Створення додатку для класифікації тексту та визначення його основних тез на основі попередніх досліджень та порівняльного аналізу

Виконала: Онищенко Є.А.

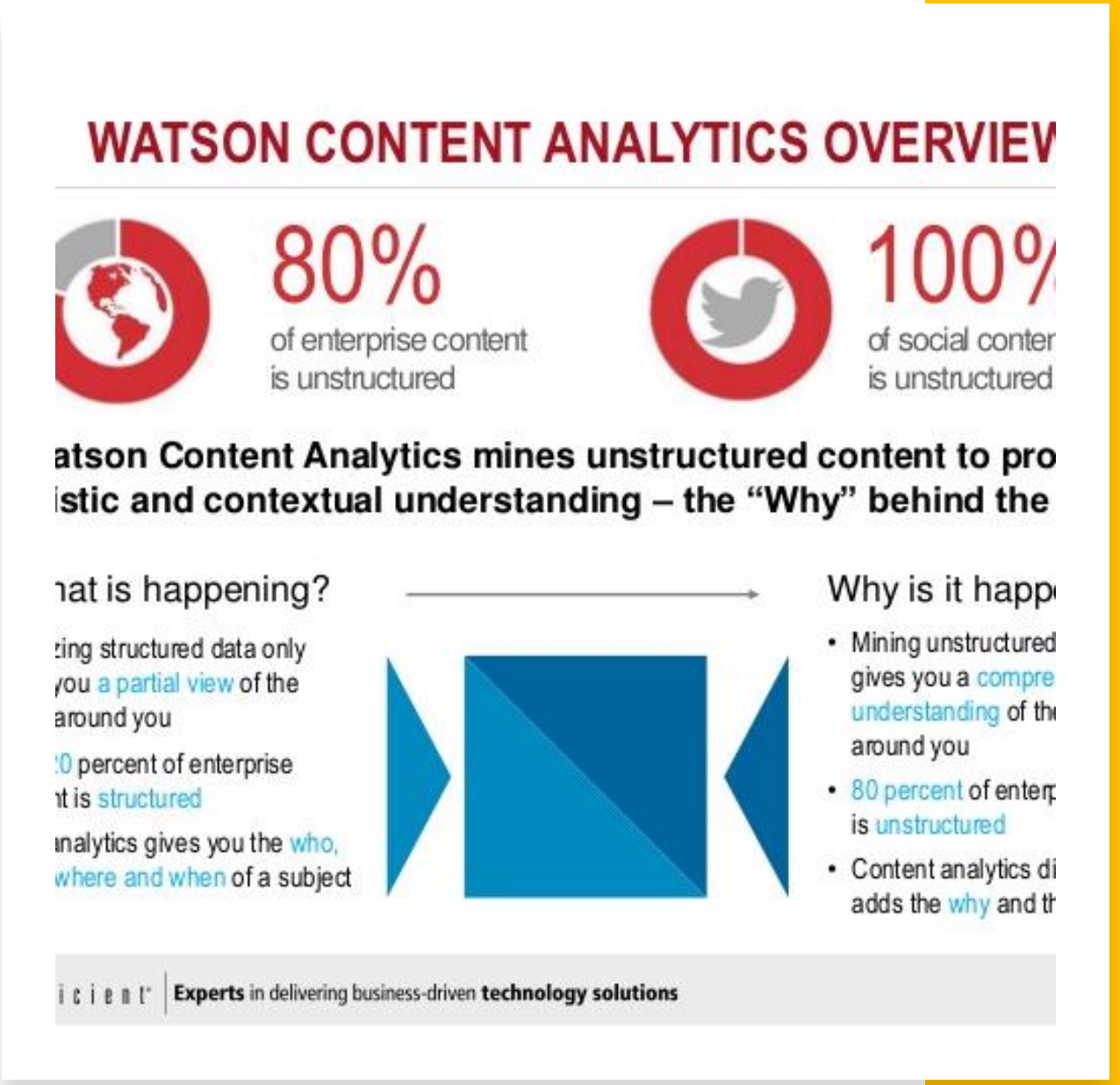
Науковий керівник: Мироненко С.С.

