Ai4Ai

Project Overview

Bill Park,Daphne He,Kacey Choi

Team Introduction

Bill Park
b2park17@uw.edu

Daphne He
daphnehe@uw.edu

Kacey Choi
kchoi22 @uw.edu

Problem Context

In today's digital landscape, the quick diffusion of information via internet platforms has made it difficult to distinguish between real news and misleading stories. With the development of advanced artificial intelligence (AI) technology, the possibility to generate realistic yet misleading material has increased, complicating the issue of ensuring news validity for the general audience. This rising conundrum has implications for public opinion, democratic integrity, and the legitimacy of media sources.

Our capstone project arises from the urgent need to address these challenges. It focuses on the increasing difficulty that individuals and platforms face in distinguishing authentic news from deceptive content created by both humans and machines. This problem context is critical as the ability to effectively identify and classify misinformation is becoming a pivotal skill in the digital era. Misinformation can sway public opinion, manipulate stock markets, and even influence election outcomes, highlighting the profound societal necessity for effective tools to combat the spread of false news.

Problem Statement

This project focuses on comparing various machine learning models, including SVM, Logistic Regression, Naive Bayes, and XGBoost, to identity and categorize news articles as Human Real, Human Fake, Machine Real, or Machine Fake. By identifying the best-performing model, we aim to enhance the detection and prevention of misinformation, ensuring the public has access to accurate and reliable news.

Objective:

- ❏ Compare the efficacy of various machine learning models to determine the most effective approach for identifying & categorizing news articles.

Goals:

- ❏ Accurately classify articles into the 4 categories.
- ❏ Enhance the detection and prevention of misinformation through advanced machine learning techniques.

Impact:

- ❏ Ensure public access to credible and reliable news by improving the accuracy of misinformation detection.
- ❏ Address the growing challenge posed by AI-generated content, which can mimic genuine news articles and potentially mislead the public.

Key Research Insights

Support Vector Machine (SVM):

- Insight: Highest accuracy and stability in most categories
- Reason: Handles high-dimensional text data effective

Logistic Regression:

- Insight: Promising in training, overall poor accuracy in validation
- Reason: Prone to overfitting and sensitive to data distribution

XGBoost:

- Insight: Overall high accuracies in all categories
- Reason: Gradient boosting framework captures patterns effectively.

Naive Bayes:

- Insight: Reasonable training performance, but poor validation performance
- Reason: Simplicity and assumption of feature independence.

Ethical Considerations

- Training Data Biases: We need to ensure that the training data does not reflect biases that could lead to unfair treatment of certain types of news.
- Accuracy: Misclassifying real news as fake news (false positive) or fake news as real news (false negative) can have significant consequences.
- Transparency: Ensuring model explainability is important for accountability. There also needs to be clear communication of how they will be used and their limitations.
- Governance: There needs to be more robust mechanisms for monitoring, evaluation, and addressing any adverse effects that arise from their use.

Next Steps Beyond Capstone

- Continue to review and revise preprocessing steps
- Further fine tune hyperparameters for each model, and perform cross validation with more data to avoid overfitting.
- Evaluate and fine tune other machine learning models and compare them to the current top performing model.
- Implement techniques to increase the explainability of the models, such as LIME or SHAP for model interpretation.
- Develop ethical guidelines for the use of the model for AI fake news detection, addressing fairness, accountability, transparency, and privacy.
- Engage with stakeholders like internet users and researchers to gather feedback and insights.