Інтелектуальна система розпізнавання тональності тексту

Шипік Данило Володимирович

ННК «ІПСА» НТУУ «КПІ ім. І. Сікорського»

2018

Актуальність роботи

- велика кількість інформації у сучасному світі, у тому числі величезна кількість новин, що з’являється на різних новинних ресурсах, змушує людей виокремлювати новини та ресурси, які варті їх уваги.
- Читання поганих новин зранку може знизити ефективність роботи на весь день

Мета роботи

- Метою даної роботи є розробка інтелектуальної системи оцінювання емоційного відклику на новини за коментарями, опис її принципів роботи та її практична реалізація.

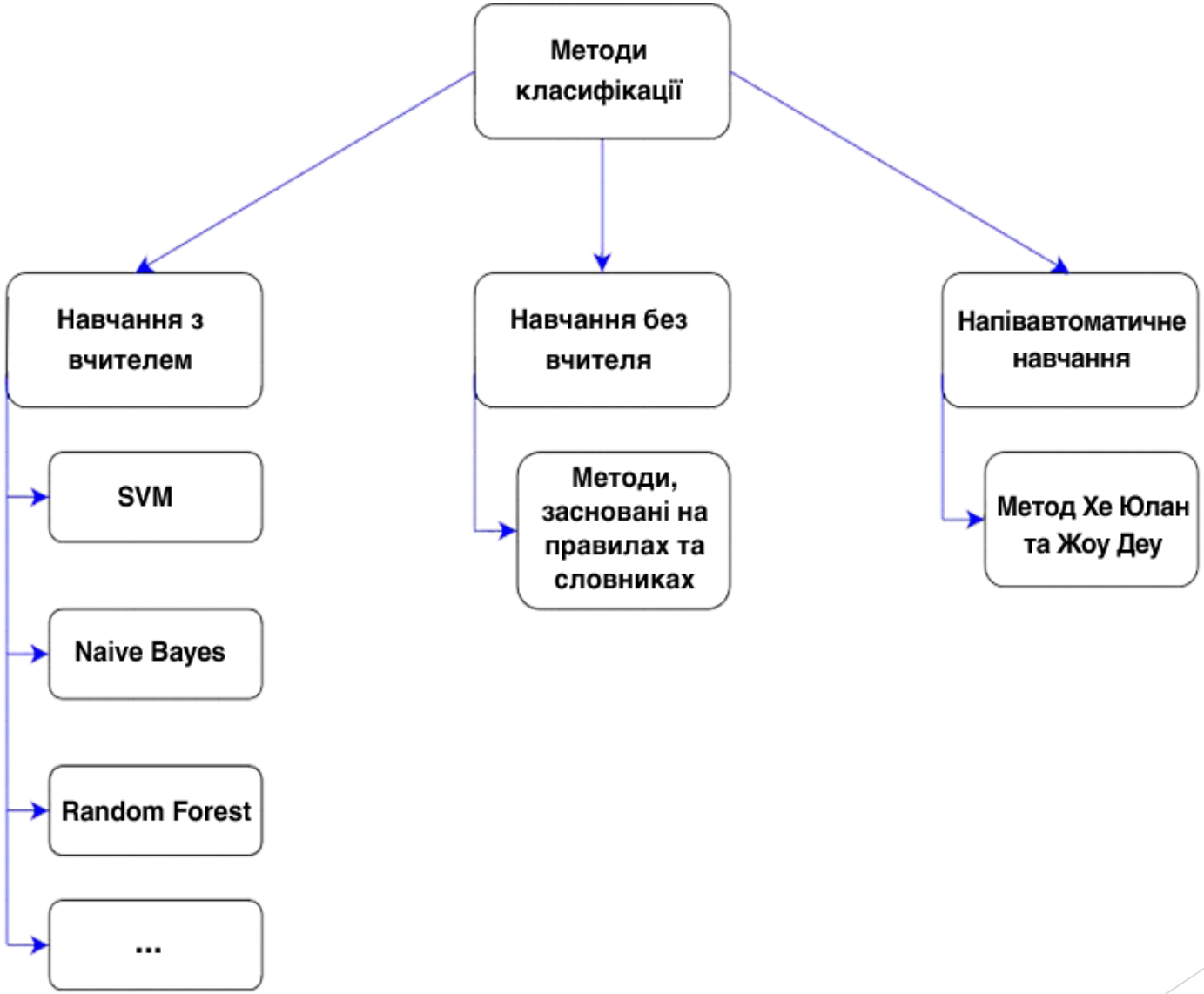
Постановка задачі магістерської дисертації

