DOI: https://doi.org/10.26565/2304-6201-2024-64-03

УДК (UDC) 004.8

Волинець Катерина Андріївна *студентка ННІ комп'ютерних наук та штучного інтелекту, Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022*
e-mail: kateryna.volynets@student.karazin.ua
https://orcid.org/0009-0003-7661-9758

Стрілець Вікторія Євгенівна *к.т.н., доцент кафедри комп'ютерних систем та робототехніки, ННІ комп'ютерних наук та штучного інтелекту, Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022*
e-mail: viktoria.strilets@karazin.ua
https://orcid.org/0000-0002-2475-1496

Яковлев Данило В'ячеславович *студент ННІ комп'ютерних наук та штучного інтелекту, Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022*
e-mail: yakovlev2020ki11@student.karazin.ua
https://orcid.org/0009-0005-4785-6361

Модель класифікації спам-повідомлень у медичних інформаційних системах

Актуальність. У сучасних медичних інформаційних системах щодень генерується значна кількість текстових записів від пацієнтів, лікарів та персоналу. Для якісної роботи такі систем потребують впровадження моделей і методів аналізу і класифікації текстових даних, зокрема виявленню спамових повідомлень і їх блокуванню. Тому розробка, удосконалення і впровадження моделей і методів класифікації спам-повідомлень є актуальним завданням.
Мета дослідження: покращення якості моделей класифікації спам-повідомлень, що дозволить з високою вірогідністю виявляти і фільтрувати небажані повідомлення у медичних інформаційних системах.
Методи дослідження: методи обробки природної мови, моделювання, машинне навчання, методи класифікації, методи аналізу даних, статистичні методи.
Результати. Були побудовані моделі класифікації спам-повідомлень із використанням таких методів машинного навчання як модель логістичної регресії, модель наївного Баєйсівського класифікатора та модель опорних векторів. Для навчання моделей використаних набір SMS Spam Collection, попередньо підготовлений із використанням CountVectorizer та TF-IDFVectorizer. Усі запропоновані моделі показали високу точність у класифікації спам-повідомлень.
Висновки: розроблені моделі класифікації повідомлень на основі машинного навчання та nlp-підходу успішно визначають небажані повідомлення. Кращою за показниками якості виявилася модель на основі методу опорних векторів з TF-IDF векторизацією, оскільки вона показала найвище значення точності (98.75%) та високу повноту (90.3%) класифікації. Подальші вдосконалення моделей та розширення навчального набору можуть сприяти подальшому покращенню якості розпізнавання спаму.

Ключові слова: *спам-повідомлення, медичні інформаційні системи, машинне навчання, обробка природної мови, класифікація текстових даних.*

Як цитувати: Волинець К. А., Стрілець В. Є., Яковлев Д. В. Модель класифікації спам-повідомлень у медичних інформаційних системах. *Вісник Харківського національного університету імені В.Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2024. вип. 64. С.25-31. https://doi.org/10.26565/2304-6201-2024-64-03

How to quote: Volynets K. A., Strilets V. Y., Yakovlev D. V. "The spam-messages classification model in a medical information system". *Bulletin of V.N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 64, pp.25-31, 2024. https://doi.org/10.26565/2304-6201-2024-64-03[in Ukrainian]

1 Вступ

У сучасному інформаційному суспільстві обробка великих обсягів текстових даних стає все більш важливим завданням для багатьох галузей, зокрема у сфері медичних послуг. Щодня генерується значна кількість текстових повідомлень у медичних інформаційних системах (МІС), включаючи запити пацієнтів, медичні записи та результати аналізів. Для ефективної роботи таких систем необхідно впровадження автоматизованих методів аналізу та класифікації текстових даних.

Одним із потужних інструментів для вирішення цієї задачі є методи обробки природної мови (NLP), які дозволяють не тільки класифікувати текстові повідомлення, але й виявляти спам або інші нерелевантні дані. На відміну від штучних мов, таких як мови програмування та математичні нотації, природні мови розвивалися в міру переходу від покоління до покоління, і їх важко визначити чіткими правилами [1].

Застосування NLP у медичних інформаційних системах дозволяє не тільки автоматизувати обробку запитів, але й забезпечити високу точність та швидкість аналізу даних, що сприяє поліпшенню якості медичних послуг. Проте, залишається низка проблем, пов'язаних із специфікою медичних текстів, таких як складність медичної термінології та необхідність врахування контексту. Це підкреслює актуальність досліджень у галузі класифікації текстових даних з використанням NLP.

