Міністерство освіти і науки України
Національний технічний університет України «Київський Політехнічний Інститут ім. І. Сікорського»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Дипломна робота на тему:

Порівняння деяких методів машинного навчання для аналізу текстової інформації

Виконала:
Камінська О.М., КА-31
Науковий керівник:
к. ф.-м. н., доц. Каніовська І. Ю.

Дипломна робота

- **Об’єкт дослідження** – алгоритми машинного навчання для аналізу текстової інформації.
- **Предмет дослідження** – застосування методів аналізу текстової інформації в задачі класифікації спаму.
- **Мета роботи** – порівняти деякі методи машинного навчання при аналізі текстової інформації на прикладі задачі класифікації спаму.

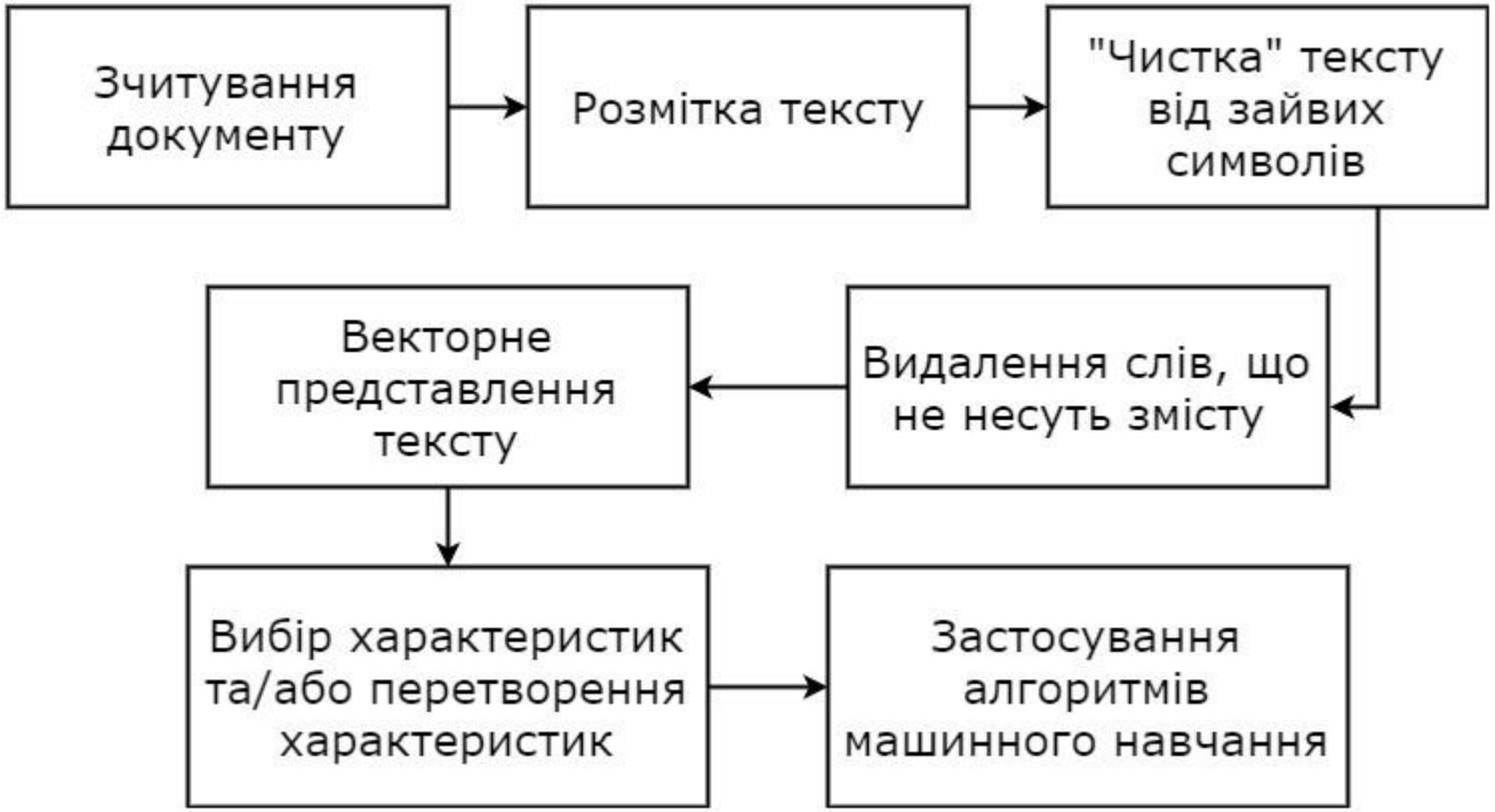
Актуальність задачі

- Технологія глибинного аналізу текстів здатна «просіювати» великі обсяги неструктурованої інформації, такої як звіт чи повідомлення, і виявляти в них лише цінну інформацію, щоб людина не витрачала час на проведення аналізу «вручну». Саме тому пошук оптимальних методів машинного навчання для розв'язання різних типів задач аналізу текстової інформації є актуальним напрямом досліджень.

Постановка задачі

- Розглядається задача класифікації, що визначає, які повідомлення з набору повідомлень є спамом, а які ні.
- Порівнюються кілька методів машинного навчання на прикладі розв'язання задачі класифікації повідомлень «**спам – не спам**», виявляються їх переваги та недоліки та вибираються найбільш оптимальні методи.

