ANKARA UNIVERSITY COM2501 – INTRODUCTION TO DATA SCİENCE

Real or Fake News Detection using Machine Learning

COMPUTER ENGİNEERING DEPARTMENT

AUTHORS
BORAN ÖMER DOĞAN : 21290594

Introduction:
In today's digital age, the rapid proliferation of information has brought with it a significant challenge: the spread of misinformation. Fake news, characterized by its intention to deceive, can have profound implications on public opinion, political processes, and societal trust. Addressing this issue is crucial for maintaining the integrity of information and supporting an informed citizenry. The "Real or Fake News Detection" project aims to tackle this challenge by developing an advanced machine learning model capable of distinguishing between authentic and fake news articles.

The project is motivated by the critical need to combat misinformation, enhance media literacy, and support informed decision-making among the public. Furthermore, it contributes to the growing field of natural language processing (NLP) research, exploring innovative approaches to text classification and model interpretability.

To achieve its objectives, the project employs Naive Bayes and Support Vector Machine (SVM) classifiers, leveraging their strengths in handling text data. The methodology encompasses comprehensive data pre-processing, including text cleaning, tokenization, stemming, and TF-IDF vectorization. These steps are essential for transforming raw text into a structured format suitable for model training.

Models are trained and evaluated on a labeled dataset, with their performance assessed through metrics such as accuracy, precision, recall, and F1-score. The project also addresses various challenges, such as class imbalance, feature engineering, model generalization, and interpretability, employing techniques to enhance robustness and reliability.

By developing a tool to identify fake news, this project aims to promote information integrity and contribute significantly to the field of machine learning and NLP, providing a foundation for future advancements in the detection and analysis of misinformation.

Dataset:
Description:
The dataset utilized in this project comprises over 14,000 instances of both real and fake news articles, providing a substantial corpus for model development and evaluation. With a significant volume of labeled data, the dataset ensures comprehensive coverage across various contexts and topics, enhancing the robustness and generalization capability of the machine learning models.

Source Diversity: Drawing from a vast array of sources, the dataset reflects the diversity of news content available across different platforms and media outlets.

Substantial Volume: With over 14,000 instances, the dataset offers a sizable collection of real and fake news articles, enabling thorough model training and evaluation.

Labeling Accuracy: Each news article is meticulously labeled as either real or fake, ensuring the accuracy and reliability of the dataset for training and testing purposes.

Scope of Data: The dataset encompasses news articles spanning a wide range of topics, including politics, health, entertainment, and more, capturing the breadth of information dissemination in digital media.
The utilization of this extensive dataset underscores the project's commitment to leveraging ample data resources to develop robust and reliable machine learning models for fake news detection.

Methodology:

Technical Background:

Text Pre-processing: *The raw text data undergoes several pre-processing steps, including lowercasing, tokenization, stemming, and removal of stopwords and non-alphabetic characters. These steps help standardize the text data and reduce dimensionality.*

Feature Extraction: *TF-IDF vectorization is used to convert the pre-processed text data into numerical feature vectors. TF-IDF (Term Frequency-Inverse Document Frequency) assigns weights to words based on their frequency in the document and inverse frequency across documents, capturing the importance of each word in the context of the entire corpus.*

Machine Learning Models: *Two classification algorithms, Naive Bayes and Support Vector Machine (SVM), are employed for fake news detection. Naive Bayes is a probabilistic classifier based on Bayes' theorem and assumes independence between features, while SVM aims to find the hyperplane that best separates the data points into different classes.*

Challenges:

Class Imbalance: *The dataset may exhibit class imbalance, with a significantly larger number of real news articles compared to fake ones. Addressing class imbalance is crucial to prevent the model from being biased towards the majority class.*

Feature Engineering: *Selecting relevant features and representing textual data effectively pose challenges. The choice of features, such as n-grams or word embeddings, impacts the model's performance.*

Model Generalization: *Ensuring that the trained models generalize well to unseen data is essential. Overfitting, where the model learns to memorize the training data rather than generalize patterns, is a common challenge.*

Interpretability: *While accuracy is important, understanding the decision-making process of the models is also valuable. Interpretable models enable users to comprehend why a particular prediction was made, enhancing trust and usability.*

Proposed Implementation Scheme:

Data Pre-processing: *Perform text cleaning, tokenization, stemming, and vectorization using TF-IDF.*

Feature Engineering: Experiment with different feature representations, such as uni-grams, bi-grams, and word embeddings, to capture semantic information effectively.

Address Class Imbalance: *Utilize techniques such as oversampling, undersampling, or generating synthetic samples to balance the distribution of classes in the dataset.*

Model Training: *Train Naive Bayes and SVM classifiers on the pre-processed and balanced dataset. Tune hyperparameters using techniques like grid search or random search to optimize model performance.*

Evaluation and Analysis: *Evaluate the trained models using metrics like accuracy, precision, recall, and F1-score on both training and test datasets. Analyze the confusion matrix to understand the types of errors made by the models.*

Interpretability: *Employ techniques such as feature importance analysis or model visualization to interpret the decision-making process of the classifiers and gain insights into the characteristics of real and fake news articles.*
By following this implementation scheme, the project aims to develop effective and interpretable models for fake news detection, addressing the technical challenges associated with text classification and model generalization.

Results & Discussion:
Evaluation Results:
The accuracy scores of the trained models are as follows:

Naive Bayes:

Training Data Accuracy: 97.82%
Test Data Accuracy: 95.50%

SVM:

Training Data Accuracy: 99.91%
Test Data Accuracy: 98.89%
These scores indicate the robust performance of both Naive Bayes and SVM classifiers in correctly classifying news articles as real or fake.

Analysis:
Further analysis includes examination of precision, recall, and F1-score metrics, along with visualization of accuracy scores on training and test datasets. Insights gained from this analysis provide valuable information for model refinement and future iterations.

Resources:
https://www.kaggle.com/c/fake-news/data?select=train.csv