2 Задачі обробки природної мови

Серед ключових праць у галузі NLP для медицини важливе місце займають роботи, присвячені автоматичному узагальненню або виокремленню інформації з електронних медичних записів, що сприяє покращенню управління даними пацієнтів [2]. NLP-моделі в цих роботах показали високу точність у задачах автоматизації обробки результатів аналізів та призначень лікарів.

Однією з основних задач NLP у МІС є класифікація текстових даних, таких як запити пацієнтів і лікарські рекомендації. Методи машинного навчання, як показано у [3], демонструють високу ефективність при класифікації великих обсягів тексту, але потребують ретельної підготовки даних через складність медичної термінології.

Дослідження [4] зосередили увагу на використанні методів опорних векторів (SVM) і байєсових класифікаторів для класифікації медичних повідомлень, підкреслюючи як переваги цих підходів, так і труднощі, зумовлені неоднозначністю медичної мови.

Останні наукові досягнення, такі як [5], продемонстрували ефективність трансформерних моделей, таких як BERT, для класифікації медичних текстів завдяки їхній здатності враховувати контекст. Проте, проблеми, пов'язані з обробкою медичних текстів, які часто містять абревіатури та спеціальні терміни, залишаються актуальними [6].

У роботі розглядається *задача* створення моделі класифікації спам-повідомлень, які надходять до медичних інформаційних систем, із використанням методів машинного навчання, які добре зарекомендували себе в цьому напрямку.

Метою роботи є покращення якості моделей класифікації спам-повідомлень, що дозволить з високою вірогідністю виявляти і фільтрувати небажані повідомлення у медичних інформаційних системах.

3 Методи класифікації текстових даних

Серед методів машинного навчання для розв’язання класифікації текстових даних застосовують такі методи:

1. Метод опорних векторів (SVM) є одним з найпопулярніших методів машинного навчання для класифікації тексту. Він працює шляхом пошуку оптимальної гіперплощини, яка найкращим чином розділяє дані у просторі ознак. SVM добре справляється з наборами даних з великою кількістю ознак та може ефективно працювати з текстовими даними;

2. Байєсівські методи класифікації використовують теорему Байєса для прогнозування класу тексту на основі ймовірностей. Найпоширенішими варіантами байєсівських методів є наївний байєсівський класифікатор (Naive Bayes Classifier), який вважає, що всі ознаки у тексті незалежні між собою при умові класифікації. Цей метод простий у реалізації та зазвичай добре працює для наборів даних з великою кількістю ознак;

3. Нейронні мережі є потужними інструментами для класифікації тексту. Вони можуть ефективно використовуватися для вирішення різноманітних завдань, таких як класифікація

тональності, визначення сутностей тощо. Два основних типи нейронних мереж, які часто використовуються для класифікації текстів:

– згорткові нейронні мережі (CNN) використовуються для аналізу послідовностей даних, таких як текст. Вони здатні автоматично виявляти важливі ознаки у тексті, що робить їх ефективними для класифікації;

– рекурентні нейронні мережі (RNN) працюють з послідовностями довільної довжини та зберігати інформацію про попередні стани. Це робить їх добрим вибором для аналізу тексту, оскільки вони здатні врахувати контекст інформації.

Для дослідження були обрані такі методи машинного навчання, як: логістична регресія, наївний байєсівських класифікатор та метод опорних векторів.

Оскільки у дослідженні розглядається задача класифікації тестових повідомлень, то для створення ефективних моделей варто застосовувати також лексичні методи NLP. Лексичні методи зосереджені на аналізі і обробці окремих слів або токенів у тексті. Вони базуються на лексичних властивостях мови, таких як частота вживання слів, визначення частин мови тощо. До основних лексичних методів відносять:

- токенізацію (tokenization) – процес розбиття тексту на окремі слова або токени. Більшість моделей обробки текстових даних спирається на попередньо виокремлені або позначені слова з тексту, що аналізується [3];

- лематизацію (lemmatization) – процес перетворення слова на його базову форму або лему. Наприклад, слово "running" буде перетворено на "run";

- стемінг (stemming) – відсічення афіксів зі слова, залишаючи лише його корінь. Наприклад, слово "beginning" може бути зведене до "begin".

