



# 9238-MANGAYARKARASI COLLEGE OF ENGINEERING

- TITLE: FAKE NEWS DETECTION USING NLP

GUIDED BY:

- MRS.BK.HEMALATHA HOD/ECE

PREPARED BY:

S.I.SYED ALI FATHIMA

G.SATHYA

G.SAKTHII PRIYA

K.PORKALAI SELVI



# PROBLEM STATEMENT:

- FAKE NEWS DETECTION USING NLP IS IMPORTANT BECAUSE IT HELPS COMBAT THE SPREAD OF MISINFORMATION AND PROMOTES MEDIA LITERACY. BY ANALYZING LANGUAGE PATTERNS AND SOURCES, NLP CAN IDENTIFY MISLEADING CONTENT AND PROTECT USERS FROM BEING DECEIVED.
- FAKE NEWS DETECTION USING NLP CAN HELP IDENTIFY MISLEADING INFORMATION. NLP TECHNIQUES CAN ANALYZE TEXT PATTERNS, SOURCES, AND CREDIBILITY TO DETERMINE THE AUTHENTICITY OF NEWS ARTICLES.
- FAKE NEWS DETECTION USING NLP IS THAT IT HELPS MAINTAIN THE CREDIBILITY OF NEWS SOURCES. BY IDENTIFYING AND FLAGGING FAKE NEWS, NLP CAN CONTRIBUTE TO PRESERVING THE TRUSTWORTHINESS OF JOURNALISM AND ENSURING THAT PEOPLE HAVE ACCESS TO ACCURATE AND RELIABLE INFORMATION.

# THINKING PROCESS OF FAKE NEWS DETECTION:

- WE CAN FOLLOW THE DESIGN THINKING PROCESS. THIS INVOLVES:
- 1. EMPATHIZE: UNDERSTAND THE NEEDS OF USERS AND THE IMPACT OF FAKE NEWS ON SOCIETY.
- 2. DEFINE: CLEARLY DEFINE THE PROBLEM STATEMENT AND THE GOALS OF THE DETECTION SYSTEM.
- 3. IDEATE: GENERATE IDEAS FOR NLP TECHNIQUES AND ALGORITHMS THAT CAN ANALYZE TEXT PATTERNS AND SOURCES TO IDENTIFY FAKE NEWS.
- 4. PROTOTYPE: DEVELOP A PROTOTYPE SYSTEM THAT CAN PROCESS AND ANALYZE NEWS ARTICLES USING NLP TECHNIQUES.
- 5. TEST: EVALUATE THE PERFORMANCE OF THE SYSTEM USING A DATASET OF KNOWN FAKE AND REAL NEWS ARTICLES.
- 6. ITERATE: CONTINUOUSLY IMPROVE THE SYSTEM BASED ON FEEDBACK AND REFINE THE NLP MODELS AND ALGORITHMS USED.
- BY FOLLOWING THIS PROCESS, WE CAN CREATE AN EFFECTIVE AND RELIABLE FAKE NEWS DETECTION SYSTEM USING NLP.

# PHASE OF DEVELOPMENT:

- THE DEVELOPMENT OF A FAKE NEWS DETECTION SYSTEM USING NLP TYPICALLY INVOLVES SEVERAL PHASES:
- 
- 1. DATA COLLECTION: GATHER A DIVERSE AND COMPREHENSIVE DATASET OF NEWS ARTICLES, INCLUDING BOTH REAL AND FAKE NEWS EXAMPLES.
- 2. PREPROCESSING: CLEAN AND PREPROCESS THE DATA BY REMOVING NOISE, FORMATTING TEXT, AND HANDLING MISSING VALUES.
- 3. FEATURE EXTRACTION: EXTRACT RELEVANT FEATURES FROM THE TEXT, SUCH AS WORD FREQUENCIES, N-GRAMS, OR SEMANTIC FEATURES, TO REPRESENT THE ARTICLES.
- 4. MODEL DEVELOPMENT: TRAIN NLP MODELS, SUCH AS MACHINE LEARNING ALGORITHMS OR DEEP LEARNING MODELS, USING THE LABELED DATASET TO CLASSIFY NEWS ARTICLES AS REAL OR FAKE.
- 5. EVALUATION: EVALUATE THE PERFORMANCE OF THE MODELS USING APPROPRIATE METRICS, SUCH AS ACCURACY, PRECISION, RECALL, OR F1 SCORE, ON A SEPARATE TEST DATASET




6. Fine-tuning: Refine the models by adjusting hyperparameters, trying different algorithms, or incorporating additional features to improve performance.