- Проаналізувати існуючі методи та рішення у галузі визначення тональності текстів
- Вивчити можливість збільшення ефективності існуючих методів
- Розробити програмний продукт (чат-бот), що втілює запропонований метод
- Проаналізувати ефективність запропонованих покращень

Об’єкт та предмет дослідження

- Об’єктом дослідження є методи розпізнавання тональності тексту.
- Предметом дослідження є методи Наївного Байеса, метод опорних векторів та метод Хе Юлан та Жоу Деу у контексті розпізнавання тональності текстів.

Аналіз існуючих методів



Критерії ефективності

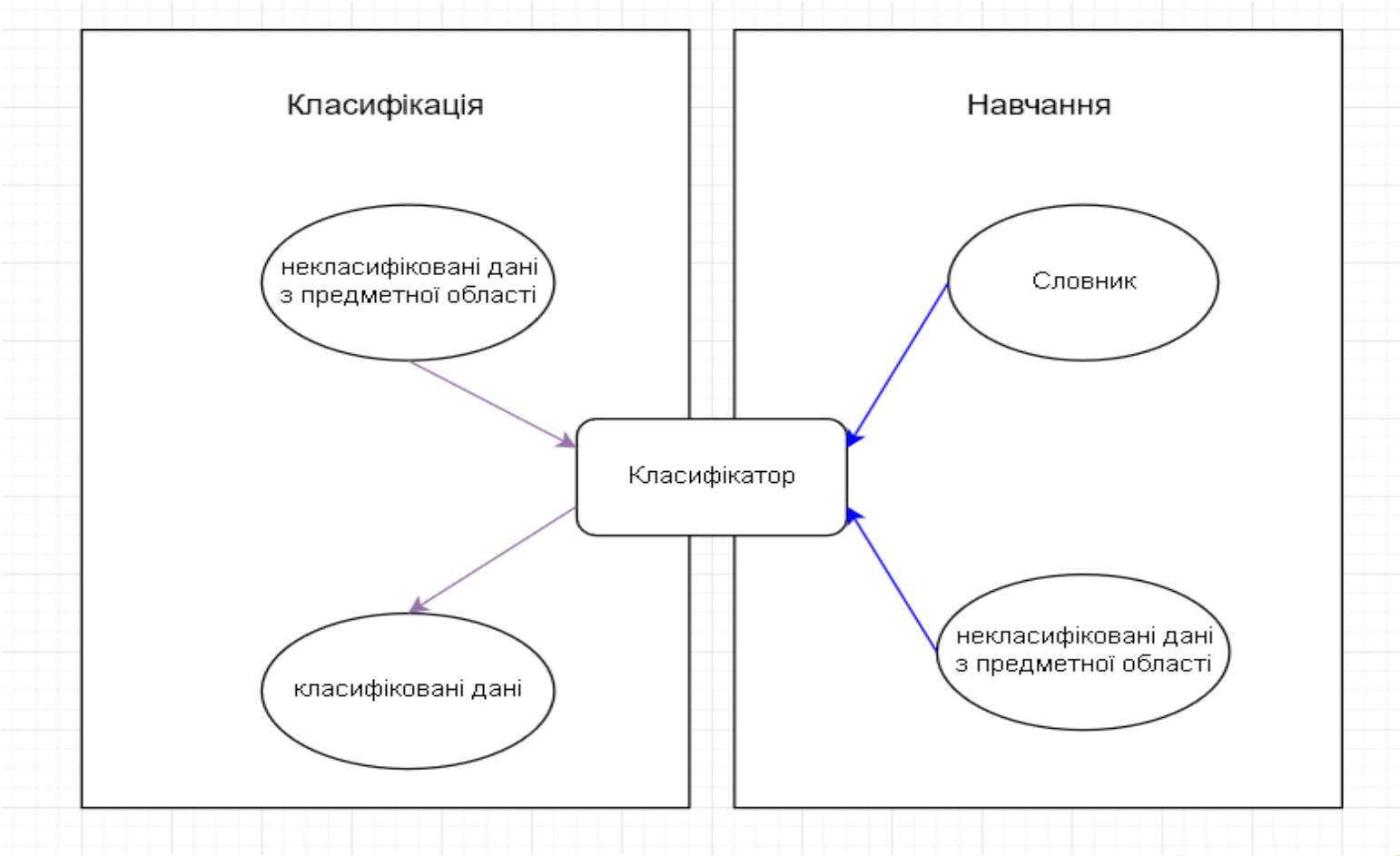
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

