Сентимент-аналіз тексту

Виконав:

Студент групи ДА-82

Муравльов А.Д.

Визначення сентимент-аналізу та тональності

Аналіз тональності тексту (сентимент-аналіз, Opinion mining) - клас методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики та емоційної оцінки авторів (думок) об'єктам, про які йдеться в тексті.

Тональність - це емоційне ставлення автора висловлювання до деякого об'єкту (об'єкту реального світу, події, процесу або їх властивостей/атрибутів), виражене в тексті. Емоційна складова, виражена лише на рівні лексеми чи комунікативного фрагмента, називається лексичною тональністю (чи лексичним сентиментом). Тональність всього тексту загалом можна з'ясувати, як функцію (у найпростішому разі суму) лексичних тональностей складових його одиниць (пропозицій) і правил їх поєднання.

Задачі сентимент аналізу

Основна задача сентимент-аналізу — знаходження думок в тексті та визначення їх властивостей.

Класифікація думок

- Тотальна оцінка
  - Позитивна
  - Негативна
  - Нейтральна
- Тип
  - Безпосередня думка
  - Порівняння
- Зміст думки
  - Entity / feature
  - orientation/polarity
  - Holder
  - Time

Види класифікації

За бінарною шкалою

Полярність документа можна визначати за бінарною шкалою. У цьому випадку для визначення полярності документа використовується два класи оцінок: позитивна чи негативна. Одним з недоліків цього підходу є те, що емоційну складову документа не завжди можна однозначно визначити, тобто документ може містити ознаки позитивної оцінки, так і негативної ознаки. Ранні роботи в цій області включають в себе праці Терні і Панга, які застосовують різні методи розпізнавання полярності оглядів товару і відгуків про фільмах відповідно. Це приклад роботи на рівні документа.

Види класифікації

Системи шкалювання

Іншим методом визначення тональності є використання систем шкалювання, за допомогою чого словами, зазвичай пов'язаних з негативними, нейтральними або позитивними тональностями, ставляться відповідно числа за шкалою від -10 до 10 (від негативного до самого позитивного). Спочатку фрагмент неструктурованого тексту досліджується з допомогою інструментів та алгоритмів обробки природної мови, а потім виділені з цього тексту об'єкти та терміни аналізуються з метою розуміння значення цих слів.

Види класифікації

За багатосмуговою шкалою

Можна класифікувати полярність документа по багатосмуговій шкалою, що було зроблено Пангом і Снайдером (серед інших). Ними було розширене основне завдання класифікації кіновідгуків від оцінки «позитивний або негативний» в бік прогнозування рейтингу по 3-х або 4-бальною шкалою. У той же час Снайдер провів поглиблений аналіз оглядів ресторанів, пророкуючи рейтинги їх різних властивостей, таких як їжа і атмосфера (за 5-бальною шкалою).

Види класифікації

Суб’єктивність/об’єктивність

Інший дослідницький напрямок — це ідентифікація суб'єктивності/об'єктивності. Це завдання зазвичай визначається як віднесення даного тексту в один з двох класі суб'єктивний й або об'єктивний. Ця проблема іноді може бути більш складною, ніж класифікація полярності: суб'єктивність слів і фраз може залежати від контексту, а об'єктивний документ може містити в собі суб'єктивні пропозиції (наприклад, новинна стаття, цитує думки людей). Більш того, як згадував Су, результати більшою мірою залежать від визначення суб'єктивності, вживаючийся в рамках анотації текстів. Як би те ні було, Панг показав, що видалення об'єктивних пропозицій з документа перед класифікацією полярності допомогло підвищити точність результатів.