7. Deployment: Integrate the trained models into a user-friendly application or system that can analyze and classify news articles in real-time

8. Continuous Improvement: Monitor the system's performance, gather user feedback, and incorporate updates and improvements to enhance the accuracy and effectiveness of the fake news detection system

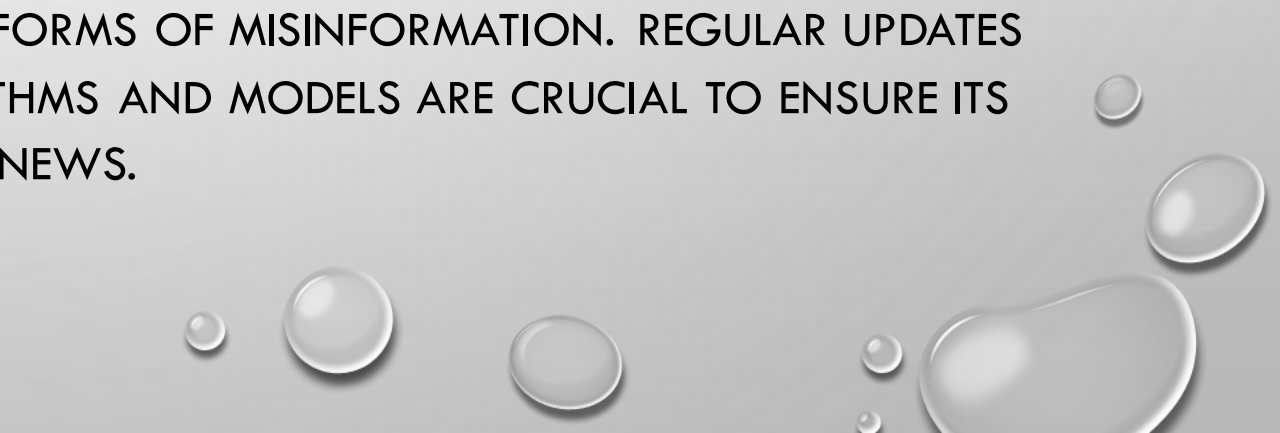
9. By following these phases, developers can create robust and reliable NLP-based fake news detection systems.

10. Another important point in the development of a fake news detection system using NLP is the need for ongoing training and updating of the models. As new types of fake news emerge, the models should be continuously trained with updated datasets to ensure their effectiveness in detecting evolving forms of misinformation. Regular updates and improvements to the system's algorithms and techniques are essential to stay ahead in the battle against fake news.





# DATA SET USED FOR FAKE NEWS DETECTION:

- THE DEVELOPMENT OF SUCH SYSTEMS TYPICALLY INVOLVES COLLECTING A DIVERSE DATASET OF NEWS ARTICLES, INCLUDING BOTH REAL AND FAKE EXAMPLES, AND LABELING THEM ACCORDINGLY. THESE DATASETS ARE OFTEN CREATED BY RESEARCH INSTITUTIONS OR ORGANIZATIONS DEDICATED TO COMBATING MISINFORMATION.
  - TO CONSIDER FOR THE DEVELOPMENT OF A FAKE NEWS DETECTION SYSTEM USING NLP IS THE NEED FOR CONTINUOUS IMPROVEMENT AND ADAPTATION. AS FAKE NEWS TACTICS EVOLVE, THE SYSTEM SHOULD BE REGULARLY UPDATED WITH NEW TECHNIQUES AND STRATEGIES TO EFFECTIVELY IDENTIFY AND COMBAT THE LATEST FORMS OF MISINFORMATION. REGULAR UPDATES AND ENHANCEMENTS TO THE SYSTEM'S ALGORITHMS AND MODELS ARE CRUCIAL TO ENSURE ITS ACCURACY AND RELIABILITY IN DETECTING FAKE NEWS.
- 