Мета роботи

- Аналіз галузі класифікації текстів.
- Дослідження задач аналізу тональності текстів та виділення ключових слів.
- Розробити додаток для класифікації тексту та визначення його основних тез.

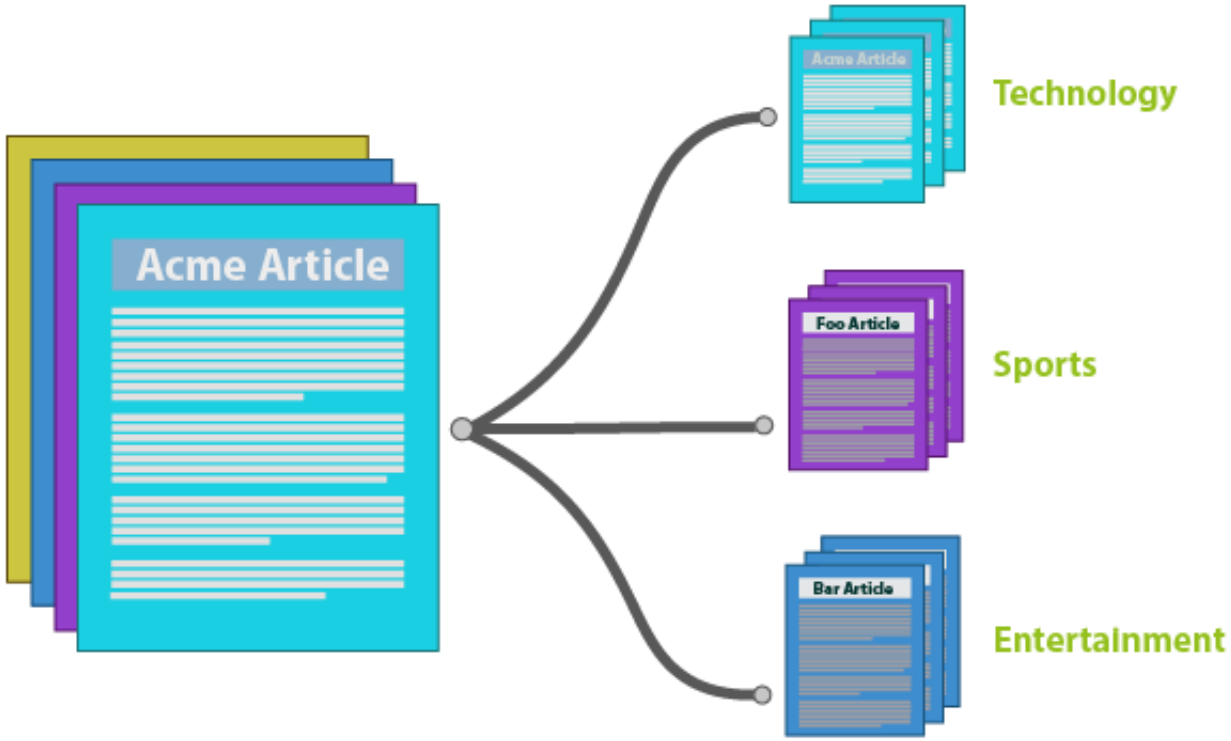
Актуальність

- За даними досліджень IBM, близько 80% даних у світі є неструктурованими.
- Кількість загальнодоступних даних постійно зростає.
- Текст являється найбільш поширеним типом неструктурованих даних.
- Робота великої кількості спеціалістів у різних областях пов’язана з постійним аналізом текстів.