Методи класифікації

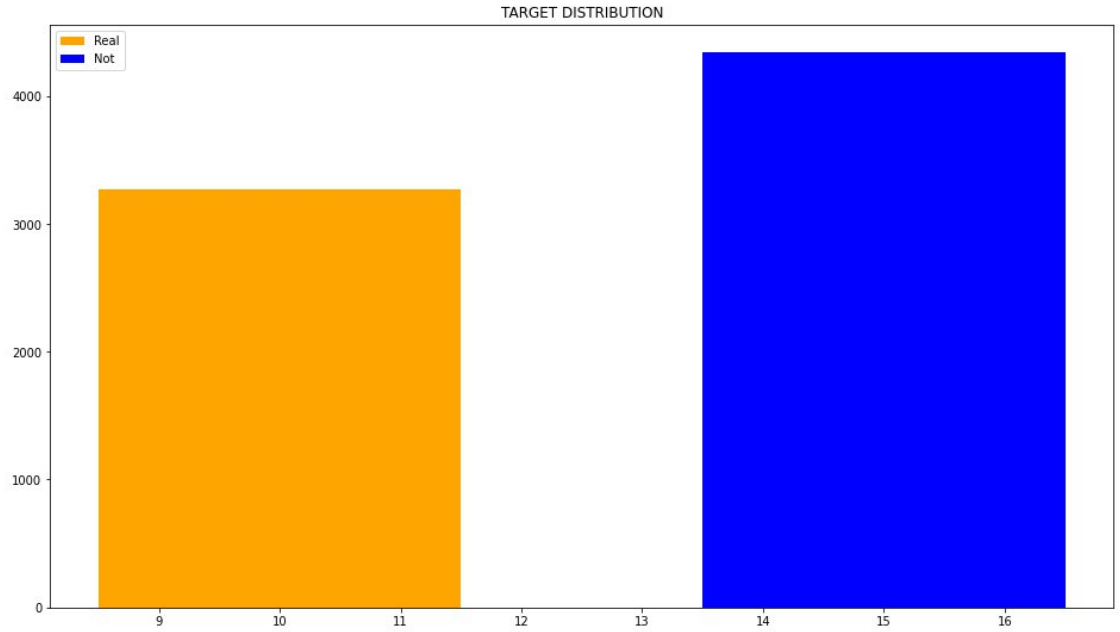
- Засновані на правилах та словниках
- Supervised learning
- Unsupervised learning
- Теоретико-графові моделі

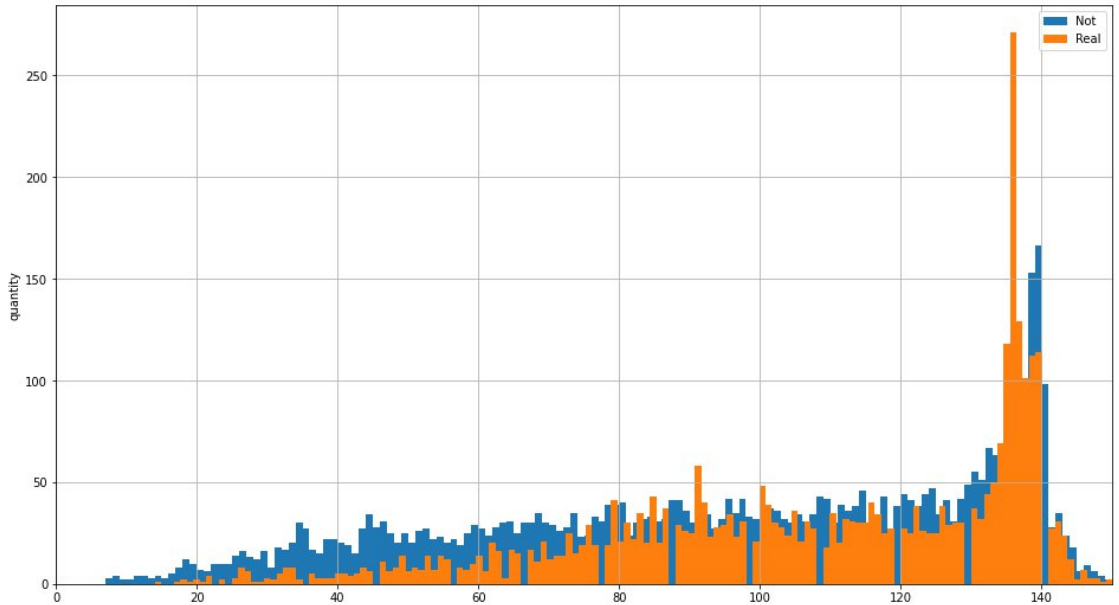
Практична реалізація

Розглянемо приклад сентимент-аналізу тексту на прикладі kaggle competition “Natural Language Processing with Disaster Tweets”. Суть роботи зводиться до визначення чи описується в твіті реальна катастрофа чи ні.