4 Розробка моделей класифікації спам-повідомлень

4.1 Збір і попередня обробка набору даних

Під час збору та обробки медичних даних для класифікації виникають специфічні проблеми, пов'язані з медичними текстами. Наприклад, медичні дані часто містять спеціалізовані терміни, абревіатури та скорочення, що ускладнює автоматичне розпізнавання та класифікацію. Неоднозначність медичної термінології та її чутливість до контексту вимагає більш ретельної попередньої обробки, включаючи лематизацію та видалення стоп-слів.

Для класифікації текстових даних важливим етапом є підготовка якісного набору даних. Для класифікації спам-повідомлень можна використати SMS Spam Collection Dataset [8], що містить записи текстових повідомлень, позначених як "спам" або "не спам" (рис. 1).

	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Рисунок 1 – Фрагмент набору даних

Під час попередньої обробки були використані:

– очищення даних – видалення непотрібних стовпців, видалення пустих записів, перевірка наявності пропущених значень;

– лематизація – приведення слів до початкової форми, видалення стоп-слів, токенізація [7] (розбиття тексту на слова), перетворення тексту в нижній регістр для узгодження.

Також попередньо була проаналізована збалансованість між класами, оскільки нерівномірний розподіл даних між класами може впливати на якість побудованих моделей. Виявилося, що у

наборі даних спостерігається нерівномірний розподіл між класами: 4825 повідомлень позначені як «ham» (не спам) і 747 як «spam» (спам) (рис. 2).

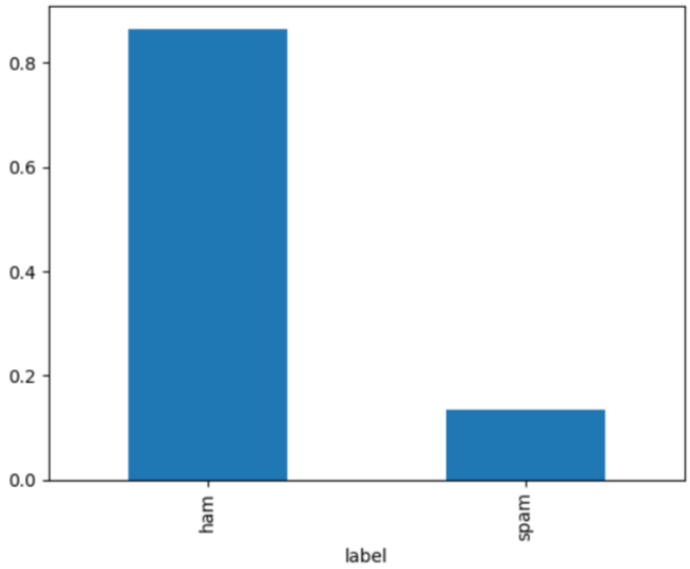

Рисунок 2 – Розподіл класів

Нерівномірність розподілу класів враховувалась під час побудови моделей та аналізі їх якості.

4.2 Створення наборів ознак

Для побудови моделей машинного навчання текстові дані необхідно перетворити у числові вектори. У досліджені були використані:

– CountVectorizer, який перетворює текст у матрицю кількостей, де кожен стовпець відповідає слову з тексту, а кожен рядок — документу (текстовому повідомленню). Кожен елемент матриці вказує кількість разів, коли слово зустрічається у документі;

– TF-IDF (Term Frequency-Inverse Document Frequency) – підхід, який враховує не лише частоту слова у документі, але й його значущість у наборі текстів. Він забезпечує більш точне представлення тексту для класифікації.

За кожним з методів векторизації було побудовано набір даних для навчання і тестування моделей класифікації.

4.3 Створення моделей класифікації та їх налаштування

Для класифікації текстових даних було побудовано декілька моделей машинного навчання: модель логістичної регресії, модель наївного Баєйсівського класифікатора та модель опорних векторів (SVM). Кожна модель була навчена на двох різних наборах ознак: з CountVectorizer та з TF-IDFVectorizer.

Процес налаштування моделей включав підбір гіперпараметрів, таких як регуляризаційні параметри для логістичної регресії та параметр ядра для SVM. Налаштування здійснювалось за допомогою крос-валідації або пошуку по сітці.

5 Аналіз якості побудованих моделей класифікації

Результати на тестовому наборі даних є основною метрикою оцінки продуктивності моделі. Проте також важливо оцінювати результати на навчальному та валідаційному наборах для перевірки наявності перенавчання або недонавчання. Розмір тренувальної і тестової вибірок був 3733 і 1839 відповідно.

Якість моделей на тестовому наборі була оцінена за метриками accuracy, precision, recall (табл. 1).