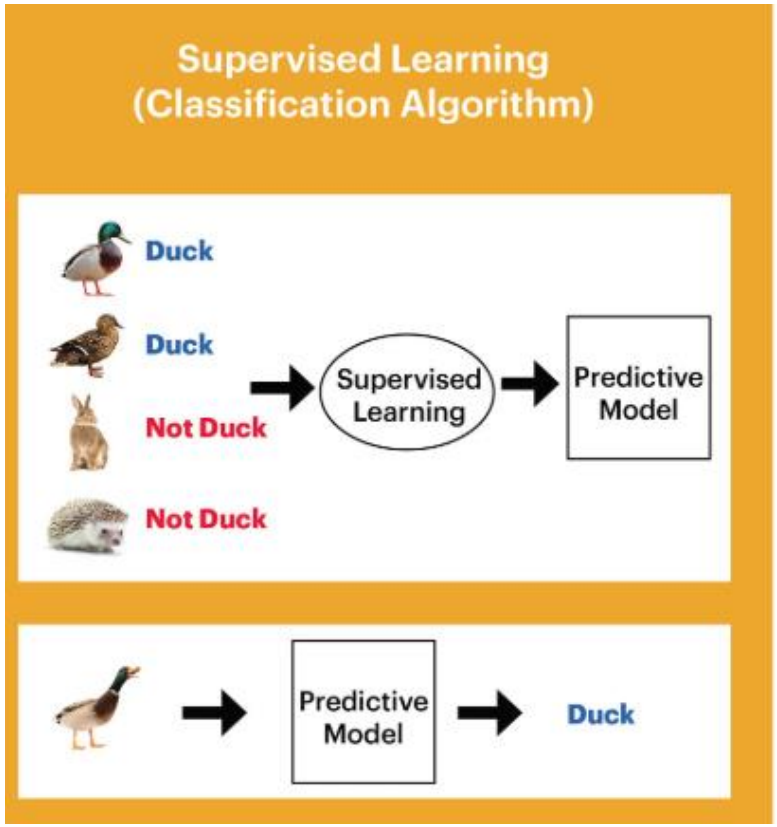
Класифікація текстів

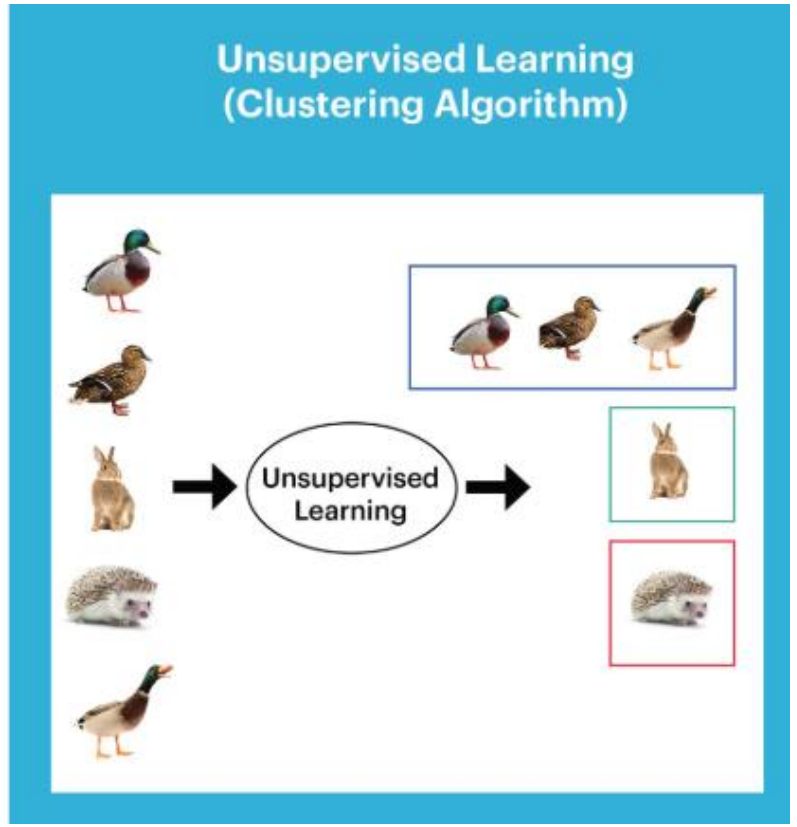
- Класифікація текстів – задача, яка полягає у віднесенні документа до однієї з декількох категорій на підставі змісту документа.
- Підходи до класифікації текстів: класифікація вручну та машинна класифікація.
- Приклади існуючих технологій, призначених для класифікації текстів: Word2Vec, fastText, GloVe.



