Project Report

SPAM NEWS DETECTION

Submitted By : Biswajit Sahoo

University : ITER-SOA

Department : Computer Science and Enginerring

Instructor: Plasmid

Abstract

This project focuses on building a machine learning model capable of detecting whether a given news article is genuine or fake. Using a dataset of over 25,000 labeled news entries — both real and fake — the model was trained with Natural Language Processing (NLP) techniques and evaluated for performance. This type of application is crucial in today's digital world where misinformation and fake news can spread rapidly through various media channels.

Introduction

In the digital era, information spreads faster than ever before. Unfortunately, the same applies to misinformation and fake news. The purpose of this project is to identify and classify news as spam (fake) or legitimate using

machine learning techniques. The model was trained using real-world news datasets and developed using Python and Jupyter Notebook. Such systems are essential in maintaining information integrity, combating fraud, and promoting media literacy.

. Motivation .

As frauds and misinformation increase, the ability to quickly and automatically identify fake or spam news is vital. By developing this model, we contribute to public awareness and help ensure that people receive credible and authentic news. This project aims to be a step toward a safer information-sharing environment.

Dataset and Tools Used

The model was trained using two primary datasets: one containing genuine news articles and the other consisting of fake or spam news. Together, these datasets comprise over 25,000 records.

Tools and technologies used:
- Programming Language: Python
- Environment: Jupyter Notebook
- Libraries: pandas, numpy, nltk, re, sklearn, TfidfVectorizer, naive_bayes

. Model Implementation .

1. The datasets were imported and inspected for completeness.
2. Missing and unusable data fields were cleaned.
3. Stop words and repetitive words were

removed to improve model clarity and performance.
4. The text data was then converted into numerical format using TfidfVectorizer.
5. A Naive Bayes classifier was trained on the processed data.
6. The model was tested for accuracy and readiness for real-world predictions.

. Challenges Faced .

- A significant portion of the dataset had missing or incomplete values that needed to be cleaned.
- Repetitive and irrelevant stop words increased data size and reduced efficiency, requiring advanced preprocessing.
- Converting text to numeric format alone was not sufficient; vectorization was required for better model understanding.

Possible Improvements

Although the model performs effectively, its performance can be further enhanced by:

- Incorporating a larger and more diverse dataset to improve accuracy.
- Developing a user-friendly front-end interface to allow real-time news input and classification.

. Applications .

- Social media platforms for content filtering
- News aggregators for fact-checking
- Browser extensions to warn users about unreliable sources
- Educational tools to teach students about media literacy

. Conclusion .

Spam and fake news pose a serious threat to society, as they can influence opinions, cause

panic, and spread misinformation. Through this project, a working spam news detection model was successfully developed and trained on real-world data. With further improvements and deployment, such systems can serve as crucial tools in combating digital misinformation.