Таблиця 1. Оцінки якості моделей класифікації

Модель	Accuracy	Precision	Recall
Логістична регресія з CountVectorizer	0.98205	0.99038	0.86919
Логістична регресія з TF-IDF	0.96628	0.99435	0.74261
Наївний Байєс з CountVectorizer	0.98586	0.95670	0.93248
Наївний Байєс з TF-IDF	0.96954	0.99453	0.76793
Метод опорних векторів з CountVectorizer	0.98096	0.97641	0.87341
Метод опорних векторів з TF-IDF	0.98749	1.0	0.90295

Аналізуючи отримані значення показників можна побачити, що модель на основі методу опорних векторів з TF-IDF має найвищий показник точності (98.75%) і precision (100%), показник recall (90.30%) є одним з кращих порівняно з іншими моделями.

Усі розроблені моделі були перевірені на розпізнавання спаму для конкретних повідомлень (рис. 3, 4).

```
Enter message for classificationIs that seriously how you spell his name?
The message is classified as ham.
```

Рисунок 3 – Визначення повідомлення як такого, що не є спамом

```
Enter message for classificationHi! Do you want to earn money? Just call me and claim your reward!
The message is classified as spam.
```

Рисунок 4 – Визначення повідомлення як спам

Протестовані моделі показали здатність до вірної класифікації спам-повідомлень. Але, оскільки початкових набір містив незбалансовані класи, то точність визначення саме спаму буде залежати від подібності аналізованих повідомлень до зразків навчального набору.

6 Розробка альтернативних моделей

Оскільки отримані результати показали високу якість моделей класифікації, було прийняте рішення також побудувати моделі на основі ансамблевих методів Gradient Boosting Classifier та Random Forest Classifier.

Отримані результати (рис. 3) свідчать про те, що моделі Gradient Boosting та Random Forest мають дуже високу якість, але їхні показники різняться в залежності від типу векторизації (CountVectorizer або TF-IDF).

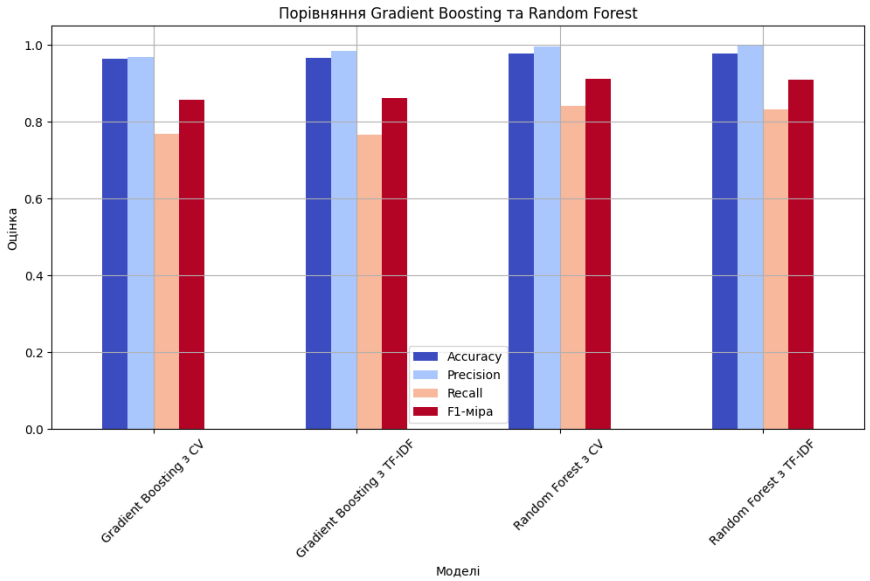

Рисунок 3 – Порівняння якості ансамблевих моделей

Random Forest з TF-IDF демонструє найкращі результати з точки зору точності, але порівнюючи результати, можна зробити висновок, що модель на основі методу опорних векторів (SVM) з TF-IDF показує трохи кращі результати.

Gradient Boosting та Random Forest зазвичай краще працюють на великих наборах даних, тому модель на основі методу опорних векторів з TF-IDF є найефективнішою для задачі класифікації з невеликою кількістю спам-повідомлень.

7 Висновок

У результаті дослідження були побудовані моделі класифікації на основі методів машинного навчання, які мають високі показники якості. Було успішно реалізовано процес класифікації текстових повідомлень з використанням моделей на основі логістичної регресії, наївного Баєйсівського класифікатора та метоуд опорних векторів (SVM). Застосування таких інструментів, як CountVectorizer і TF-IDF, дозволило точно та швидко перетворити текстові дані у числові ознаки, що сприяло підвищенню якості класифікації. Оцінка моделей за метриками якості показала, що метод опорних векторів у поєднанні з TF-IDF демонструє найвищу ефективність.

