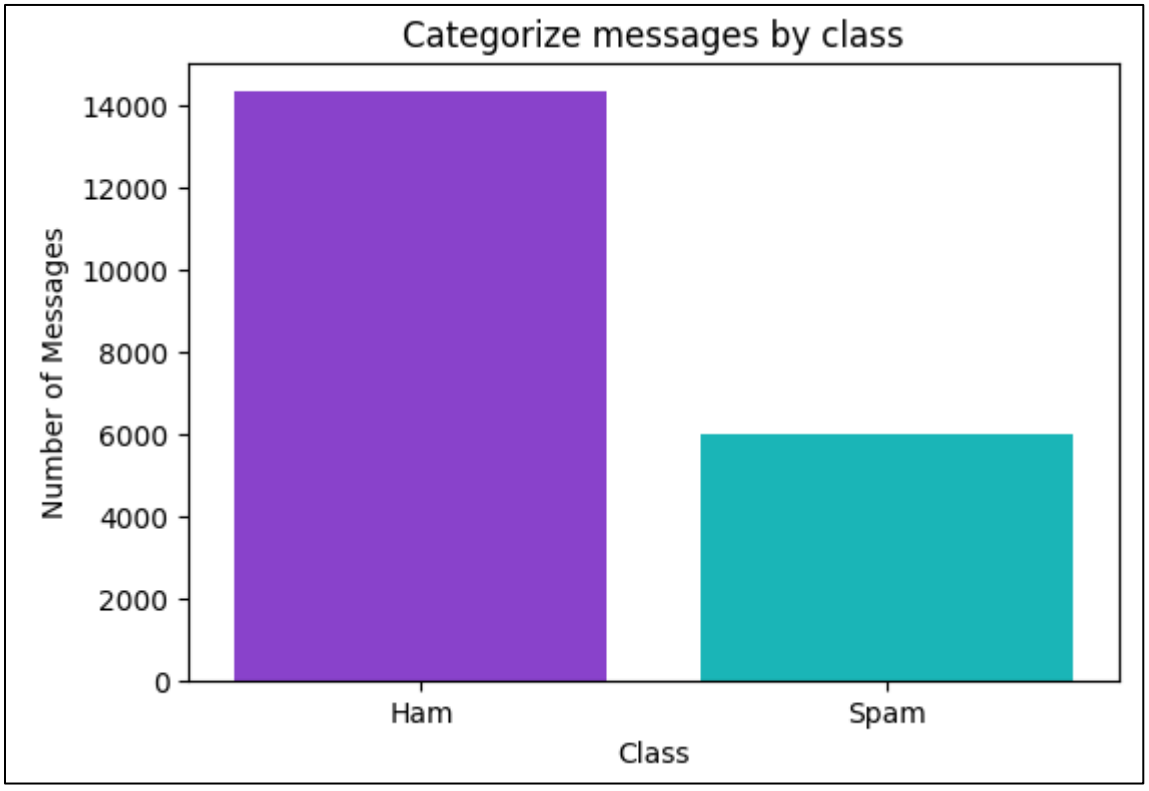

Рис. 1. Категоризація повідомлень за класами

Рисунок 2 демонструє кругову діаграму, яка показує співвідношення між повідомленнями «хам» (не спам) та повідомленнями «спам». Вона показує, що 70,5% повідомлень - це «хам», тоді як лише 29,5% - спам. Ця візуалізація дозволяє швидко оцінити баланс класів у наборі даних. Видимий дисбаланс вказує на те, що класифікатор може мати тенденцію частіше передбачати більш представлений клас, тому важливо враховувати цю нерівномірність при побудові моделі та вживати заходів для її компенсації, якщо це необхідно.

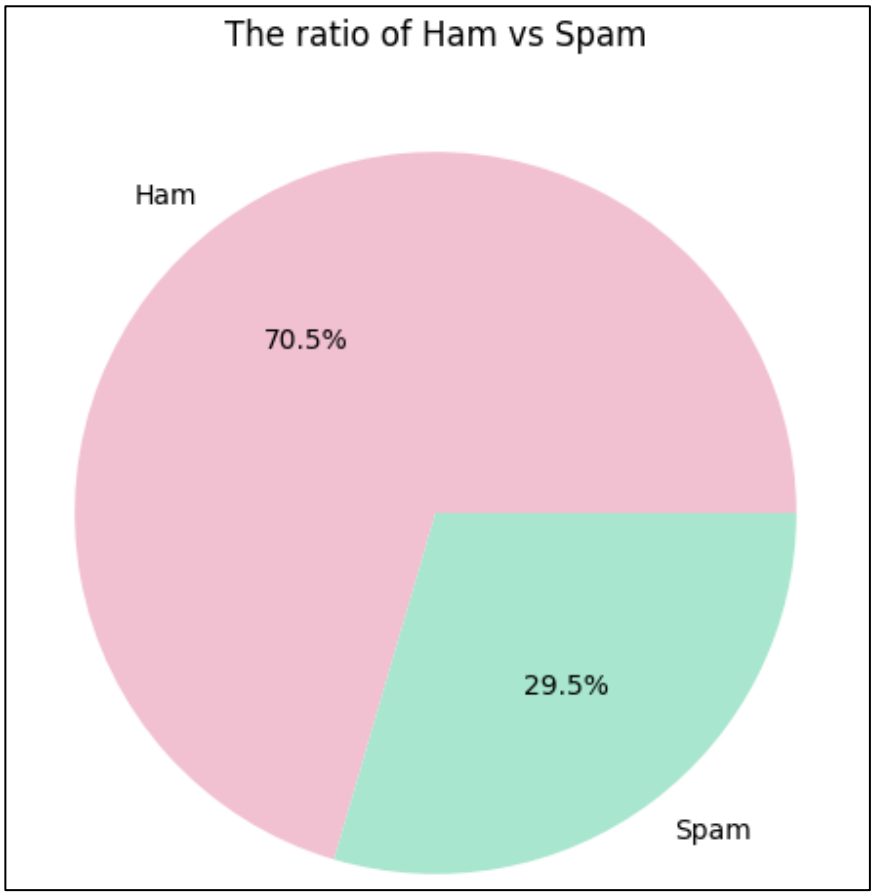

Рис. 2. Графік розподілу цільової величини якості сну (Quality of Sleep)