- **TP** – істинно-позитивне рішення;
- **TN** – істинно-негативне рішення;
- **FP** – хибно-позитивне рішення;
- **FN** – хибно-негативне рішення.

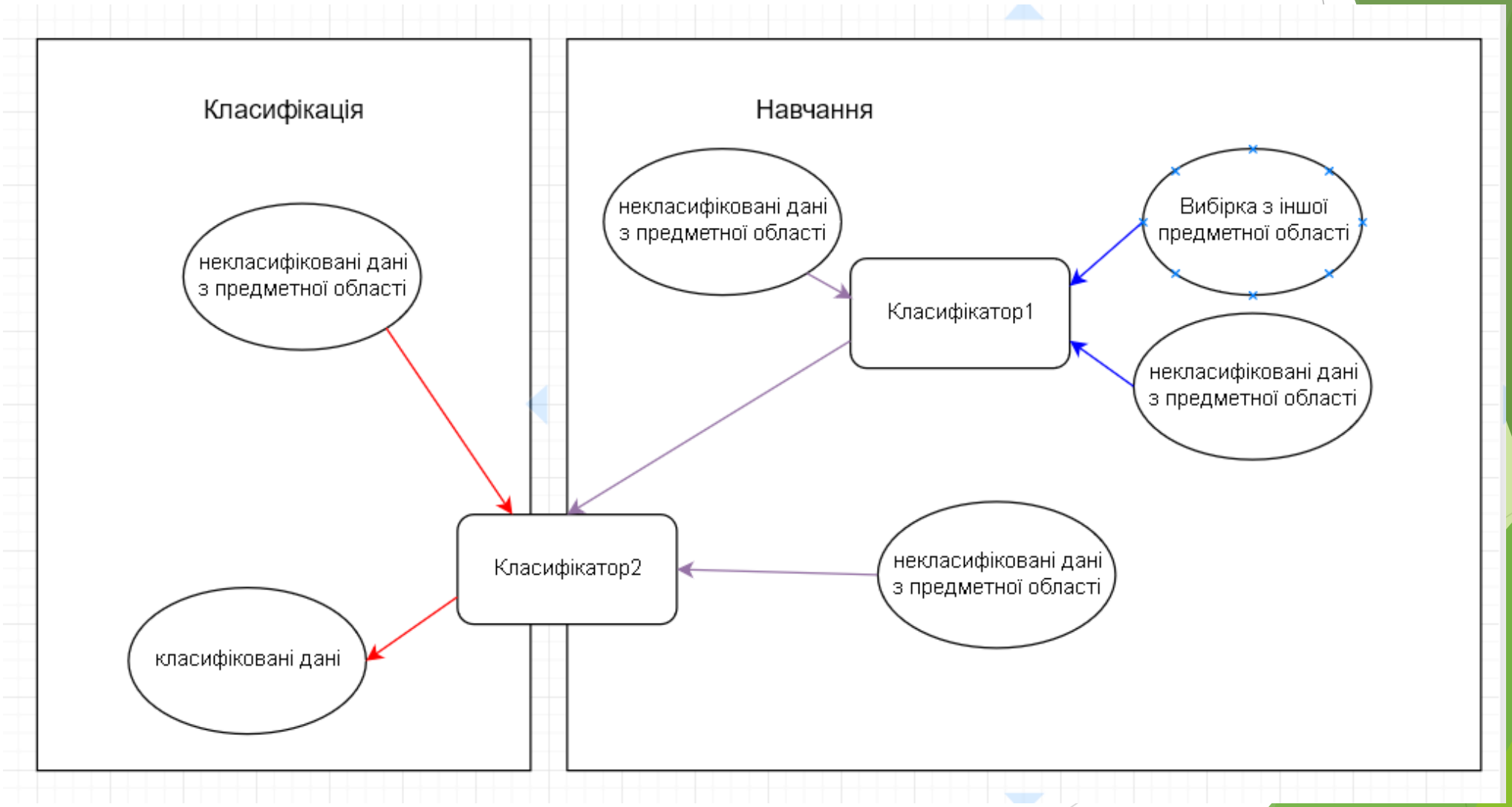
Метод Хе Юлан та Жоу Деу



Аналіз існуючих методів

	Precision	Recall	Accuracy
Словниковий метод	0.59	0.685	0.681
НБК	0.643	0.771	0.772
SVM	0.65	0.769	0.77
НБК з використанням біграм	0.683	0.782	0.791
SVM з використанням біграм	0.684	0.778	0.793
Метод Хе Юлан та Жоу Деу (з класифікатором НБК)	0.653	0.78	0.781

Запропонований метод



Порівняння можливих модифікацій за запропонованими критеріями

Модифікація методу	Precision	Recall	Accuracy
НБК + НБК	0.654	0.78	0.781
НБК + НБК з біграмами	0.661	0.778	0.783
НБК з біграмами + НБК з біграмами	0.683	0.789	0.799

Порівняння можливих модифікацій за часом навчання та класифікації

Метод	Швидкість навчання (с на 1млн прикладів)	Швидкість класифікації (с на 1 млн прикладів)
НБК з біграмами	1159	1084
Модифікація (НБК + НБК)	1796	794
Модифікація (НБК + НБК з біграмами)	2060	1073
Модифікація (НБК з біграмами + НБК з біграмами)	2303	1091

Результати роботи за запропонованими критеріями

Метод	Precision	Recall	Accuracy
Словниковий метод	0.59	0.685	0.681
НБК	0.643	0.771	0.772
SVM	0.65	0.769	0.77
НБК з використанням біграм	0.683	0.782	0.791
SVM з використанням біграм	0.684	0.778	0.793
Метод Хе Юлан та Жоу Деу (з класифікатором НБК)	0.653	0.78	0.781
Запропонована модифікація	0.683	0.789	0.799

Аналіз результатів

- Найкраща модифікація у якій в обох шарах використовується НБК з використанням біграм має ефективність роботи вищу за найкращий з інших розглянутих (НБК з використанням біграм) на 0.8% у Accuracy, 0.7% у Recall і 0% у Precision, проте працює майже (з точністю до похибки) у два рази повільніше у навчанні за звичайний НБК з використанням біграм.