Моделі на основі CountVectorizer мали вищу F1-міру в порівнянні з моделями на основі TF-IDF:

- F1-міра для CountVectorizer: 0.9211;
- F1-міра для TF-IDF: 0.8208;
- Середня точність перехресної валідації (CountVectorizer): 0.9769.

Це свідчить про те, що CountVectorizer забезпечує більш збалансовану продуктивність між прецизійністю та реколом, що може бути важливим у реальних умовах для уникнення надмірного кількості помилкових позитивних або негативних результатів.

REFERENCES

1. Steven Bird, Ewan Klein, Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media Inc. 2009. 504 p.
2. Wang, M., Sun, Z., Jia, M. et al. Intelligent virtual case learning system based on real medical records and natural language processing. *BMC Med Inform Decis Mak* 22, 60 (2022). https://doi.org/10.1186/s12911-022-01797-7.
3. Robert M. Cronin, Daniel Fabbri, Joshua C. Denny, S. Trent Rosenbloom, Gretchen Purcell Jackson. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*, Vol. 105, P. 110-120, 2017. https://doi.org/10.1016/j.ijmedinf.2017.06.004
4. Elbattah M., Arnaud É., Gignon M., Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies* (*BIOSTEC 2021*), Volume 5: HEALTHINF, P. 825-832, 2021. DOI: 10.5220/0010414508250832.
5. Turchin Alexander, Masharsky Stanislav, Zitnik Marinka. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36, 2022. DOI: 101139. 10.1016/j.imu.2022.101139.
6. Zhou Binggui, Yang Guanghua, Shi Zheng, Ma Shaodan. Natural Language Processing for Smart Healthcare. *IEEE Reviews in Biomedical Engineering*, 2021. DOI: 10.48550/arXiv.2110.15803.
7. Jurafsky Daniel, Martin James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition. Prentice Hall, 2019, 621 p. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (Last accessed: 20.11.2024)
8. Almeida T., Hidalgo J. SMS Spam Collection [Dataset]. *UCI Machine Learning Repository*. 2011. https://doi.org/10.24432/C5CC84.

Volynets Kateryna *student of Education and Research Institute of Computer Sciences and Artificial Intelligence, V.N. Karazin Kharkiv National University, 6 Svobody sq., Kharkiv, Ukraine, 61022*

Strilets Viktoriia *Ph.D, associate professor of the Department of Computer Systems and Robotics, Education and Research Institute of Computer Sciences and Artificial Intelligence, V.N. Karazin Kharkiv National University, 6 Svobody sq., Kharkiv, Ukraine, 61022*

Yakovlev Danylo *student of Education and Research Institute of Computer Sciences and Artificial Intelligence, V.N. Karazin Kharkiv National University, 6 Svobody sq., Kharkiv, Ukraine, 61022*

The spam-messages classification model in a medical information system

Relevance. In modern medical information systems, a significant number of text records are generated daily from the service, doctors and staff. For high-quality work, such systems require the implementation of models and methods for analyzing and classifying text data, in particular, detecting spam messages and blocking them. Therefore, the development, improvement and implementation of models and methods for classifying spam messages is a relevant task.
Research objective: increasing the efficiency of the spam message recognition process in medical information systems; developing and implementing spam classification models based on machine learning methods.
Research methods: natural language processing methods, modeling, machine learning, classification methods, data analysis methods, statistical methods.
Results. Spam message classification models were built using such machine learning methods as the logistic regression model, the national Bayesian classifier model and the support vector model. The SMS Spam Collection set, previously prepared using CountVectorizer and TF-IDFVectorizer, was used to train the models. All proposed models showed high accuracy in spam message classification and the ability to correctly determine the type of message.
Conclusions: The developed message classification models based on machine learning and nlp approach successfully generated unwanted messages. The best model for quality indicators was the model based on the support vector method with TF-IDF vectorization, after which it showed the highest accuracy value (98.75%) and high value of recall (90.3%) of classification. Further improvements of the models and expansion of the training set can contribute to further improvement of the quality of spam recognition.

Keywords: *spam messages, medical information systems, machine learning, natural language processing, text data classification.*

Надійшла у першій редакції 27.08.2024, в останній - 05.10.2024.
The first version has been received on 27.08.2024, the final version - on 05.10.2024.