Phase-2

Exposingthetruthwithadvancedfake news detection powered by natural language processing

Student Name: S. Sembaruthi

Register Number: 620123106104

Institution:AVSEngineeringCollege

Department: ElectronicsandCommunication Engineering

DateofSubmission:10.05.2025

GithubRepositoryLink:https://github.com/Sembaruthi-S/Nan-Muthalvan-Project-

---

1.ProblemStatement

In the digital age, misinformation spreads rapidly through socialmediaandonlineplatforms,influencingpublicopinion and decision-making. Detecting and mitigating the spread of fake news iscrucial for safeguarding truth and integrity in information sharing. This project aims to develop arobust NLP-powered system to classify news articles as real or fake,

aidingplatformsandusersinidentifyingmisinformation.

2.ProjectObjectives

Tobuildamachinelearningmodelthataccuratelydetectsfake news using natural languageprocessing.

- Toprocessandanalyzetextualdatafromnewssourcesfor pattern recognition.

- Togenerateinsightsthathelpunderstandcharacteristicsof fake news.

3.FlowchartoftheProjectWorkflow

DataCollection->DataPreprocessing->ExploratoryData Analysis -> Feature Engineering ->Model Building and Evaluation -> Visualization and Insights -> Deployment

Data Collection

- Gathernewsarticlesandheadlinesfromreliabledatasets (e.g., Kaggle, news APIs)

Data Preprocessing

- Cleantext(removepunctuation,lowercasing)
- Tokenization
- Stopword removal
- Lemmatizationorstemming

ExploratoryDataAnalysis(EDA)

- Visualizewordfrequency
- Word clouds
- Classbalancecheck(realvsfake)

FeatureEngineering

- *TF-IDFvectorization*
- *Wordembeddings(e.g.,Word2Vec,GloVe)*
- *N-grams*

ModelBuilding

- *Trainmodels:LogisticRegression,RandomForest,SVM, LSTM*
- *Splitintotrainingandtestsets*

ModelEvaluation

- *Evaluatewithmetrics:Accuracy,Precision,Recall,F1-Score*
- *ConfusionmatrixDeployment*
- *IntegratewithawebapplicationorAPI*
- *Allowusertoinputanewsheadline/articletotest authenticity*

4.Data Description

Source:Kaggle,orscrapedfromnewswebsitesandsocial media

- Features:

- title:Newstitle

- text:Fullnewscontent

- subject:Topiccategory(e.g.,politics,worldnews)

- label:1forfake,0forreal

5.Data Preprocessing

Removalofstopwords,punctuation,andspecial charactersTokenization and lowercasing

- Lemmatizationorstemming
- VectorizationusingTF-IDForWordEmbeddings
- Handlingmissing/nullvalues

6.ExploratoryDataAnalysis(EDA)

Distributionofrealvsfakelabels:

- Mostfrequentwordsinfakevsrealnews
- Word clouds
- Articlelengthdistribution
- N-gramanalysis

7. FeatureEngineering

TF-IDFvectors

- Countvectors
- Sentimentscores
- Wordembeddings(e.g.,Word2Vec,GloVe)
- Readabilityscores

8. ModelBuilding

Train/testsplit(e.g.,80/20)

- Classificationmodels:
  - LogisticRegression
  - NaiveBayes
  - RandomForest
  - SupportVectorMachine(SVM)

- *XGBoost*
- *LSTM/GRU(DeepLearningwithKerasorPyTorch)*

Evaluationmetrics:

- *Accuracy*
- *Precision,Recall,F1-Score*
- *ConfusionMatrix*
- *ROC-AUCCurve*

9.VisualizationofResults&ModelInsights

Confusionmatrixheatmap

- *ROC-AUCcurve*
- *Precision-Recallcurves*
- *Barplotscomparingmodelperformances*
- *Wordcloudsandtokenfrequencycharts*

10. ToolsandTechnologiesUsed

Languages: Python

- *Libraries:Pandas, NumPy, NLTK, Scikit-learn, Matplotlib,Seaborn,TensorFlow/Keras,XGBoost*
- *Platforms:JupyterNotebook,GoogleColab,Kaggle*
- *Visualization: Matplotlib, Seaborn, WordCloud*
- *VersionControl:GitHub*

11. TeamMembersandContributions

Hereisaparagraphdescribingtherolesandcontributions of Sembaruthi, Shabana, Shakthi, and Naga Ishwarya in theproject"ExposingtheTruthwithAdvancedFakeNews Detection Powered by Natural Language Processing": Inthisproject,eachteammemberplayedavitalrolein ensuringacomprehensiveandaccurateapproachtofake news detection.

S.SEMBARUTHI-datacleaning

Sembaruthi took the lead in data cleaning and preprocessing, handling the removal of noise, missingvalues,andformattinginconsistenciesto ensure the dataset was ready for analysis.

J. Shabana Mirza - EDA and Feature EngineeringShabanawasprimarilyresponsiblefore xploratory dataanalysis(EDA)andfeatureengineering—she extractedmeaningfulpatterns,visualizedtrends,and transformed textual data into relevant features for the models.

G.Shakthi-ModelDevelopment

Shakthifocusedonmodeldevelopment,trainingand evaluating several machine learning algorithms, including ensemble methods and deep learning models, to classify news as real or fake with high accuracy.

R. Naga Ishwarya - Documentation and Reporting

NagaIshwaryamanageddocumentationand reporting, compiling the project's findings, preparing visual summaries,andensuringthatallaspectsoftheworkflowwere clearly presented for stakeholders and final submission.