|  |  |  |
| --- | --- | --- |
|  | **Софийски университет „Св. Кл. Охридски”**  Факултет по математика и информатика |  |

**Курсов Проект**

на тема: „Откриване на фалшиви новини с използване на Python Scikit-learn

”

**Иван Мишев Михов 9МI3400516**

**Николай Георгиев Тодоров 1МI3400515**

**Специалност ИИОЗ**

**Преподавател : проф. Дмн Галя Ангелова**

Курс: „1“, Учебна година: 2024/2025

**Съдържание**

[1 Въведение 2](#_Toc189861827)

[2 Избор на корпус и допълнителни данни (Данни) 2](#_Toc189861828)

[3 Избор на подход (Метод) : 2](#_Toc189861829)

[3.1 Какви основни обработки от КЛ са необходими 2](#_Toc189861830)

[3.2 Какви допълнителни алгоритми са приложени 3](#_Toc189861831)

[4 Експерименти върху данните (Експерименти) 3](#_Toc189861832)

[4.1 Подготовка на данните 3](#_Toc189861833)

[4.2 Обучение на моделите 3](#_Toc189861834)

[4.3 Подготовка на данните 4](#_Toc189861835)

[5 Как ще се валидират резултатите, анализ на грешките (Резултати, Дискусия) 4](#_Toc189861836)

[6 Бъдеща работа (Заключение и бъдеща работа) 4](#_Toc189861837)

[7 Свършена работа : 5](#_Toc189861838)

[Иван Михов: 5](#_Toc189861839)

[Преработване на данните и разработване на Flask приложение 5](#_Toc189861840)

[Николай Тодоров: 5](#_Toc189861841)

[Векторизация и обучение на модели 5](#_Toc189861842)

[Линк към гитхъб репо : 6](#_Toc189861843)

# Въведение

Проектът има за цел разработването на система за откриване на фалшиви новини, използвайки техники за машинно обучение и обработка на естествен език (NLP). Основната мотивация зад този проект е нарастващият обем на фалшива информация в онлайн медиите и социалните мрежи, което води до значителни последици върху общественото мнение, политическите процеси и личните решения.

# Избор на корпус и допълнителни данни (Данни)

За реализиране на проекта се използват анотирани текстови корпуси с фалшиви и истински новини. Основният използван корпус е **[Fake News Detection Dataset](https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets)** от Kaggle, който предоставя богата колекция от новинарски статии с ясно обозначени етикети за фалшиви и истински новини.

Датасетът съдържа разнообразие от текстове, обхващащи различни тематики и източници, което го прави подходящ за изграждане на модели с добра генерализация. Включените данни са структурирани в няколко ключови полета, като заглавие, съдържание на статията и етикет за класификация (истинска или фалшива новина).

За подобряване на качеството на модела е извършена допълнителна обработка на данните, включваща премахване на дублиращи се записи, почистване на шума (ненужни символи, HTML тагове и др.) и нормализиране на текста.

# Избор на подход (Метод) :

## Какви основни обработки от КЛ са необходими

* **Предобработка на текста:** Включва почистване на текста чрез премахване на стоп-думи, пунктуация и специални символи. Също така се извършва лемматизация, за да се намалят думите до техните базови форми, и нормализиране на текста за осигуряване на консистентност.
* **Извличане на характеристики:** Използва се TF-IDF (Term Frequency-Inverse Document Frequency) за представяне на текста като числови вектори. Тази техника позволява открояване на значими думи, които често се срещат във фалшиви новини.
* **Анализ на настроения:** Използва се VADER Sentiment Analyzer за идентифициране на емоционални тоналности в текста. Това е важно, тъй като фалшивите новини често се характеризират със силно емоционално съдържание.

## Какви допълнителни алгоритми са приложени

* **Класификационни модели:** Използвани са алгоритми като Logistic Regression, Naive Bayes и Support Vector Machines (SVM) за класификация на новините. Всеки от тези модели има своите предимства при обработка на текстови данни и се прилага в зависимост от спецификата на задачата.
* **Методи за балансиране на данните:** При наличие на дисбаланс в класовете (повече истински новини спрямо фалшиви или обратно) се използват техники като SMOTE (Synthetic Minority Over-sampling Technique) за създаване на синтетични примери и подобряване на представянето на модела.
* **Интеграция с Flask:** За демонстриране на резултатите е разработено просто уеб приложение с помощта на Flask, което позволява лесно тестване на модела в реална среда.

# Експерименти върху данните (Експерименти)

В тази секция са описани проведените експерименти, целящи оценка на ефективността на използваните методи и модели за откриване на фалшиви новини.

## Подготовка на данните

Първата стъпка в експерименталния процес е разделянето на данните на тренировъчен и тестови набор, като е използвано съотношение 80:20. Това осигурява достатъчно данни за обучение на модела, като същевременно се запазва представителна извадка за тестване.

* **Почистване на данните:** Премахнати са дублиращи се записи, ненужни символи и празни стойности.