# DATA PREPROCESSING STEPS:

- TO PREPROCESS DATA FOR FAKE NEWS DETECTION USING NLP, SEVERAL STEPS CAN BE FOLLOWED:
- 1. TEXT CLEANING: REMOVE ANY SPECIAL CHARACTERS, PUNCTUATION, OR UNNECESSARY SYMBOLS FROM THE TEXT.
- 2. TOKENIZATION: SPLIT THE TEXT INTO INDIVIDUAL WORDS OR TOKENS TO ANALYZE THEM SEPARATELY.
- 3. STOPWORD REMOVAL: ELIMINATE COMMON WORDS (SUCH AS “THE,” “IS,” “AND”) THAT DO NOT CARRY SIGNIFICANT MEANING.
- 4. LOWERCASING: CONVERT ALL TEXT TO LOWERCASE TO ENSURE CONSISTENT ANALYSIS.
- 5. LEMMATIZATION OR STEMMING: REDUCE WORDS TO THEIR BASE OR ROOT FORM TO NORMALIZE THE TEXTT



6. REMOVING NOISE: REMOVE ANY IRRELEVANT INFORMATION, SUCH AS URLS, NUMBERS, OR HTML TAGS.

7. HANDLING NEGATIONS: IDENTIFY AND HANDLE NEGATIONS APPROPRIATELY TO PRESERVE THE INTENDED MEANING OF THE TEXT.

8. HANDLING IMBALANCED DATA: IF THE DATASET IS IMBALANCED (MORE REAL OR FAKE NEWS SAMPLES), APPLY TECHNIQUES LIKE OVERSAMPLING OR UNDERSAMPLING TO ADDRESS THE IMBALANCE.

THESE PREPROCESSING STEPS HELP TO CLEAN AND STANDARDIZE THE TEXT DATA, MAKING IT MORE SUITABLE FOR ANALYSIS AND IMPROVING THE ACCURACY OF THE FAKE NEWS DETECTION SYSTEM.