Системи, на основі машинного навчання

- У підході навчання під наглядом, для того щоб модель могла робити передбачення, її необхідно натренувати на вибірці даних.
- При використанні підходу навчання без нагляду, моделі не потрібно попереднє навчання, використовуються лише властивості та характеристики текстового документа.





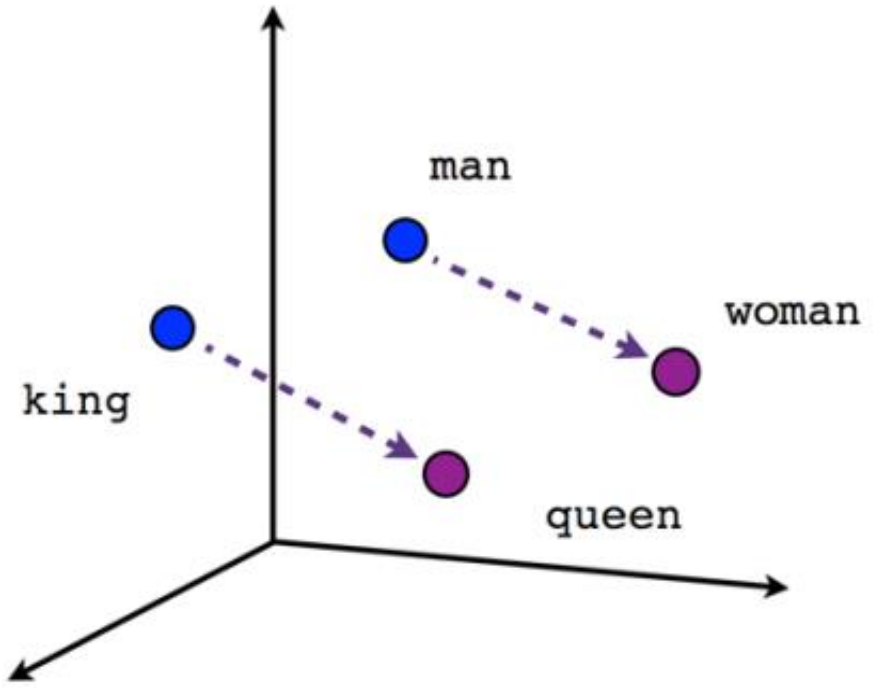
Векторна репрезентація слів

- Векторна репрезентація слів – вектори дійсних чисел, які репрезентують слова зі словника та застосовуються в галузі обробки природних мов.