Процес класифікації тексту



Перетворення тексту у числові ознаки

- Міра **TF-IDF** (англ. TF — term frequency, IDF — inverse document frequency) — статистичний показник, що використовується для оцінки важливості слів у контексті документа.

- $\text{TF-IDF} = \text{TF} \cdot \text{IDF} = \frac{n_i}{\sum_k n_k} \cdot \lg \frac{D}{d_i}$

- де n_i — число входжень слова в документ;
  $\sum_k n_k$ — загальна кількість слів в документі;
  D – кількість документів колекції;
  d_i – кількість документів, в яких зустрічається слово i.

Методи, що досліджувалися

- Метод логістичної регресії
  (logistic regression method, **LR**);
- Метод k найближчих сусідів
  (k-nearest neighbor method, **kNN**);
- Метод опорних векторів
  (Support Vector Machine, **SVM**);
- Метод дерев рішень
  (decision trees method, **DT**);
- Наївна Байєсівська класифікація
  (naive-Bayes approach, **NBA**).

Вибір параметрів тренування

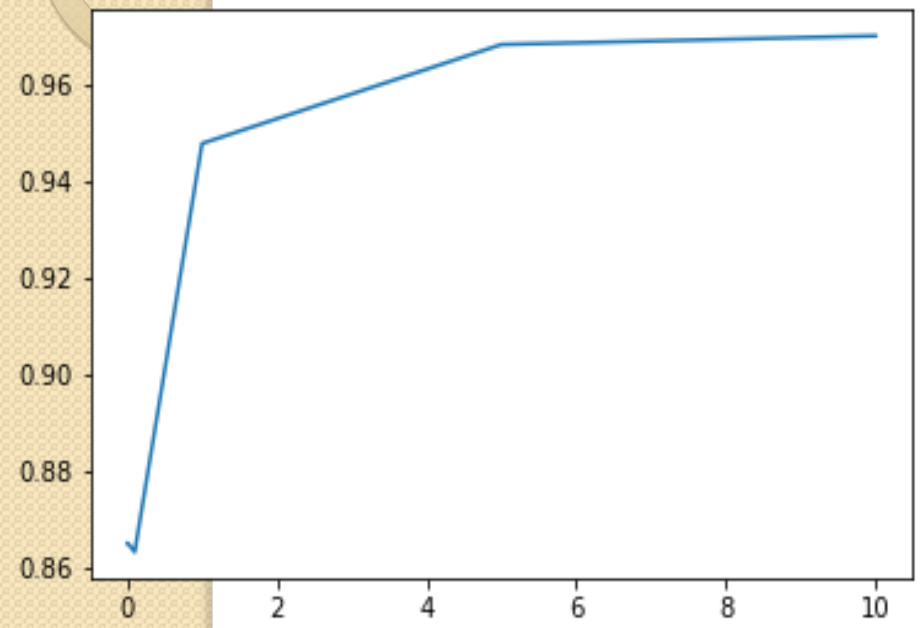
- **LR**:
  - *Penalty* (укр. «*функціонал втрат*») – визначення норми, яка застосовується для обчислення похибки.
  - *C* – параметр згладжування.
- **kNN**:
  - *k* – кількість сусідів, які використовуються в алгоритмі.
- **SVM**:
  - *Kernel* (укр. «*ядро*») – тип ядра, яке використовується в алгоритмі.
- **DT**:
  - *k* – максимальна глибина дерева рішень.
- **NBA**:
  - α - додатковий параметр згладжування.