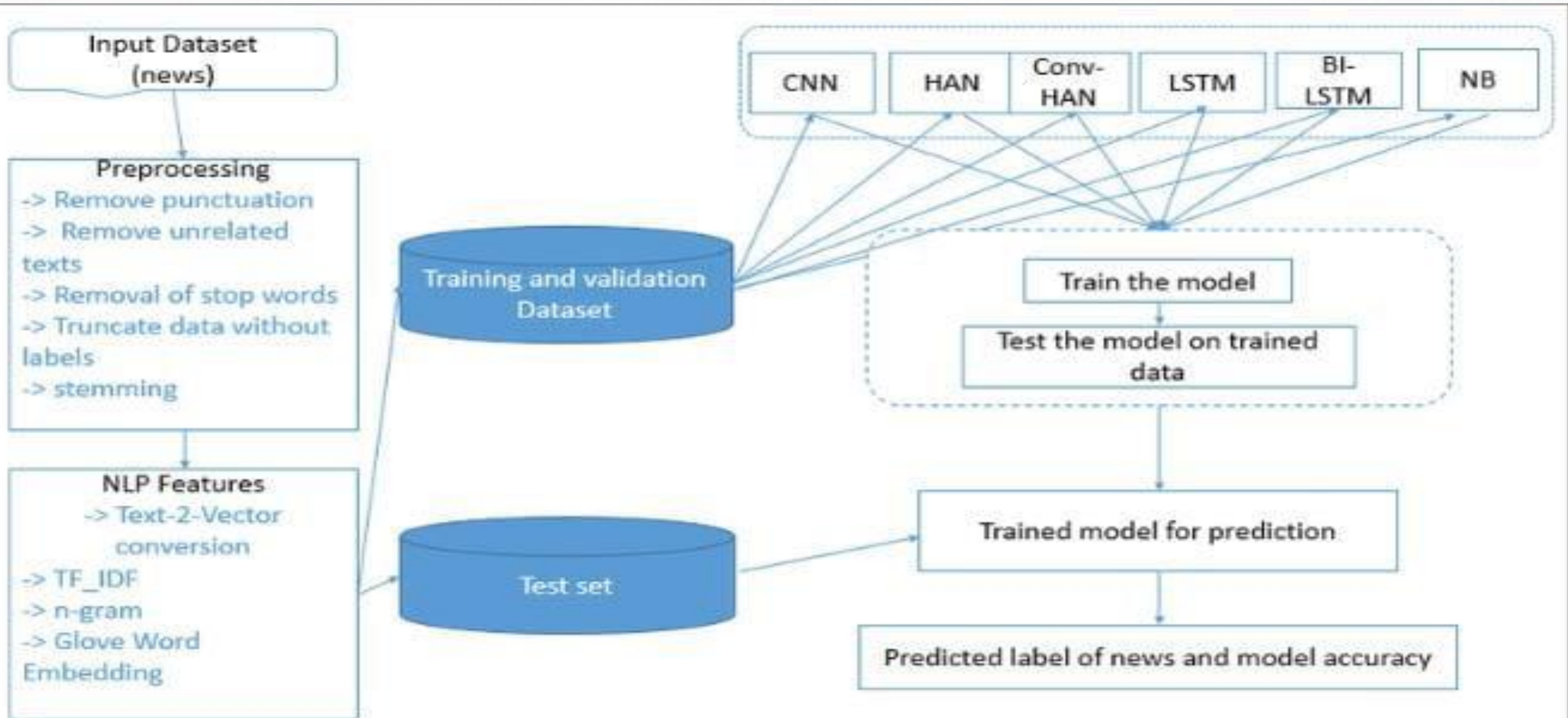


# FEATURED EXTRACTION TECHNIQUES

- ONE COMMON FEATURE EXTRACTION TECHNIQUE FOR FAKE NEWS DETECTION USING NLP IS TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY). IT CALCULATES THE IMPORTANCE OF EACH WORD IN A DOCUMENT BY CONSIDERING BOTH ITS FREQUENCY IN THE DOCUMENT AND ITS RARITY ACROSS THE ENTIRE DATASET.
- THIS TECHNIQUE HELPS TO IDENTIFY WORDS THAT ARE MORE INDICATIVE OF FAKE NEWS OR REAL NEWS. OTHER TECHNIQUES INCLUDE BAG-OF-WORDS, WORD2VEC, AND GLOVE EMBEDDINGS, WHICH CAPTURE SEMANTIC RELATIONSHIPS BETWEEN WORDS. THESE EXTRACTED FEATURES PROVIDE VALUABLE INFORMATION TO TRAIN MACHINE LEARNING MODELS FOR FAKE NEWS DETECTION.

- ANOTHER IMPORTANT POINT FOR FEATURE EXTRACTION IN FAKE NEWS DETECTION USING NLP IS THE USE OF N-GRAMS. N-GRAMS ARE CONTIGUOUS SEQUENCES OF N WORDS THAT CAN CAPTURE CONTEXTUAL INFORMATION AND IMPROVE THE UNDERSTANDING OF THE TEXT.
- BY CONSIDERING NOT ONLY INDIVIDUAL WORDS BUT ALSO COMBINATIONS OF WORDS, THE MODEL CAN BETTER IDENTIFY PATTERNS AND DISTINGUISH BETWEEN REAL AND FAKE NEWS ARTICLES. N-GRAMS CAN BE USED ALONGSIDE OTHER FEATURE EXTRACTION TECHNIQUES TO ENHANCE THE PERFORMANCE OF THE FAKE NEWS DETECTION SYSTEM.

