Title Slide:

My project topic is Fake News Detection using machine learning.

Introduction Slide:

I chose this topic to understand the concept of machine learning techniques and algorithms. The end result of this project will be a User Interface in which the user will enter an article link and the system will provide a prediction on the authenticity of the article in percentage. A supervised learning approach has been taken to achieve this.

The project has been split into 3 major phases. First is the exploratory data analysis and data cleaning / preparation phase. Second phase is building the machine learning model. Third phase is to utilise the model in a UI.

Research ML Slide:

Machine Learning and AI has seen a rapid rise in the last few years. It is being used in various different industries such as healthcare, retail and robotics to provide new and better solutions to existing problems. Machine Learning algorithms can be split into two main groups: Supervised and Unsupervised. In supervised machine learning, all data is labelled and the algorithms learn to predict the output from the input data. In unsupervised machine learning, all data is unlabelled and the algorithms learn the inherent structure of data without any explicitly provided labels. The chart shows the basic machine learning workflow.

Research Algorithms and Library Slide:

Technologies Used Slide:

Anaconda is a distribution of python that simplifies package management. It has a lot of tools used for machine learning in a single installation.

JupyterNotebook is an open source web application which is useful for machine learning projects.

Talend is an ETL tool which I used for cleaning and preparing the data.

Streamlit is a python library which will be used to create the UI.

Screenshots Slide 1:

The left screenshot shows the scatter plot. This is a data visualization technique which shows the biased and unbiased labelled data in the form of a scatter plot.

The right screenshot shows the creation of the training dataset. This training data is an initial set of data which helps the algorithm to learn and make predictions. Separating data into training and testing sets can minimize effects of data discrepancies.

Screenshots Slide 2:

The pandas describe function shows some basic statistical details like percentile, mean etc of the data. It is basically a statistical summary.

The screenshot on the right shows the steps to build a confusion matrix. The confusion matrix is a table which describes the performance of the classification model on the set of test data for which the true values are already known.

Screenshots Slide 3:

Using multinomial naïve bayes on the training data results in an accuracy of 97.4%.

The confusion matrix basically tells us 4 different combinations of predicted and actual values.

True Positive:

Interpretation: You predicted positive and it’s true..

True Negative:

Interpretation: You predicted negative and it’s true.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it’s false.

False Negative: (Type 2 Error)

Interpretation: You predicted negative and it’s false.