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

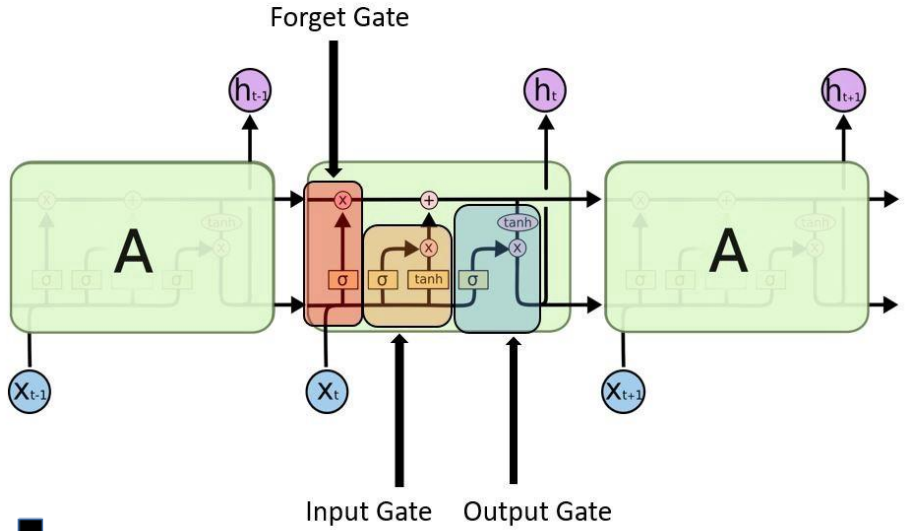


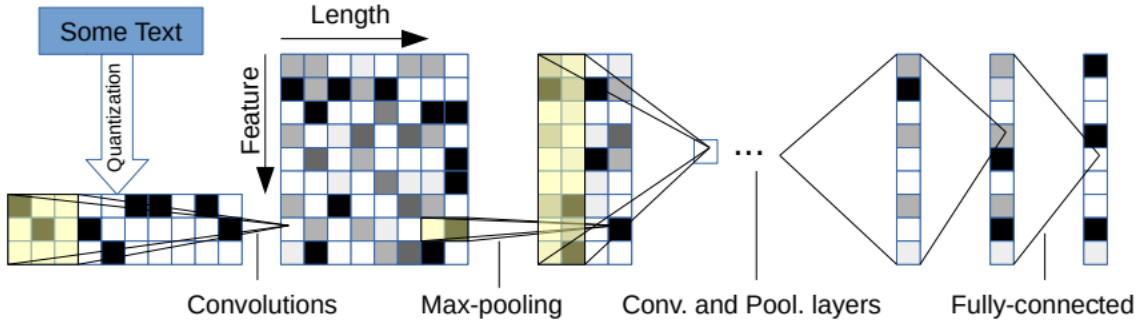
Задача аналізу тональності тексту

- Аналіз тональності тексту є процесом ідентифікації та визначення емоційного забарвлення тексту.
- Існуючі підходи: словниково-орієнтований підхід та підхід на основі машинного навчання.
- Для порівняльного аналізу були обрані наступні моделі:
  - Naïve Bayes Classifier
  - LSTM
  - 1D CNN
  - 3D CNN + BERT