# FLOW CHART:



# STANCE ALGORITHM

- STANCE DETECTION ALGORITHMS ARE USED TO DETERMINE THE PERSPECTIVE OR STANCE EXPRESSED IN A PIECE OF TEXT TOWARDS A PARTICULAR TOPIC OR CLAIM. THESE ALGORITHMS ANALYZE THE TEXT AND CLASSIFY IT AS SUPPORTING, OPPOSING, OR NEUTRAL TOWARDS THE GIVEN TOPIC.
- VARIOUS TECHNIQUES LIKE MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, AND DEEP LEARNING ARE USED TO DEVELOP STANCE DETECTION ALGORITHMS. THEY CAN BE HELPFUL IN UNDERSTANDING THE OPINIONS AND ATTITUDES EXPRESSED IN TEXT DATA.

# BENEFIT FOR STANCE ALGORITHM

- STANCE DETECTION ALGORITHMS HAVE SEVERAL BENEFITS. THEY CAN HELP IN:
- 1. UNDERSTANDING PUBLIC OPINION: STANCE DETECTION ALGORITHMS CAN ANALYZE LARGE VOLUMES OF TEXT DATA TO GAUGE PUBLIC SENTIMENT AND ATTITUDES TOWARDS SPECIFIC TOPICS OR CLAIMS.
- 2. IDENTIFYING MISINFORMATION: BY DETECTING THE STANCE EXPRESSED IN TEXT, THESE ALGORITHMS CAN HELP IDENTIFY INSTANCES OF MISINFORMATION OR BIASED CONTENT.
- 3. ENHANCING CONTENT MODERATION: STANCE DETECTION ALGORITHMS CAN ASSIST IN CONTENT MODERATION BY FLAGGING POTENTIALLY HARMFUL OR INAPPROPRIATE CONTENT BASED ON THE EXPRESSED STANCE.
- 4. SUPPORTING DECISION-MAKING: STANCE DETECTION ALGORITHMS CAN PROVIDE VALUABLE INSIGHTS FOR DECISION-MAKERS BY ANALYZING PUBLIC OPINION ON VARIOUS TOPICS, HELPING THEM MAKE INFORMED CHOICES.
- 5. IMPROVING CUSTOMER FEEDBACK ANALYSIS: THESE ALGORITHMS CAN BE USED TO ANALYZE CUSTOMER FEEDBACK AND REVIEWS, HELPING BUSINESSES UNDERSTAND CUSTOMER SENTIMENTS TOWARDS THEIR PRODUCTS OR SERVICES.
- THESE ARE JUST A FEW EXAMPLES OF THE BENEFITS OF STANCE DETECTION ALGORITHMS. LET ME KNOW IF YOU'D LIKE MORE INFORMATION OR HAVE ANY OTHER QUESTIONS!

# WHY NEED FOR STANCE ALGORITHM FOR DETECTON?

- STANCE ALGORITHMS ARE IMPORTANT FOR FAKE NEWS DETECTION BECAUSE THEY HELP DETERMINE THE PERSPECTIVE OR STANCE EXPRESSED IN A PIECE OF TEXT TOWARDS A PARTICULAR CLAIM OR TOPIC. BY ANALYZING THE STANCE, THESE ALGORITHMS CAN IDENTIFY WHETHER THE TEXT IS SUPPORTING, OPPOSING, OR NEUTRAL TOWARDS THE CLAIM, WHICH CAN BE VALUABLE IN IDENTIFYING AND FLAGGING POTENTIAL INSTANCES OF FAKE NEWS. STANCE DETECTION ADDS AN ADDITIONAL LAYER OF ANALYSIS TO IMPROVE THE ACCURACY OF FAKE NEWS DETECTION SYSTEMS.

