Софийски Университет
”Св. Климент Охридски”

Извличане на информация

16.12. 2024 г.

Класификатор на фалшиви новини(чрез Spark)

Ради Радев

8MI3400507

Разпределени Системи и Мобилни технологии

Съдържание:

Мотивация

1. Липса на критичност на читателите на новини(особено оналайн)
2. Фалшивите новини са нарастващ глобален проблем.
3. Влияят негативно на обществото, демокрацията и доверието.

Предложено решение

Използване на Apache Spark за обработка на големи набори от данни.

Основни характеристики:

- Разпределена обработка на данни.
- Толерантен към грешки
- Лесно скалируем

Може да се използва и за:

- Извличане на текстови характеристики чрез NLP.
- Обучение и оценка на машинни модели.

Преглед на набора от данни

Примерни източници:

- *LIAR Dataset*, Kaggle Fake News Dataset.
- https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets?resource=download

Основни атрибути:

- Заглавие, съдържание, етикет (*истински/фалшиви*).

Обем: Над 80000 статии, с разнообразие от теми(около 8).

Архитектура на системата

1. Събиране на данни (CSV or Parquet).
2. Предварителна обработка (премахване на шум, токенизация).
3. Извличане на характеристики (TF-IDF, Word2Vec).
4. Обучение на модел (логистична регресия, случайни гори).
5. Оценка и прогноза.
6. Контейнеризация и преносимост на прокета

Основни моменти от реализацията

Предварителна обработка:

- Премахване на стоп думи и пунктуация.
- Токенизация и стеминг.

Характеристики:

- TF-IDF (важност на думите).
- N-грам модели и анализ на дължината на текста.

Основни моменти от реализацията

Spark MLlib:

- Логистична регресия за класификация.
- Случайни гори за по-сложни модели.

Оценка:

- Метрики: Точност, прецизност, F1-скор.

Резултати и предизвикателства

- **Резултати**:
  - Постигната точност: 95+%.
  - Подобрение след оптимизация на данните.
- **Предизвикателства**:
  - Балансиране на класовете.
  - Мащабируемост на обработката на данни.

Заключение

Постигнато:

- Изградена система за откриване на фалшиви новини.
- Подобрени резултати чрез използване на Spark.

Бъдещи подобрения:

- Въвеждане на реални времеви прогнози.
- Използване на модели като BERT за по-точно разпознаване.

Технологии

Python

Poetry

Pyspark

Docker

Декалрация за плагиатство

- Тази курсова работа е моя работа, като всички изречения, илюстрации и програми от други хора са изрично цитирани.
- Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.
- Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

Ради Стефчев Радев