Fake News Detection with Apache Spark

BHAVYA JAIN 2022UCA1838

DIVYANSH SAHARAN 2022UCA1840

KUSHAGRA AGARWAL 2022UCA1842

The Problem: Fake News

Impact

- Fake news undermines trust and influences public opinion.
- It spreads disinformation, often with harmful consequences.

Real-world Consequences

- Social unrest
- Economic losses
- Public health crises

Data Acquisition and Overview

Dataset Source

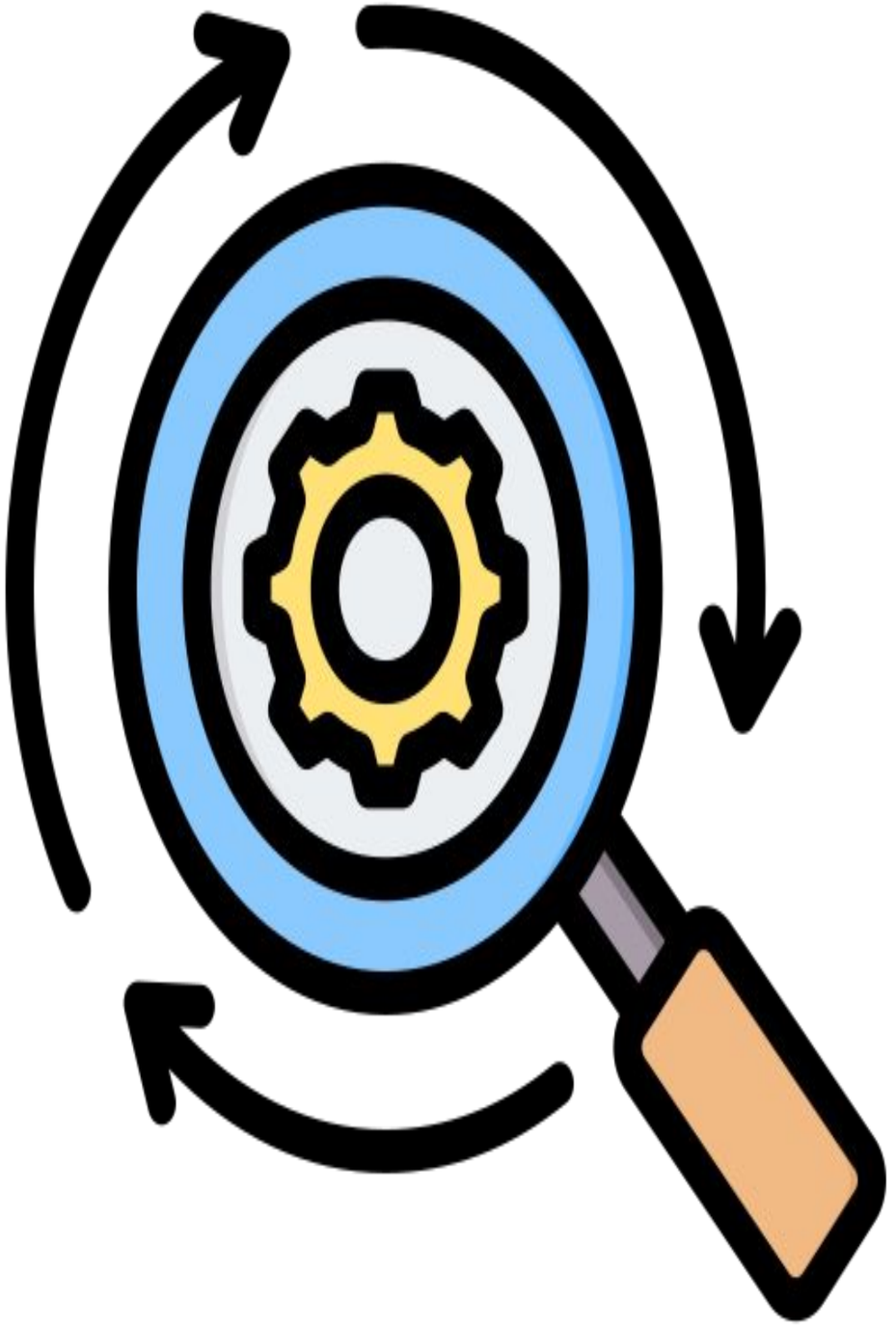
Kaggle datasets of William LiFFERTH

Scale

The dataset includes 20,000 articles, totaling over 100MB.