# ALGORITHM FOR FAKE NEWS DETECTION:

- CERTAINLY! HERE'S A HIGH-LEVEL FLOWCHART FOR DETECTING FAKE NEWS USING NLP IN PYTHON:
- 1. START
- 2. PREPROCESS THE TEXT DATA (REMOVE STOP WORDS, PUNCTUATION, AND PERFORM STEMMING/LEMMATIZATION)
- 3. EXTRACT RELEVANT FEATURES FROM THE PREPROCESSED TEXT (SUCH AS WORD FREQUENCY, N-GRAMS, OR TF-IDF)
- 4. SPLIT THE DATASET INTO TRAINING AND TESTING SETS
- 5. TRAIN A MACHINE LEARNING MODEL (SUCH AS NAÏVE BAYES, LOGISTIC REGRESSION, OR RANDOM FOREST) USING THE TRAINING DATA
- 6. EVALUATE THE MODEL'S PERFORMANCE USING THE TESTING DATA
- 7. IF THE PERFORMANCE IS SATISFACTORY, PROCEED TO STEP 8. OTHERWISE, GO BACK TO STEP 3 AND TRY DIFFERENT FEATURES OR MODELS.
- 8. USE THE TRAINED MODEL TO PREDICT THE AUTHENTICITY OF NEW NEWS ARTICLES
- 9. END
- HOPE THIS HELPS! LET ME KNOW IF YOU HAVE ANY MORE QUESTIONS.



# TOOL AND LIBRARIES:

- IN THE PYTHON FAKE NEWS DETECTION PROJECT WE USE FOLLOWING LIBRARIES
- PYTHON-3.X
- PANDAS-1.2.4
- SCIKIT-LEARN-0.24.1
- SPACY
- STREAMLIT
- MATPLOTLIP



# PROGRAM:

```
IMPORT PANDAS AS PD
```

```
IMPORT MATPLOTLIB.PYPILOT AS PLT
```

```
IMPORT SPACY
```

```
FROM SPACY.UTIL IMPORT MINIBATCH, COMPOUNDING
```

```
IMPORT RANDOM
```

```
NLP = SPACY.LOAD('EL__CORE__NEWS__MD')
```

```
DF1 = PD.READ__CSV('../DATA/JTP__FAKE__NEWS.CSV')
```

```
DF1.REPLACE(TO__REPLACE='[ \ N \ R \ T]', VALUE='', REGEX=TRUE, INPLACE=TRUE)
```

- `IMPORT PANDAS AS PD`
- `IMPORT MATPLOTLIB.PYPILOT AS PLT`
- `IMPORT SPACY`
- `FROM SPACY`
- 
- `DEF LOAD__DATA(TRAIN__DATA, LIMIT=0, SPLIT=0.8):`
- `RANDOM.SHUFFLE(TRAIN__DATA)`
- `TRAIN__DATA = TRAIN__DATA[-LIMIT:]`
- `TEXTS, LABELS = ZIP(*TRAIN__DATA)`
- `CATS = [{“REAL”: NOT BOOL(Y), “FAKE”: BOOL(Y)} FOR Y IN LABELS]`
- `SPLIT = INT(LEN(TRAIN__DATA) * SPLIT)`
- 
- `RETURN (TEXTS[:SPLIT], CATS[:SPLIT]), (TEXTS[SPLIT:], CATS[SPLIT:])`
- `# ----- EVALUATE FUNCTION DEFINED BELOW -----`

- 
- DEF EVALUATE(TOKENIZER, TEXTCAT, TEXTS, CATS):
- DOCS = (TOKENIZER(TEXT) FOR TEXT IN TEXTS)
- TP = 0.0 # TRUE POSITIVES
- FP = 1E-8 # FALSE POSITIVES
- FN = 1E-8 # FALSE NEGATIVES
- TN = 0.0 # TRUE NEGATIVES
- FOR I, DOC IN ENUMERATE(TEXTCAT.PIPED(DOCS)):
- GOLD = CATS[I]
- FOR THE LABEL, SCORE IN DOC.CATS.ITEMS():
- IF THE LABEL IS NOT IN GOLD:
- CONTINUE
- IF LABEL == "FAKE":
- CONTINUE
- IF SCORE >= 0.5 AND GOLD[LABEL] >= 0.5:
-