Огляд обраних методів для аналізу

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$



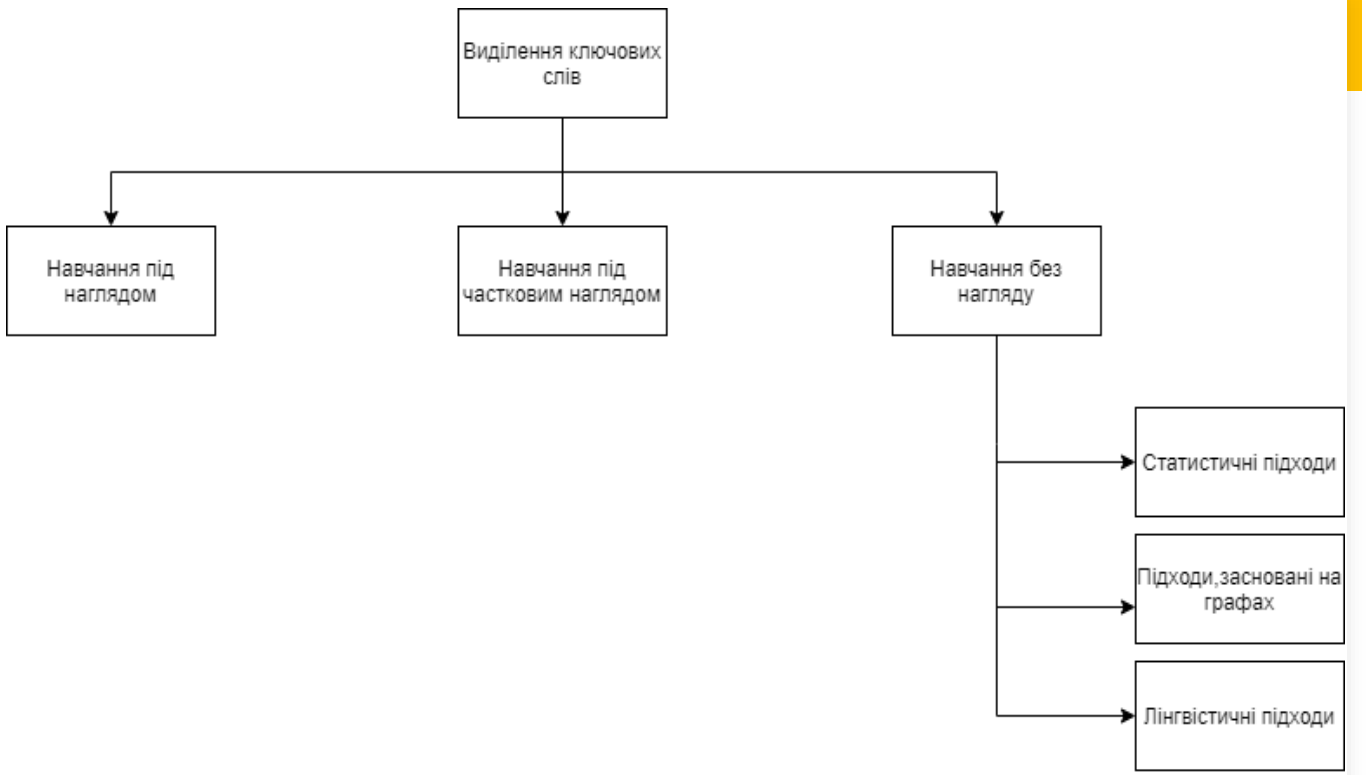


Класифікатор	Час тренування, с	Точність тренування, %	Час тестування, с	Точність тестування, %
Naïve Bayes Classifier	~ 1 с	98.0 %	~ 1 с	88.06 %
LSTM	2 829 с	95.03 %	97 с	75.32 %
1D CNN	461 с	99.0 %	21 с	88.57 %
3D CNN + Bert Tokenizer	2 304 с	99.27 %	4 с	89.72 %