Критерії оцінювання методів

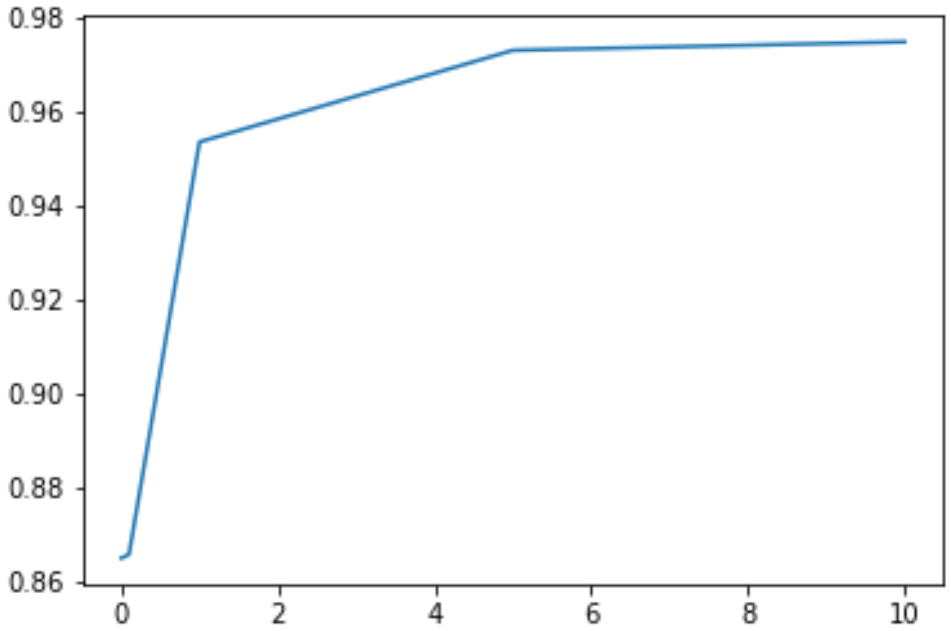
- Час виконання алгоритму;
- Значення точності;
- Похибки першого і другого роду (False positive & True negative).

Всі значення критеріїв обчислюються при тих значеннях ключових параметрів кожного методу, що забезпечують його найбільшу точність.

