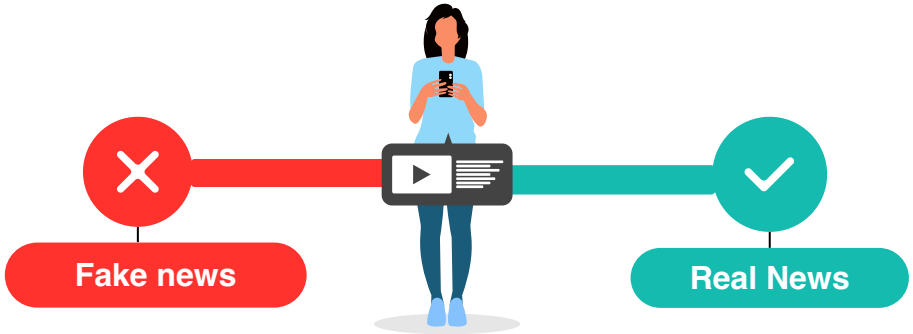
Fake News Classification

By: Group 62
Sanjitha Reddy Pathuri [A20524383]
Mohan Babu Kunchala [A20524765]

Introduction

- In today's world, Social media has become an important source of news and information.
- Despite its many advantages, large amounts of fake news/misinformation are created online for various reasons.
- Fake news/misinformation can have major consequences for individuals and society.
- It is important to detect fake news and misinformation in social media.



- In this project, we built a ML classifier that detects fake news/ misinformation based on the titles of a fake news article A and coming news article B

Tools and Methodology

Tools:

- Python language
- Sk-learn, pandas, numpy libraries
- They offer a powerful set of tools for building classification models

Methodology:

- Data preprocessing using NLP techniques
- Bag of words and TF-IDF transformer
- Model training using machine learning algorithms
- Model testing
- Model evaluation using different evaluation metrics

Steps followed

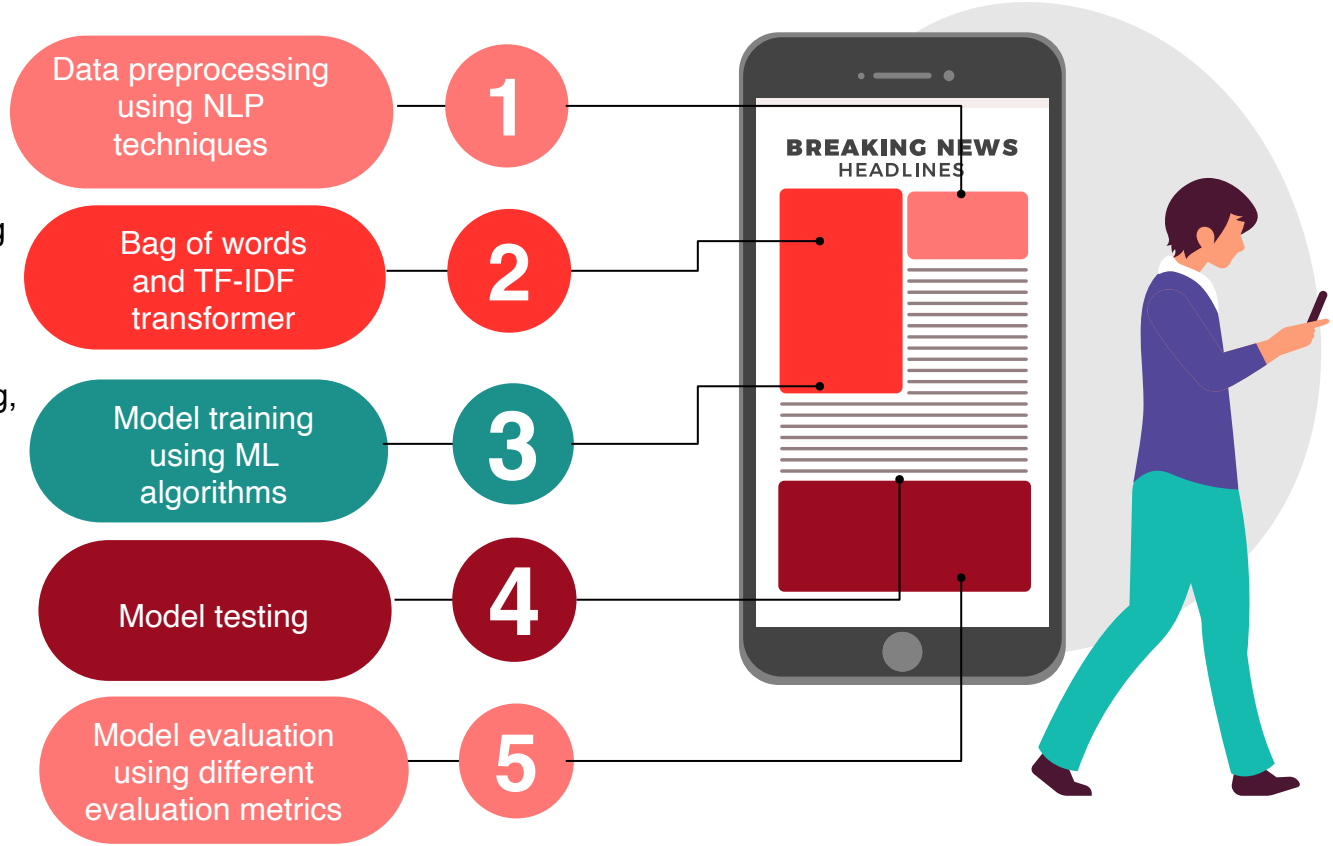
Remove duplicates, stop words, punctuations, clean text, Lemmatization

