SAPIENZA
UNIVERSITÀ DI ROMA

Big Data Computing
2021-2022

Fake News Detection using Pyspark

A presentation by: Albiona Guri

SAPIENZA
UNIVERSITÀ DI ROMA

Outline

1. Introduction
2. Dataset
3. Models
4. Experimental results
5. Conclusion
6. A functional Application
7. Contribution

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

- The main objective of the project is to determine if a given text data by the user is fake or real.
- It consists of a binary classification problem.
- In order to automate the news qualifying procedure I implemented 4 machine learning algorithms to predict the news.

SAPIENZA
UNIVERSITÀ DI ROMA

Dataset

- The dataset is derived from Kaggle resource and is composed by text data.
- It consists of around 44900 training examples.
- The ratio between minority class and majority is particularly small.

SAPIENZA
UNIVERSITÀ DI ROMA

Models

1. Logistic Regression
   - Logistic Regression
2. Random Forest
   - Random Forest Classifier
3. Multilayer Perceptron
   - MultilayerPerceptronClassifier
4. Transformers
   - Happy Transformer