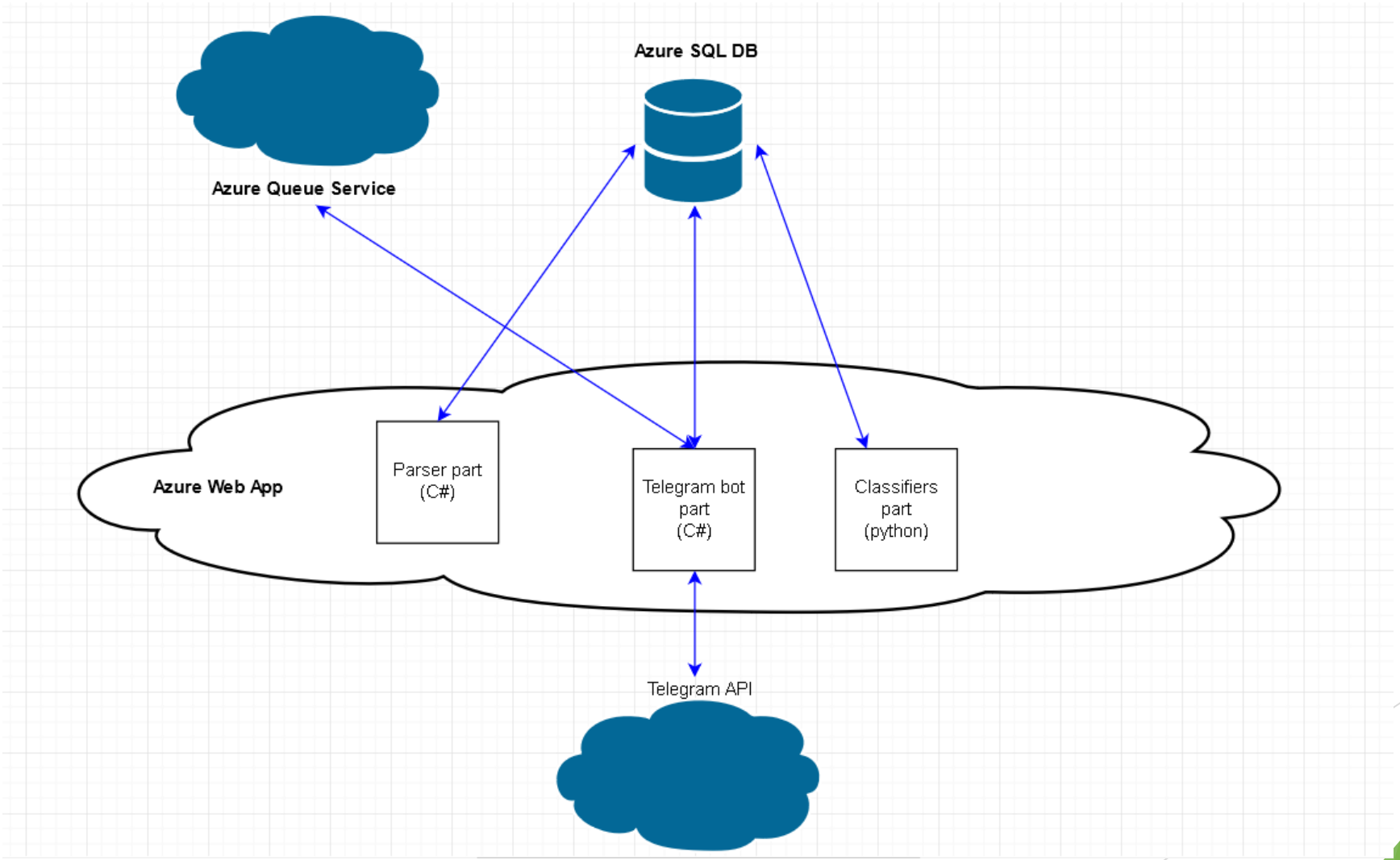
Вибір платформи та мови програмування

- В якості мов та інструментів були обрані C#, python, nltk, scikit-learn та Azure.

C#

Azure

Архітектура системи



Висновки

- Новизна роботи:
  - запропоновано принципово новий спосіб використання аналізу тональності, а саме аналіз емоційної реакції на певні новини за їх коментарями та подальше надсилання новин за допомогою чат-бота;
  - обґрунтовано використання власних підходів до розв’язку задачі аналізу тональності.
  - запропоновано модифікацію алгоритму аналізу тональності для коментарів, як коротких, емоційно забарвлених текстів зі змінюваною лексикою;

- Практична цінність роботи:
  - реалізовано запропоновану модифікацію класифікатора;
  - створено чат бота із зручним для користувача інтерфейсом, який використовує

модифікований алгоритм Хе Юлан та Жоу Деу, що демонструє результати

на 3% у Accuracy, 0.9% у Recall і 1,8% у Precision краще ніж оригінальний алгоритм

та 0.8% у Accuracy, 0.7% у Recall і 0% у Precision краще за найкращий розглянутий