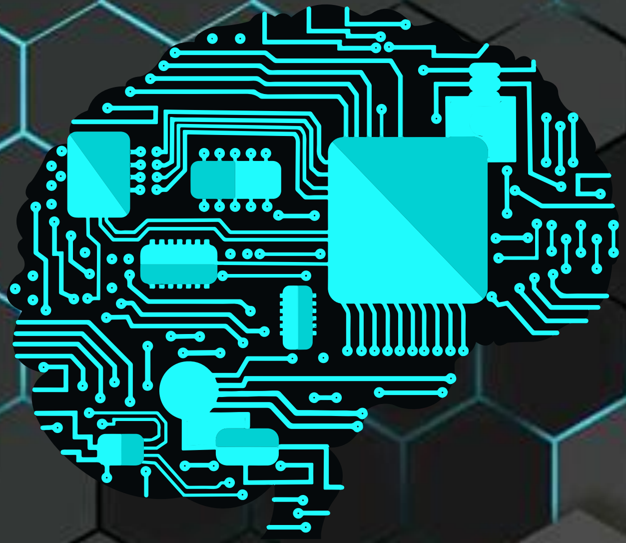
Email-Spam-Classification_prediction

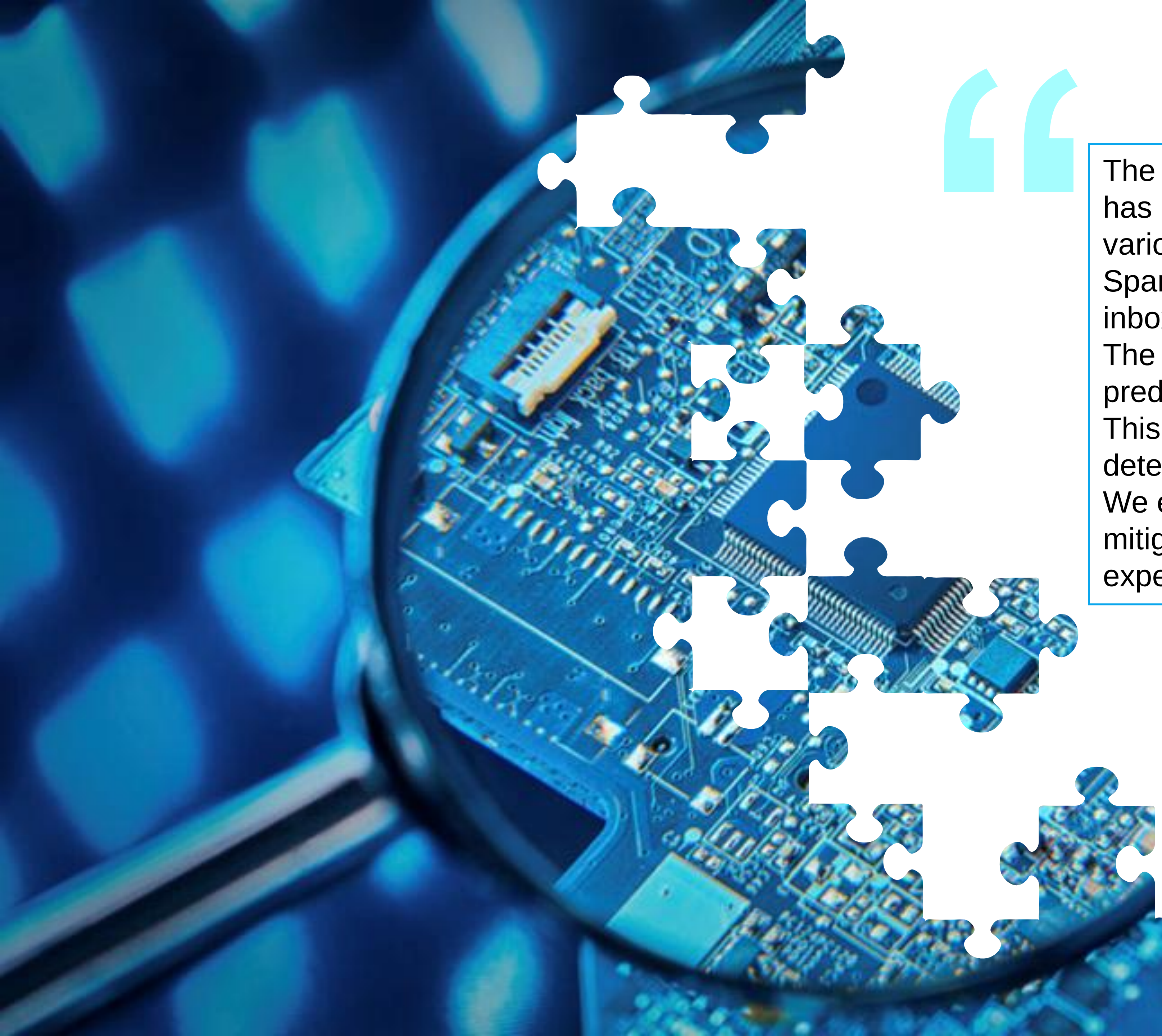
A Machine Learning Approach

SAHIL KUMAR



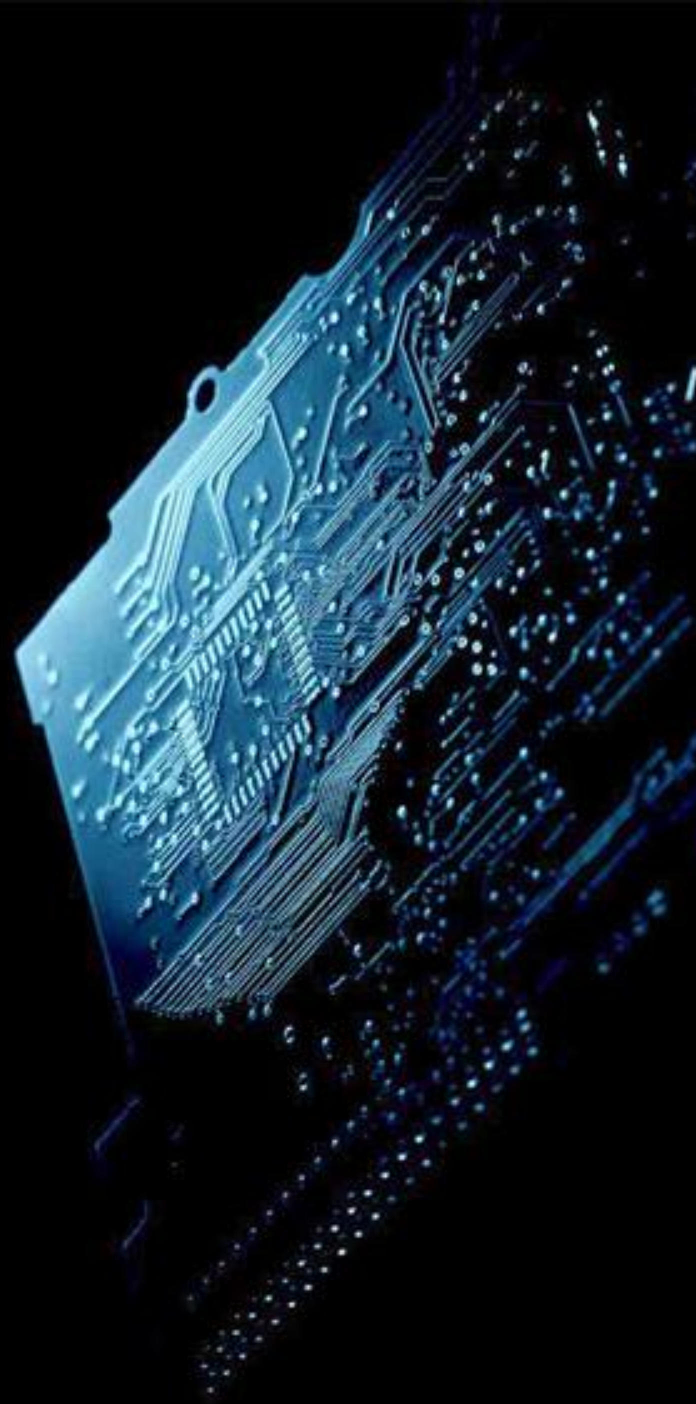
Introduction

Welcome the audience to the presentation on "Spam Prediction on Sparse Text."
Briefly introduce the topic and its significance in the digital age.
Highlight the ever-increasing volume of textual data and the need for automated spam detection.
Explain that this presentation will explore a machine learning approach to address this challenge.
Set the tone for an informative and engaging discussion on spam prediction using sparse text data.



Problem Statement

The exponential growth of digital communication has led to a surge in spam messages across various platforms.
Spam poses a significant threat by cluttering inboxes, spreading malware, and deceiving users.
The challenge is to develop effective spam prediction models to filter out unwanted content.
This presentation addresses the problem of spam detection in sparse text data.
We explore machine learning techniques to mitigate the impact of spam and enhance user experience.



Data Collection

Collecting reliable data is a crucial step in building an effective spam prediction model.

We obtained a diverse dataset consisting of text messages, emails, and social media content.

Data sources include user reports, publicly available corpora, and web scraping.