Key Fields

- Title
- Author
- Article text
- Labels (real/fake)



Exploratory Data Analysis (EDA)

Feature Identification

Identification of relevant features.
Example: The feature ‘Author’ is irrelevant in the context of NLP.

Missing Value Analysis

Checked for NULLs across features and assessed their influence on class label distribution.

Key findings: News Articles that have at least one of the key features missing have a high chance of being fake.

Preprocessing Steps

1 **Text Cleaning**

Removal of non-alphanumeric characters and conversion to lowercase.

2 **Stopword Removal**

Using NLTK stopword list to remove common words.

3 **Stemming**

PorterStemmer algorithm applied to reduce words to their root form.

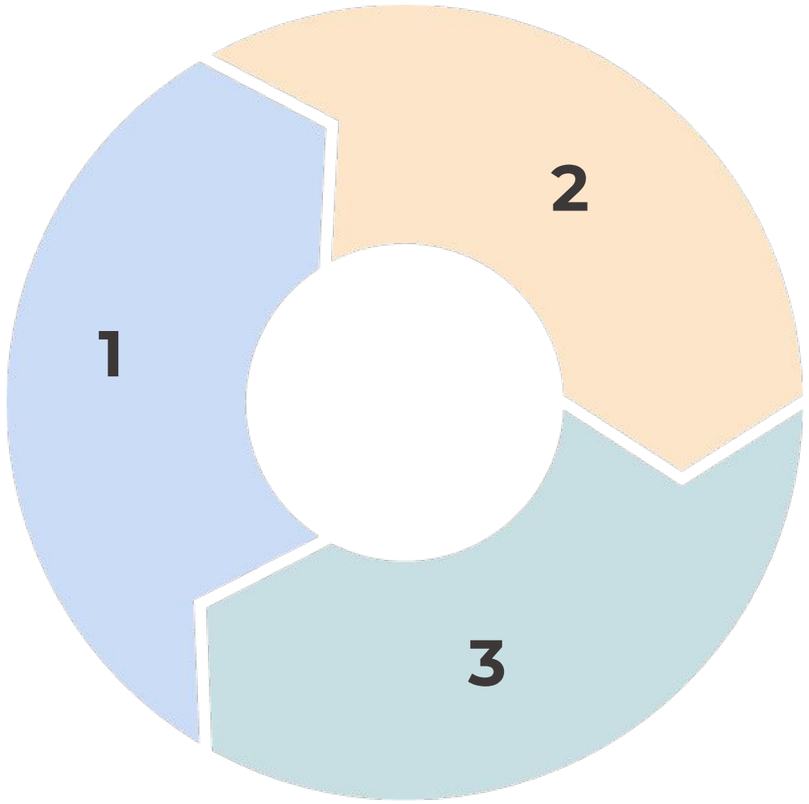
4 **Vectorization**

TF-IDF with specific parameters

Main Model

Algorithm

Naive Bayes, suited for text data.



Implementation

Spark MLlib library used for distributed training.

Hyperparameter Tuning

Grid search and cross-validation for optimization.

Apache Spark: The Engine Behind Scalability

Spark Core

Distributed and parallel processing

Spark MLlib

Scalable machine learning library.

Spark SQL

Structured data processing with DataFrames.

Apache Spark is a unified analytics engine. It provides speed, scalability, and fault tolerance. It's 10x faster than Hadoop.

Results and Performance

92.5%

Accuracy

97.2%

Precision

93%

F1-Score

Impact

1

Improved Public Awareness

Helps people access accurate information and make informed decisions.

2

Enhanced Trust in Media Platforms

Platforms that actively combat fake news are perceived as more trustworthy.

3

Safer Online Environment

Reduces the chances of inciting violence, scams, or harmful ideologies