Результати порівняльного аналізу методів сентимент аналізу тексту

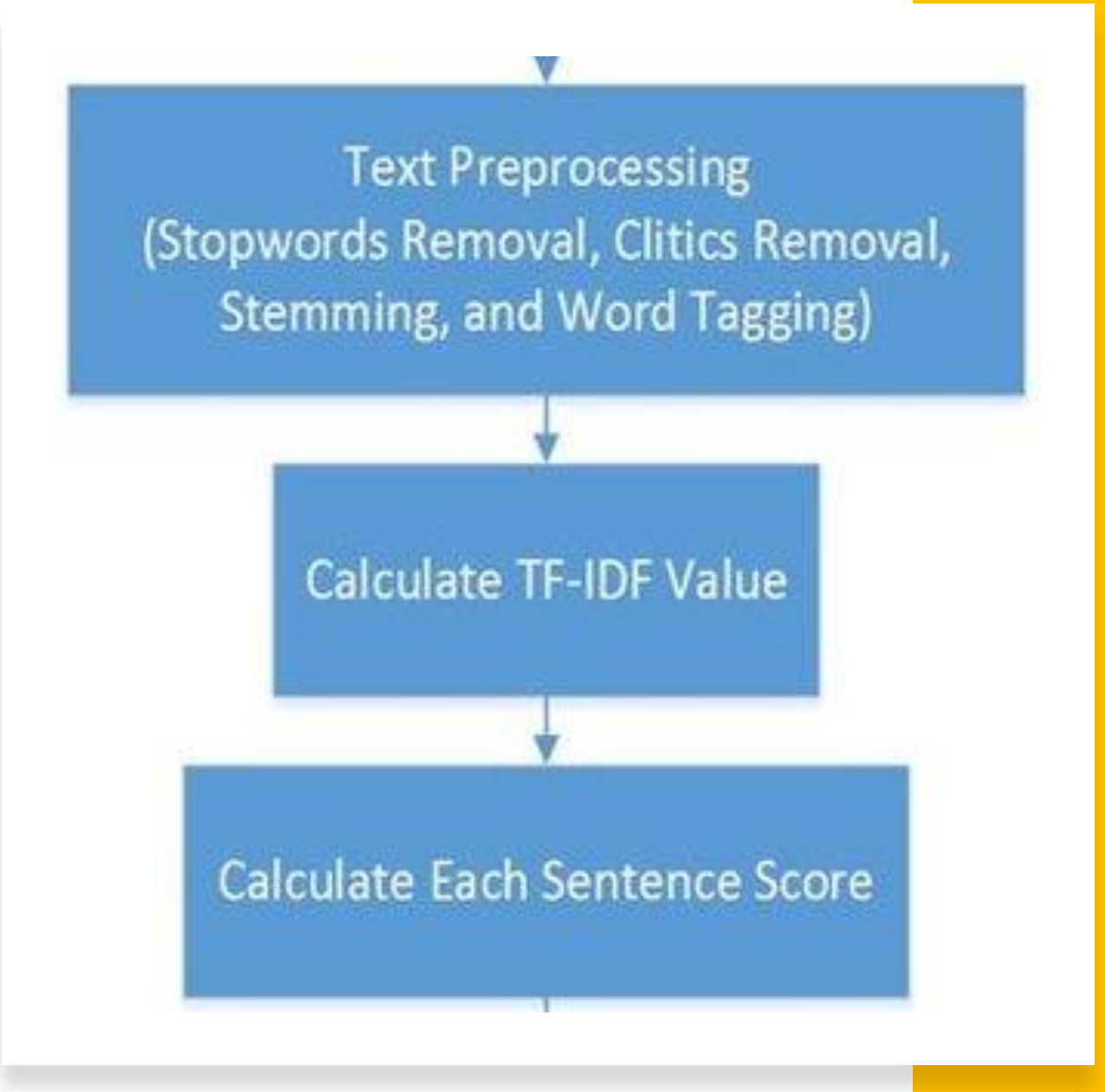
Задача виділення ключових слів

- Ключові слова – послідовність одного або декількох слів, які виражають загальний зміст тексту.
- Підходи для вирішення задачі: мануальний та автоматичний.
- Обрані методи для порівняльного аналізу : TF-IDF, RAKE та TextRank.



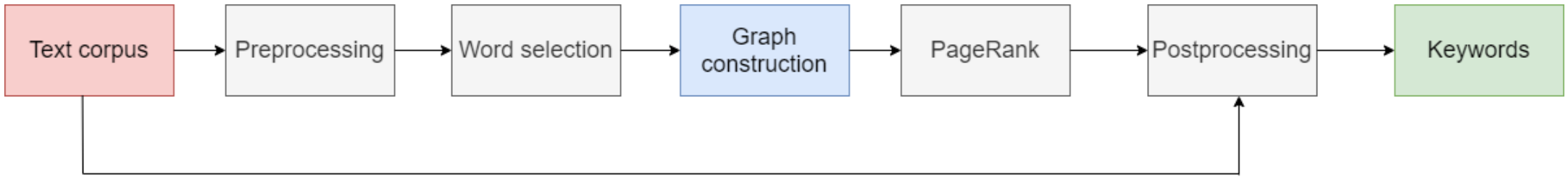
TF-IDF

- TF-IDF це формула, яка визначає ступінь важливості слова в текстовому документі.
- $TF - IDF_{score} = TF_{x,y} * IDF = TF_{x,y} * \log\frac{N}{df}$,
- де $TF_{x,y}$ – це частота ключового слова X в текстовому файлі Y, N – загальна кількість текстових документів в корпусі, df – кількість документів, в яких наявне ключове слово X .
- Добуток величин TF та IDF дає результат TF – IDF слова у документі. Чим цей показник вище, тим важливіше слово.



Методи RAKE та TextRank

- Rapid Automatic Keyword Extraction (RAKE) – це алгоритм автоматичного вилучення ключових слів з тексту.
- TextRank - графова технологія заснована на підході навчання без вчителя, яка використовується для отримання змісту тексту. Ця технологія заснована на алгоритму PageRank, який використовується для класифікації веб-сторінок.



Алгоритм	Середня точність, %	Середній час на аналіз
TextRank	80%	1.5 с
RAKE	81%	1.7 с
TF-IDF	77%	1 с

Результати порівняльного аналізу методів виділення ключових слів тексту

Створення додатку

- Мета додатку: підвищення ефективності статистики та зменшення кількості часу, необхідного на проведення попередньої класифікації.
- Цільова галузь застосування додатку: маркетинг