## Обучение на моделите

За обучение са използвани три основни класификационни модела:

* **Logistic Regression:** Подходящ за бинарна класификация с висока ефективност при текстови данни.
* **Naive Bayes:** Работи добре с високодименсионални текстови данни, използвайки вероятностен подход.
* **Support Vector Machines (SVM):** Известен със своята точност при разделяне на данни в различни класове.

Всеки модел е обучаван с различни хиперпараметри, за да се оптимизира производителността.

## Подготовка на данните

Оценката на моделите е извършена с помощта на следните метрики:

* **Точност (Accuracy):** Процент на правилно класифицираните примери.
* **Прецизност (Precision):** Оценка на броя на правилно идентифицираните фалшиви новини спрямо всички, класифицирани като фалшиви.
* **Чувствителност (Recall):** Способността на модела да открива всички действителни фалшиви новини.
* **F1-скор:** Хармонично средно между прецизност и чувствителност.

# Как ще се валидират резултатите, анализ на грешките (Резултати, Дискусия)

* **Classification Report:** Показва прецизност, чувствителност, F1-скор и точност за всеки клас.
* **Accuracy Score:** Използван за оценка на общата точност на модела.
* **Train-Test Split (80:20):** Основният метод за разделяне на данните, използван за обучение и тестване.
* **RandomizedSearchCV:** За оптимизация на хиперпараметрите и валидиране на резултатите чрез крос-валидация.

# Бъдеща работа (Заключение и бъдеща работа)

* **Интеграция на дълбоко обучение:** Използване на усъвършенствани модели като BERT (Bidirectional Encoder Representations from Transformers), които ще позволят по-добро разбиране на контекста и семантиката на текстовете.
* **Обогатяване на датасета:** Добавяне на нови данни от различни източници, включително социални мрежи и новинарски платформи, за да се подобри разнообразието и представителността на тренировъчния корпус.
* **Анализ на сложни езикови конструкции:** Разработване на техники за разпознаване на ирония, сарказъм и други форми на езикови нюанси, които често се срещат във фалшиви новини.
* **Подобрение на уеб приложението:** Разширяване на функционалността на текущото Flask приложение с възможности за анализ в реално време и интеграция с външни API източници за автоматично откриване на новини.
* **Автоматично откриване на тематични модели:** Използване на тематично моделиране (Topic Modeling) за анализ на тенденции и идентифициране на теми, които са по-склонни да бъдат обект на дезинформация.

# Свършена работа :

# Иван Михов:

# Преработване на данните и разработване на Flask приложение

В рамките на проекта е извършена обстойна преработка на данните и разработване на уеб приложение с помощта на Flask. Процесът започва със зареждане и обединяване на два основни набора от данни – „Fake.csv“ и „True.csv“, съдържащи примери за фалшиви и истински новини. След обединяването на данните е приложено почистване, включващо премахване на липсващи стойности и базова текстова обработка с цел подготвяне на данните за последващ анализ. Както и работа по премахване на стоп думи, lemmanization както и работа по pos tagging

За по-добро разбиране на структурата и разпределението на данните са създадени визуализации, включително диаграми за разпределение на категориите (истински и фалшиви новини) и облаци от думи (word clouds). Тези визуализации помагат за идентифициране на най-често срещаните думи и тенденции в текста.

Втората част на задачата включва разработването на уеб приложение с помощта на Flask. Приложението предоставя интуитивен потребителски интерфейс, който позволява въвеждането на текст на новина и получаването на моментално предсказание дали тя е фалшива или истинска. За тази цел обучените модели са интегрирани в приложението, като е осигурена стабилна работа и съвместимост с реални потребителски сценарии.

# Николай Тодоров:

# Векторизация и обучение на модели

Друг основен компонент на проекта е свързан с преобразуването на текстовите данни в числов вид (векторизация) и обучението на модели за класификация. За целта е използван методът TF-IDF (Term Frequency-Inverse Document Frequency), който позволява извличане на значими характеристики от текстовете, като се определя важността на всяка дума спрямо контекста на новината.

С цел подобряване на ефективността на модела и намаляване на размерността на данните е приложена техника за понижаване на размерността – Truncated Singular Value Decomposition (SVD). Освен това е реализиран персонализиран клас за извличане на допълнителни езикови характеристики, включително броя на съществителните, глаголите, прилагателните и наречията, съотношението на стоп думите, средната дължина на думите и анализ на полярността на текста чрез VADER анализатор.

След приключване на фазата на извличане на характеристики са обучени няколко модела за машинно обучение, сред които Multinomial Naive Bayes. За оптимизация на хиперпараметрите е използван RandomizedSearchCV, който подпомага откриването на най-подходящите настройки за всеки модел. Накрая ефективността на моделите е оценена чрез метрики като точност (accuracy) и F1-скор, което позволява прецизно измерване на тяхната ефективност при класифициране на новините като фалшиви или истински.

# Линк към гитхъб репо :

<https://github.com/Mihov01/fakeNewsDetection>