The dataset contains both spam and non-spam examples for training and evaluation.

Data preprocessing involves text cleaning, tokenization, and feature extraction.

Anonymization and privacy considerations are maintained throughout the data collection process.

Model Evaluation

Assessing the performance of our spam prediction model is essential to ensure its effectiveness.
We utilize various evaluation metrics, including accuracy, precision, recall, and F1-score.
Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) are used for model comparison.
Cross-validation helps in estimating the model's generalization performance.
We emphasize minimizing false positives to avoid classifying legitimate messages as spam.
Continuous monitoring and reevaluation are part of our model maintenance strategy to adapt to evolving spam tactics.

Results

Presenting the outcomes of our spam prediction model.
Highlighting the key performance metrics such as accuracy, precision, recall, and F1-score.
Displaying a comparison of our model's performance against baseline models.
Demonstrating the effectiveness of our approach in accurately detecting spam messages.
Providing real-world examples of messages correctly classified as spam and legitimate.
Sharing insights into the practical implications and benefits of our model's results.

CODE IN USE

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report

# Load the CSV file
data = pd.read_csv("emails.csv")

# Split the dataset into features (X) and target (y)
X = data.drop(columns=["Email No.", "Prediction"])  # Exclude non-
relevant columns
y = data["Prediction"]

# Split the dataset into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Create and train a Multinomial Naive Bayes classifier
classifier = MultinomialNB()
classifier.fit(X_train, y_train)
```

```
# Make predictions on the testing data
y_pred = classifier.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")

# Generate a classification report with more metrics
report = classification_report(y_test, y_pred,
target_names=["Not Spam", "Spam"])
print("Classification Report:")
print(report)
```

Acknowledgments

We would like to express our gratitude to the individuals and resources that made this project possible:

Kaggle: We are thankful to Kaggle for providing the Titanic dataset, which served as the foundation of our analysis and model building.

Open-Source Community: Our project was greatly enhanced by the open-source tools, libraries, and resources contributed by the data science community.

Educational Platforms: Special thanks to online learning platforms, tutorials, and courses that equipped us with the necessary skills to complete this project.

https://github.com/Sahil-Kumar0/Email-Spam-Classification_prediction.git

THANK YOU

SAHIL KUMAR
EMAIL-ID:sahilsourabh1@gmail.com