Extract relevant features using the bag-of-words model and TF-IDF vectorizer

Split the train data into training, validation (80:20) Train the model using 5 different ML Algorithms

Test the trained model using the 20% validation data

Evaluate performance of the model using metrics like accuracy, precision, recall, F1-score



Algorithms Implemented

- Logistic Regression
- Naive Bayes
- Random Forest
- Linear Support Vector Machine
- Stochastic Gradient Descent

Importance of the Algorithms Chosen

Some of the most popular and widely used classifier models in ML.

1

Each has its own unique characteristics and assumptions about the data

2

3

However, there is no one-size-fits-all approach in ML

4

The choice of a classifier model depends on the specific problem at hand and the characteristics of the data.

Comparison of the Algorithms

Algorithm	Pros	Cons
Logistic Regression	Simple and fast to train and predict. Works well with linearly separable data.	Assumes a linear relationship between the features and the outcome.
Naive Bayes	Very fast to train & predict. Performs well with high-dimensional data, small training sets.	Assumes that features are independent of each other, which may not be true.
Random Forest	Can handle complex, non-linear relationships between features and outcome, reduce overfitting by using multiple trees.	Can be slow to train and predict, especially with large datasets.
Linear Support Vector Machine (SVC)	Can handle high-dimensional data. Works well with linearly separable data.	Can be sensitive to outliers.
Stochastic Gradient Descent (SGD)	Can handle large datasets and high-dimensional data, converge quickly, especially with sparse data.	Can be sensitive to the learning rate and the initial weights.

Result metrics

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**0.81**	0.80	0.81	0.80
Naive Bayes	0.77	0.76	0.77	0.76
Random Forest	**0.85**	0.85	0.85	0.84
Linear Support Vector Machine (SVC)	**0.81**	0.81	0.81	0.81
Stochastic Gradient Descent (SGD)	0.76	0.77	0.76	0.71

We achieved the best accuracy with Random Forest
followed by Linear Support Vector Classifier , Logistic Regression

Conclusion

The final model can be used to detect fake news and misinformation in social media and help combat its negative impacts on individuals and society

THANK YOU