У результаті аналізу частотності слів у спам-повідомленнях було виявлено 23267 унікальних слів. Найбільш часто в спамі зустрічаються маркетингові терміни та заклики до дії, такі як "get", "free", "u", "new", "call", "link", "one", "click", "best" та "offer". Ці слова мають найвищі показники ймовірності появи у спам-повідомленнях (від 0,0056 до 0,0030), що відображає типові стратегії спамерів привернути увагу користувачів через обіцянки безкоштовних пропозицій, знижок та спонукання до негайних дій. Ці результати підтверджують, що спам-повідомлення часто містять слова, пов'язані з маркетинговими пропозиціями, знижками та закликами до дії.

Результати функціонування наївного баєсівського класифікатора, представлені на рисунку 3, демонструють значну ефективність алгоритму при ідентифікації спам-повідомлень. Оцінювання проводилось на двох різних тестових прикладах: один було сформовано вручну, а другий обрано випадковим чином із наявного набору даних. В обох випадках класифікатор продемонстрував високу точність, зокрема визначивши перше повідомлення як спам з максимальною впевненістю у 100%.

```
🔍 Test email (manual): Exclusive offer just for you! Get 50% off on all our products. Hurry u
p, limited time only!
Email is SPAM with confidence of 100.00%

🔍 Test email (random з test_emails): desire celebrated affectionsamp tradition ideal occasion
reflection ur christmas value
Email is SPAM with confidence of 92.18%
```

Рис. 3. Результати роботи наївного алгоритму байєсівського класифікатора

Перший лист – це типовий спам, в якому пропонується значна знижка (50%) на товари з обмеженим терміном дії. Лист вимагає від одержувача термінових дій, що є характерною ознакою спаму. Через

надмірний маркетинговий тиск та обіцянки великих знижок цей лист класифікується як спам з максимальною впевненістю (100%).

Друге повідомлення складається з емоційно забарвлених та святкових слів, які можуть асоціюватися з привітаннями або маркетинговими кампаніями, зокрема до Різдва («Різдво», «привід», «цінність», “бажання”, «святкувати»). Формулювання виглядають узагальненими і не мають чіткого таргетингу або змісту, що часто трапляється в рекламних або автоматично згенерованих листах. Через це повідомлення було класифіковано як спам з високою ймовірністю 92,18%.

Загалом, байєсівський алгоритм продемонстрував свою здатність ефективно класифікувати навіть за наявності різних форматів спаму. Такі результати підтверджують ефективність наївного баєсівського підходу для класифікації електронних повідомлень, навіть при роботі з різними форматами спаму. Алгоритм успішно використовує умовні ймовірності та припущення про незалежність ознак для створення потужного класифікатора, який може стати основою для автоматичних систем фільтрації небажаних повідомлень з високою точністю.

Висновки

У результаті проведеного дослідження розроблено модель наївного баєсівського класифікатора для виявлення спаму в текстових повідомленнях. Дослідження підтвердило, що навіть базуючись на відносно простих математичних принципах, наївний баєсівський класифікатор забезпечує високу точність у виявленні небажаних повідомлень. Попередня обробка тексту, включаючи токенізацію, видалення стоп-слів та лематизацію, значно підвищує якість класифікації. Типові маркери спаму включають слова, пов'язані з комерційними пропозиціями ("free", "offer"), закликами до дії ("get", "call", "click") та створенням відчуття терміновості.

Застосування згладжування Лапласа та логарифмічного масштабу при обчисленні ймовірностей дозволяє уникнути проблем з нульовими ймовірностями та переповненням. Розроблений алгоритм може ефективно виявляти спам у різних форматах та з різним контекстом, що робить його цінним інструментом для систем фільтрації повідомлень.

Отримані результати мають практичне значення для розробки систем захисту від спаму в меседжерах та електронній пошті. Подальші дослідження можуть включати розширення набору даних, покращення методів попередньої обробки тексту та порівняння ефективності наївного баєсівського класифікатора з іншими методами машинного навчання.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Telegram Spam or Ham. 2024 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/mexwell/telegram-spam-or-ham
2. Naive Bayes Classifiers. 2025 [Електронний ресурс]. – Режим доступу: https://www.geeksforgeeks.org/naive-bayes-classifiers/
3. Рассел С., Норвіг П. Штучний інтелект: сучасний підхід. — Київ: Видавнича група BHV, 2022. — 1408 с.
4. Matplotlib Pyplot Documentation. 2025 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
5. Naive Bayesian classifier. 2025 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/code/mariana65/naive-bayesian-classifier/notebook?scriptVersionId=236035831

Білецька Мар’яна Володимирівна – студентка групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: marjnabilecka@gmail.com

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Biletska Mariana – student of Faculty of Intelligent Information Technologies and Automation, SA-22b, Vinnytsia National Technical University, Vinnytsia, e-mail marjnabilecka@gmail.com

Zhukov Serhii O. – Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com