Отримані графіки залежностей

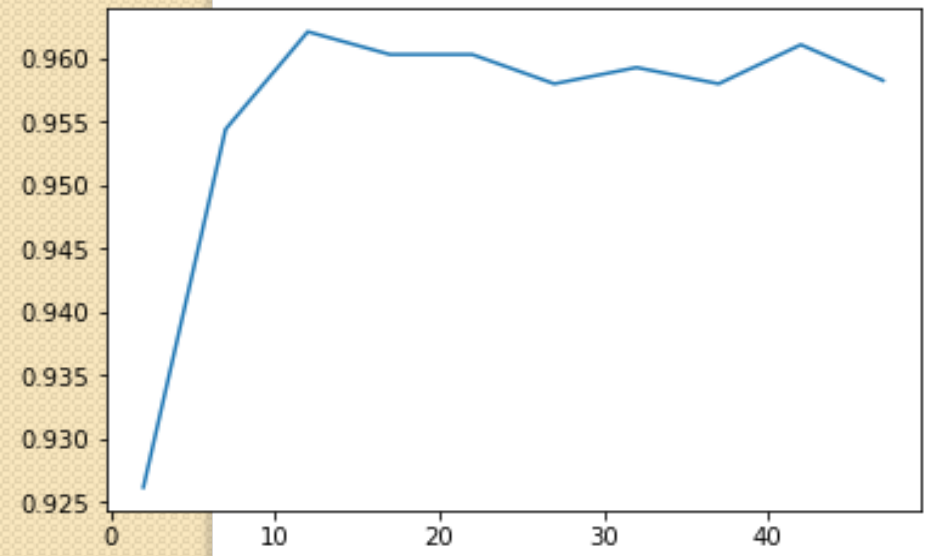


Залежність значення точності від значення C для LR (penalty = l1)

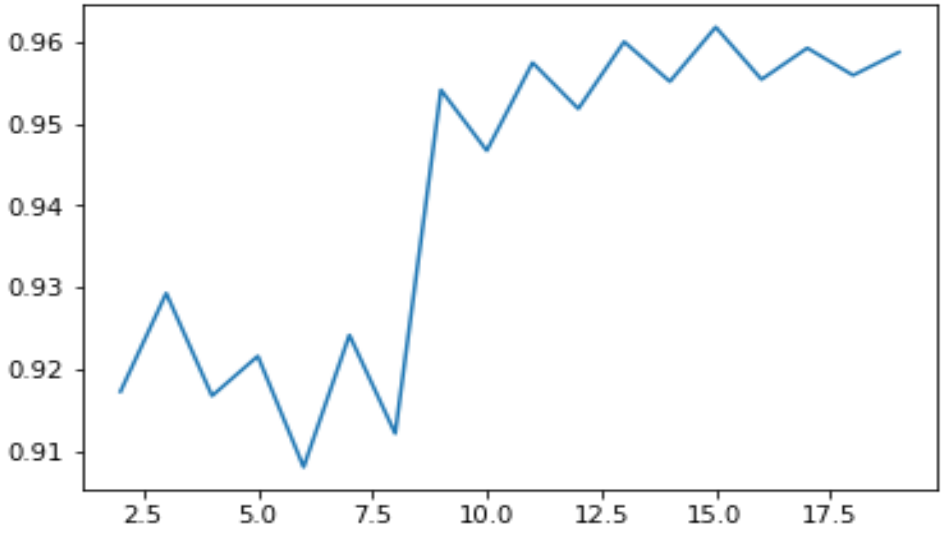


Залежність значення точності від значення C для LR (penalty = l2)

