News Classification Project Documentation

Chapter 1: Introduction

The News Classification Project is a machine learning initiative focused on distinguishing between real and fake news articles using various classification models. This project addresses the critical challenge of misinformation in today's digital age by implementing multiple classification approaches to identify authentic news from fabricated content.

Project Overview

The primary objective is to develop and compare different machine learning models that can effectively classify news articles as either real (1) or fake (0). This binary classification task utilizes the Fake_real_news dataset, which contains a collection of news articles with their corresponding labels.

Dataset Description

The Fake_real_news dataset consists of:

- News articles with varying lengths and topics

- Binary labels: 1 for real news, 0 for fake news

- Text features that include headlines and article content

Project Goals

- Implement and compare various machine learning algorithms for news classification

- Evaluate model performance using standard metrics

- Develop robust preprocessing techniques for text data

- Create an efficient classification system for practical applications

Chapter 2: Team Roles and Contributions

The project was executed through collaborative effort, with team members sharing responsibilities for data preprocessing and loading while individually focusing on specific model implementations.

Shared Responsibilities

- Data loading and preprocessing pipeline development

- Feature engineering and text vectorization

- Dataset splitting and validation strategy

Individual Contributions

Ahmed Abdelnaser

Models implemented:

- Random Forest Classifier

- Logistic Regression

- Naive Bayes

- LSTM (Long Short-Term Memory) Neural Network

John Ashraf

Models implemented:

- Linear Support Vector Machine (SVM)

- Random Forest Classifier

- Logistic Regression

Mikel Somual

Models implemented:

- Random Forest Classifier

- Logistic Regression

- PyTorch-based Neural Network

Abdelrahman Said

Models implemented:

- Random Forest Classifier

- Logistic Regression

- PyTorch-based Neural Network

Mahmoud Samy

Focused on:

- Data preprocessing and cleaning

- Feature extraction techniques

- Text normalization methods

Workflow Structure

While the team shared the initial data preprocessing pipeline, each member worked independently on their assigned models, allowing for:

- Parallel development of different approaches

- Independent model optimization

- Diverse perspective on solution strategies

- Comprehensive comparison of model performances