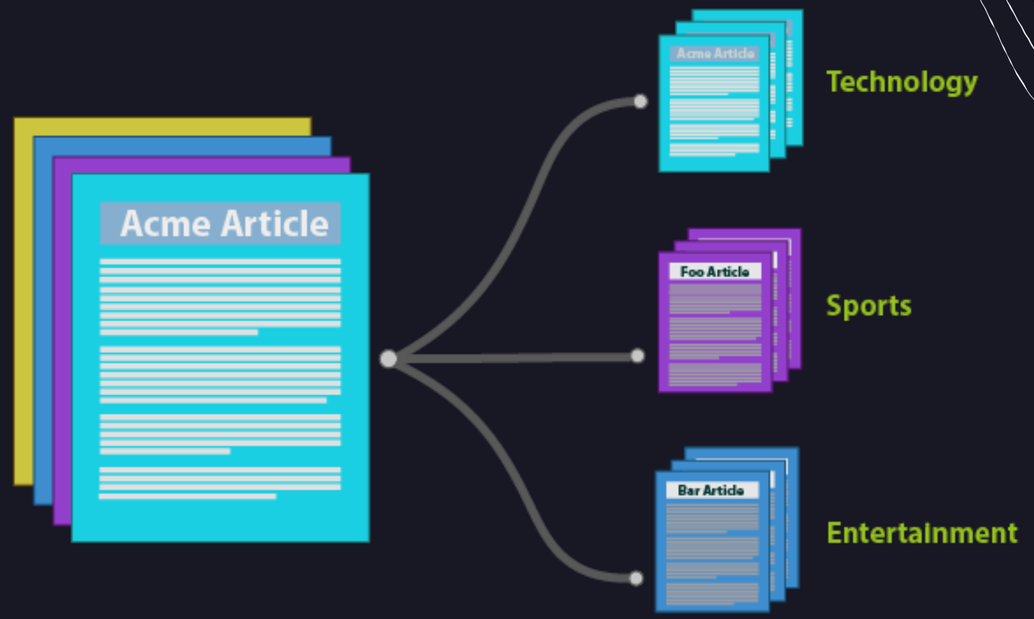
Балансування категорій у цільовому класі в тренувальних даних для задач класифікації тексту

Доповідач: Кучьменко Юрій Андрійович
Науковий керівник : Смиш Олег Русланович
Національний університет «Києво-Могилянська академія»

Класифікація текстів

Присвоєння категорій текстів, які можуть бути веб-сторінкою, книгою, статтею для ЗМІ та іншим, має безліч застосувань, наприклад:

- аналіз настроїв
- позначення тем
- класифікація новин
- визначення мови та інше



Автоматизація оцінювання есе

Датасет: The Hewlett Foundation: Automated Essay Scoring

Анонімісовані есе англійською мовою

Росмір датасету: 10323 сапncn (15.5 MB)

Значення цільового класу: 1-10

Попередня обробка тексту

Для попередньої обробки використано:

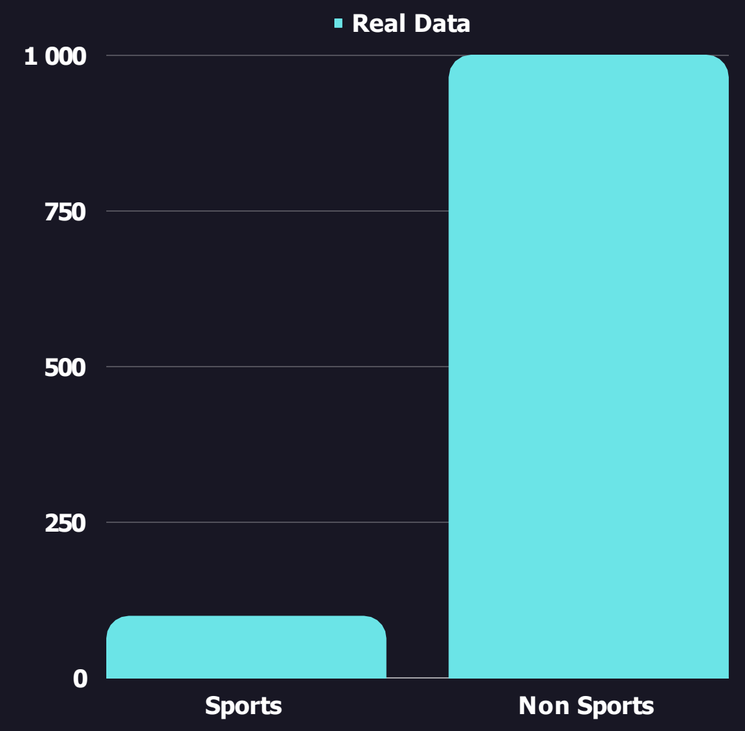
01 Токенізацію

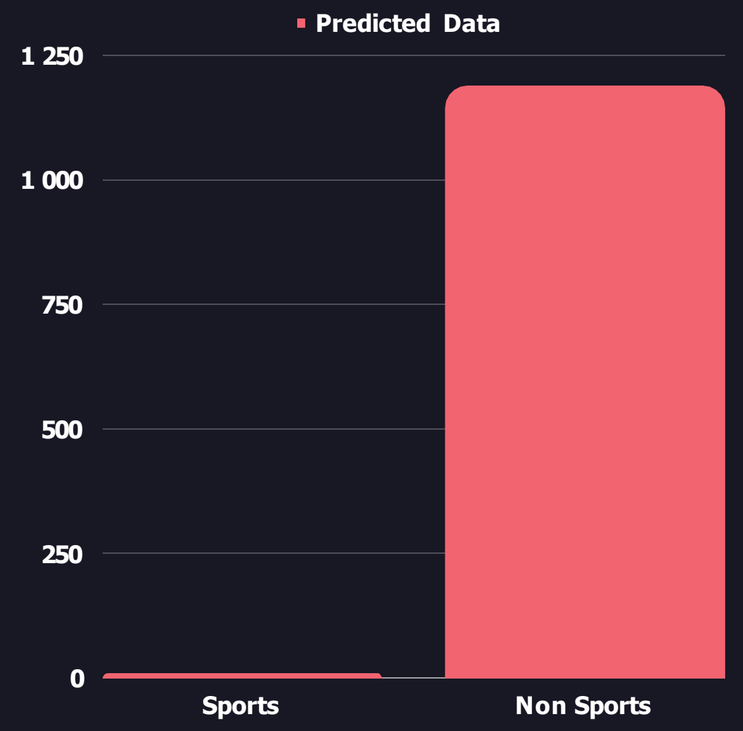
02 Вилучення стоп слів

03 Лематизацію

04 Видалення пунктуації

Проблема незбалансованих даних при класифікації тексту

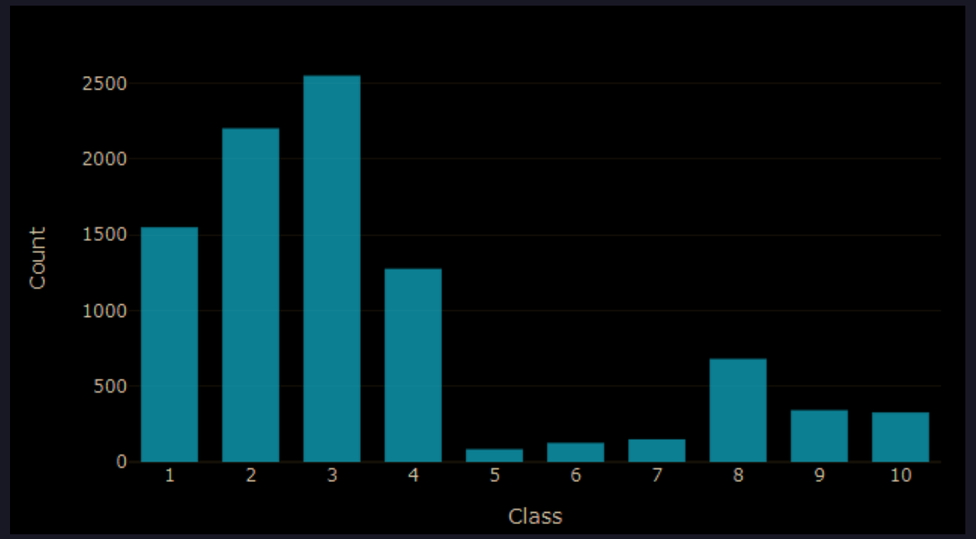




Вирішення проблеми незбалансованих даних

RandomOverSampler - алгоритм передискретизації даних, який використовується для балансування наборів даних з незбалансованими класами шляхом збільшення кількості зразків в меншому класі

До балансування даних



Збалансовані дані
за допомогою RandomOverSampler

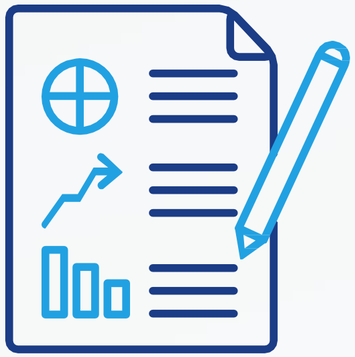
Результати

Після балансування модель навчалась довше, але метрики F1 покращились:

ВИСНОВОК

У дослідженні продемонстровано, що балансування розподілу класів у навчальних даних є ефективним підходом для покращення продуктивності моделі класифікації тексту

Клас **RandomOverSampler** зі стратегією вибірки "не більшості" може бути використано для балансування набору даних.



Q|A