Отримані графіки залежностей

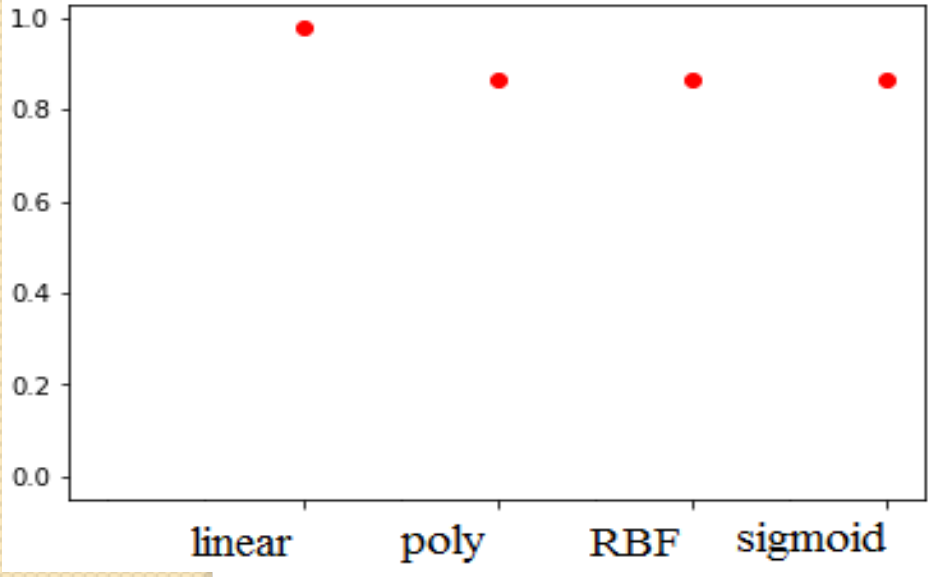


Залежність значення точності від значення **C** для **DT**

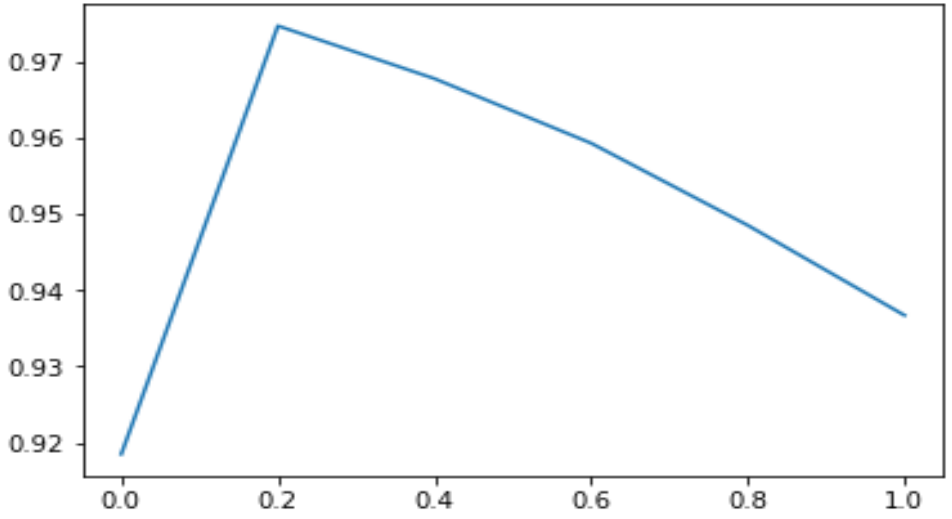


Залежність значення точності від значення **k** для **kNN**

Отримані графіки залежностей



Залежність значення точності від значення **kernel** для **SVM**



Залежність значення точності від значення **α** для **NBA**

Результати обчислень

Назва методу	Найкраще значення параметра	Точність при найкращому параметрі	Час виконання (с)	False positive	True negative	True positive	False negative
LR	C = 10 penalty = l1	0.9700	0.160	0.0	0.02451	0.81448	1.0
	C = 10 penalty = l2	0.9746					
kNN	k = 15	0.9618	0.655	0.00119	0.04065	0.69231	0.99862
DT	C = 12	0.9621	28.355	0.0	0.02451	0.81448	1.0
SVM	kernel = linear	0.9774	1.592	0.00059	0.02032	0.84615	0.99931
NBA	α = 0.2	0.9746	0.034	0.01016	0.01614	0.87783	0.98829

Аналіз результатів

- Найкращу точність і значення True negative показали метод опорних векторів та Байєсівської класифікації.
- Найменш ефективними методами виявилися методи дерева рішень та k-найближчих сусідів – вони обидва дають найменші значення точностей.
- Метод логістичної регресії дає посередні результати для обох значень параметру «penalty» по всім критеріям.
- Порівнюючи усі отримані результати, можна зробити висновок, що **метод Байєсівської класифікації** є найбільш оптимальним і ефективним для задачі класифікації спаму.

Висновки по роботі

- Описано основні поняття області, алгоритми роботи з методами машинного навчання та наведені основні методи, з їх програмною реалізацією.
- Проведено порівняння цих методів аналізу текстової інформації і виділено недоліки та переваги кожного з них.
- Проаналізовано доцільність використання методів при розв’язанні певних задач машинного навчання.
- Для розв'язання задачі класифікації спаму, згідно з результатами досліджень, рекомендується використовувати метод Байєсівської класифікації.

Подальші шляхи розвитку

- В перспективі є можливість подальшого розвитку даної роботи і використання найбільш оптимальних методів на підприємстві та в особистих цілях для автоматичної обробки текстової інформації й економії ресурсів і часу користувача.

Дякую за увагу!