Використаємо навчання з учителем та класифікацію за бінарною шкалою.

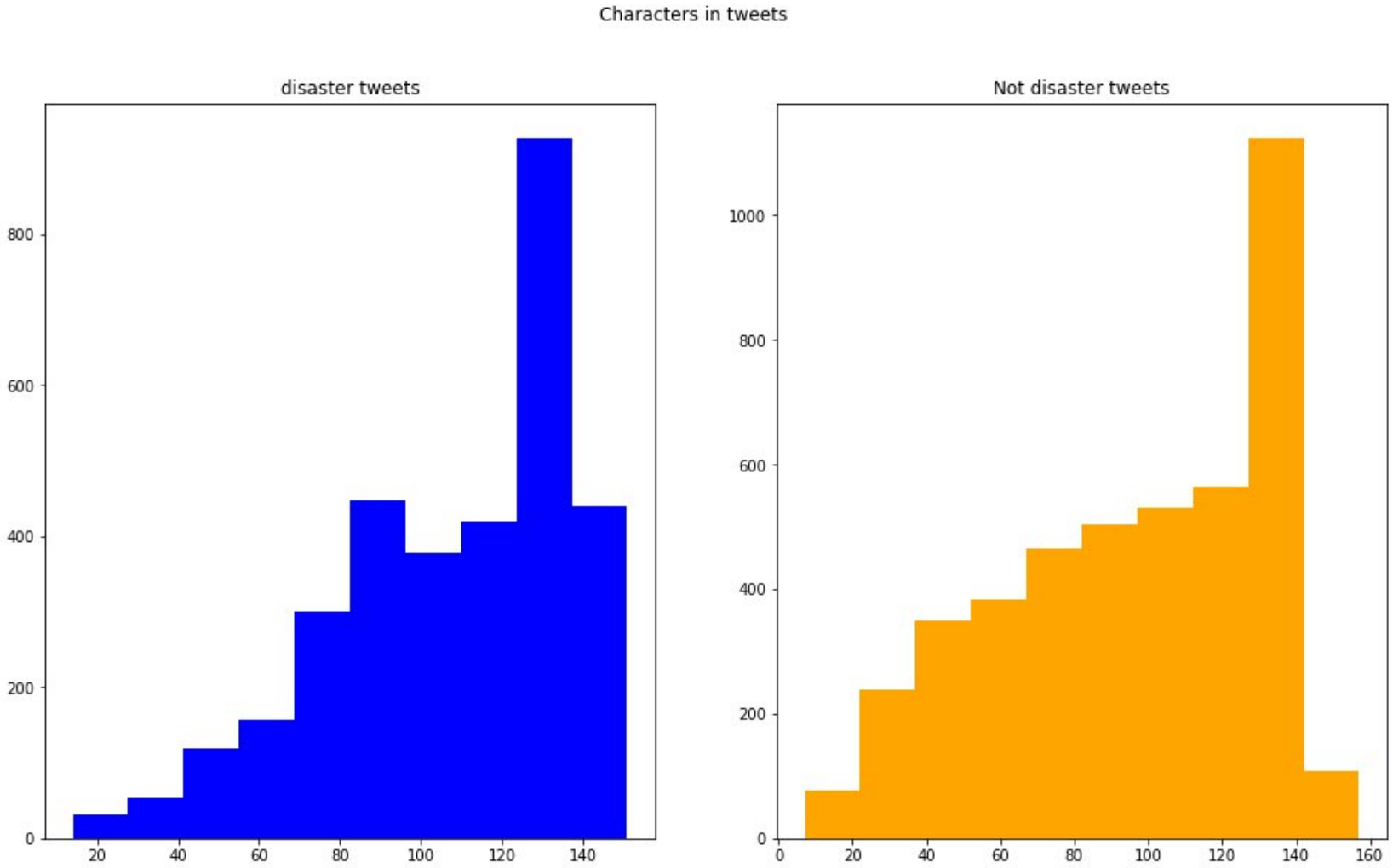
Аналіз тренувального датасету



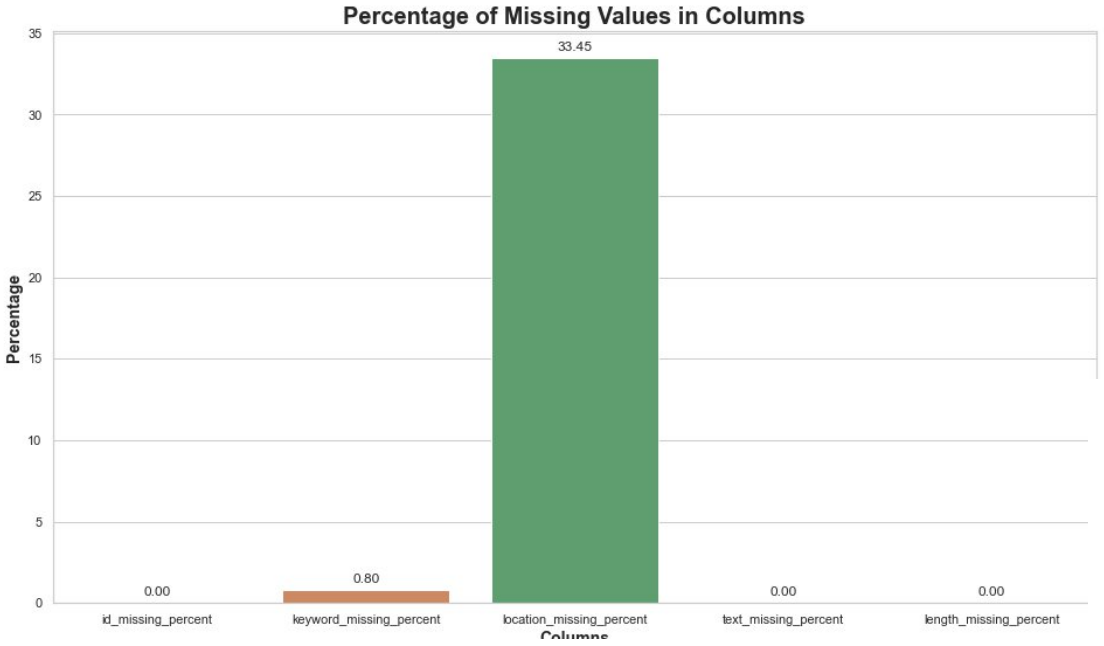


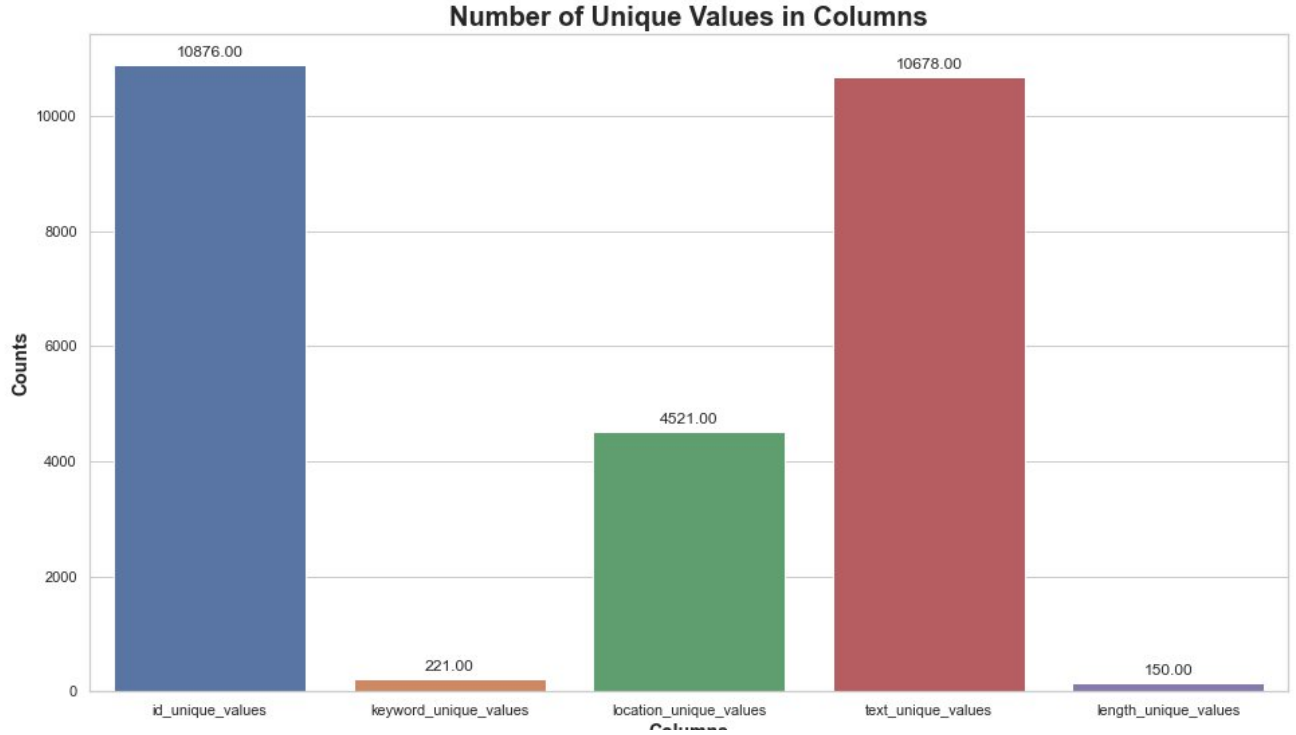
Here we can se an imbalance towards the negative class in dataset. This information could have a severe impact on classifier we create. For example, the classifier may be inclined to predict predominant class, which would ,ean higher accuracy. In our case, the imbalance isn't to serve yet it's still something worthy of taking notes of.