Вибір методів для реалізації у додатку

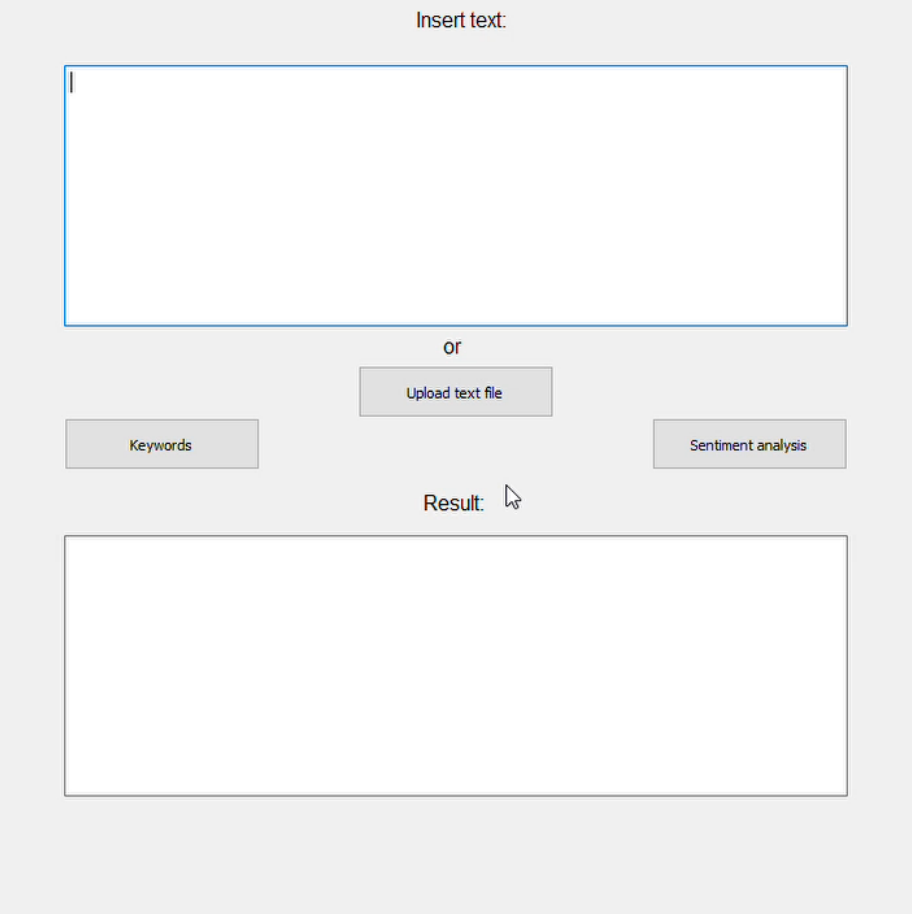
Аналіз емоційної складової тексту

- Обрана модель: модель, заснована на 3D Convolutional Neural Network та попередня обробка тексту здійснена за допомогою Bert токенізатора.
- Перевага моделі у точності класифікації та часу, необхідному на тестування.
- Показники моделі:
  - ***Час тренування***: 2 304 с
  - ***Час тестування***: 4 с
  - ***Точність тренування:*** 99.27 %
  - ***Точність тестування***: 89.72%

Виділення ключових слів

- Обрана модель: TextRank
- Основна перевага моделі над іншими у можливості оновлення словника стоп-слів.
- Показники моделі:
  - ***Середня точність:*** 80%
  - ***Середній час на аналіз:*** 1.5 с

Приклад роботи додатку



Висновки

В ході виконання роботи було проведене теоретичне та практичне дослідження галузі класифікації текстів.

Теоретичне дослідження містило в собі окрім загального вивчення галузі класифікації текстів, детальне дослідження задач аналізу тональності тексту та виділення ключових слів тексту.

Під час практичного дослідження було проведено порівняльний аналіз для задач аналізу тональності тексту та виділення ключових слів з метою визначення найбільш ефективних методів для реалізації у додатку.

Результатом роботи є додаток для класифікації тексту та виділення основних тез тексту, який використовує найбільш оптимальні та ефективні методи з доступних.

Можливі напрями подальшого розвитку

Розробка мобільного варіанту додатку

Регулювання конфігурацій BERT з метою покращення точності аналізу

Пошук шляхів вирішення проблеми виділення ключових слів з коротких текстів

Перехід від бінарного аналізу тональності до мультикласового

Додавання інформаційного пошуку до функціоналу

Розширення функціоналу, шляхом додавання автоматичної генерації питань до тексту

Дякую за увагу!