Аналіз тренувального датасету

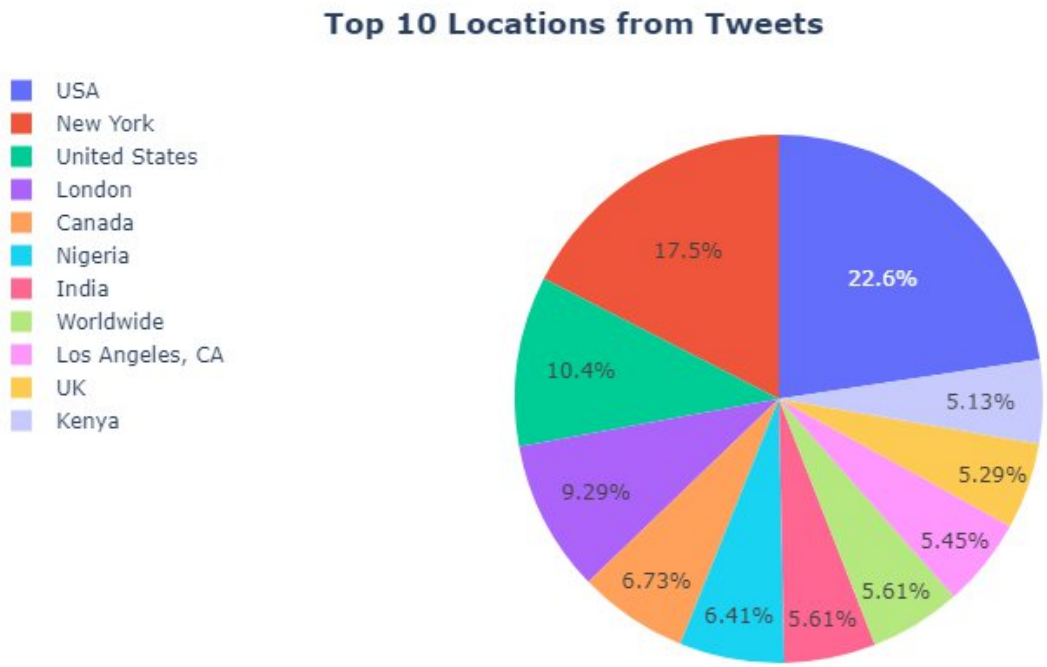


Аналіз тренувального датасету



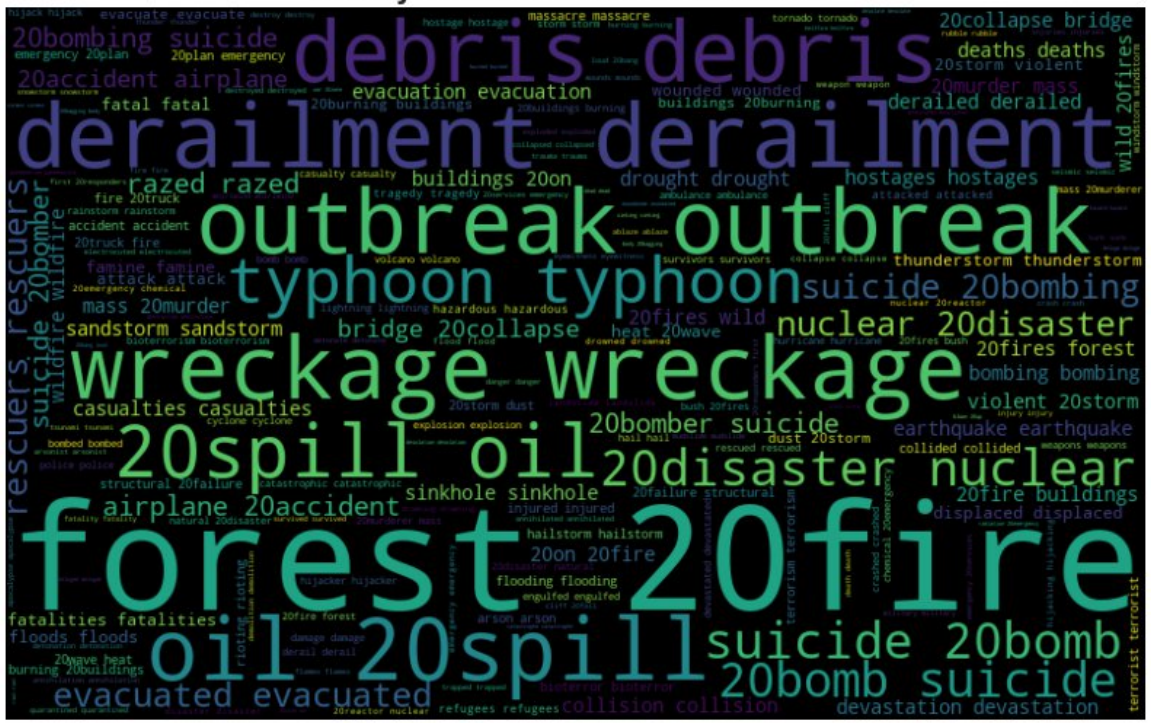


Аналіз тренувального датасету



Аналіз тренувального датасету

All keywords associated to Disasters



All keywords associated to Non-Disasters

Кроки навчання

- Заповнення пустих значень
- Видалення шумових слів
- Видалення посилань, лемматизація та токенізація тексту
- Векторизація даних та підгонка даних методом TFIDF для створення класифікатора
- Навчання

Методи навчання обрані для роботи

- Naive Bayes
- Логістична регресія
- Support vector machine
- Random forest

Результати навчання

Naive Bayes
- -AUC ROC: 0.7966
- -Kaggle: 0.7956

Random Forest
- -AUC ROC: 0.7778
- -Kaggle: 0.7833

Логістична регресія
- -AUC ROC: 0.8113
- -Kaggle: 0.7934

SVM
- -AUC ROC: 0.8111
- -Kaggle: 0.7962

Висновки

У ході виконання лабораторної роботи розглянули поняття сентимент аналізу тексту, методи та види класифікації у сентимент аналізі тексту та реалізували працюючу модель сентимент-аналізу тексту на реальних даних. Повний лістинг коду моделі та аналізу викладено у системі контролю версій Github за посиланням https://github.com/lakub-muravlov/fourth-course-projects/blob/main/AI/RGR/Code/RGR.ipynb