- 
- IF THE LABEL IS NOT IN GOLD:
- 
- CONTINUE
- IF LABEL == "FAKE":
- CONTINUE
- IF SCORE  $\geq 0.5$  AND GOLD[LABEL]  $\geq 0.5$ :
- TP += 1.0
- ELIF SCORE  $\geq 0.5$  AND GOLD[LABEL]  $< 0.5$ :
- FP += 1.0
- ELIF SCORE  $< 0.5$  AND GOLD[LABEL]  $< 0.5$ :
- TN += 1
- ELIF SCORE  $< 0.5$  AND GOLD[LABEL]  $\geq 0.5$ :
- FN += 1
- PRECISION = TP / (TP + FP)
- RECALL = TP / (TP + FN)

# OUTPUT:

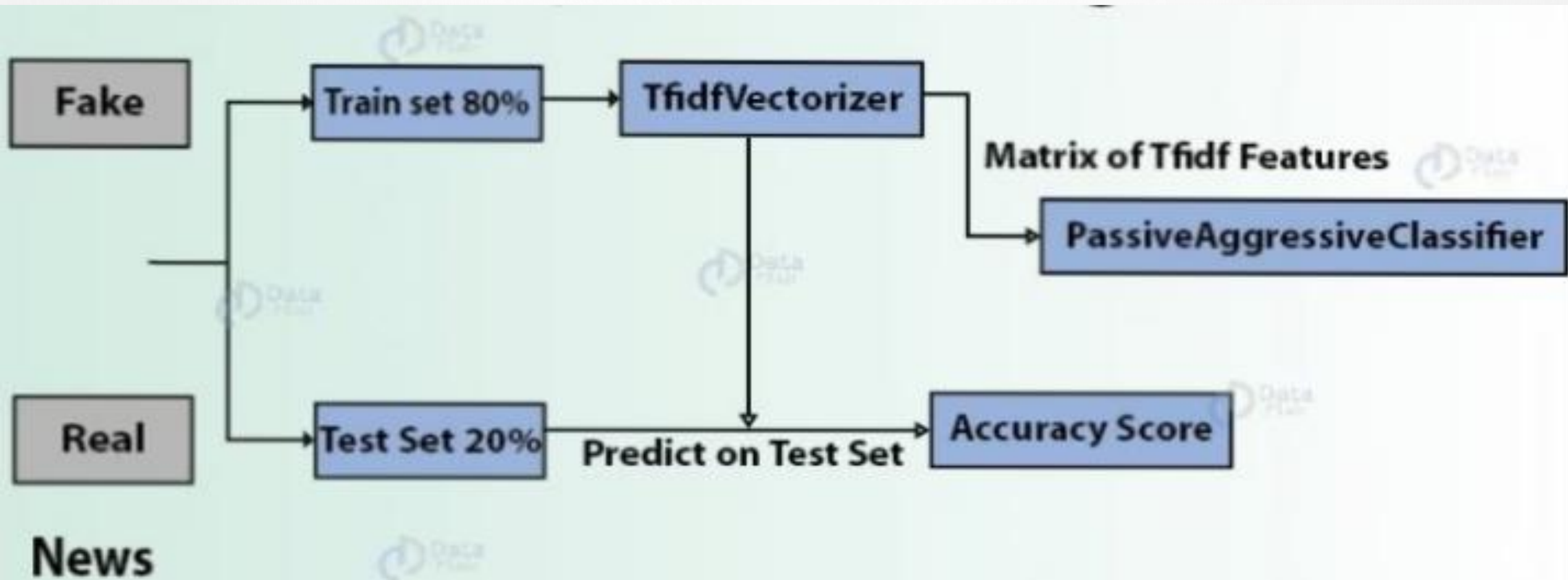
```
[2]: #Read the data
df=pd.read_csv('D:\\DataFlair\\news.csv')

#Get shape and head
df.shape
df.head()
```

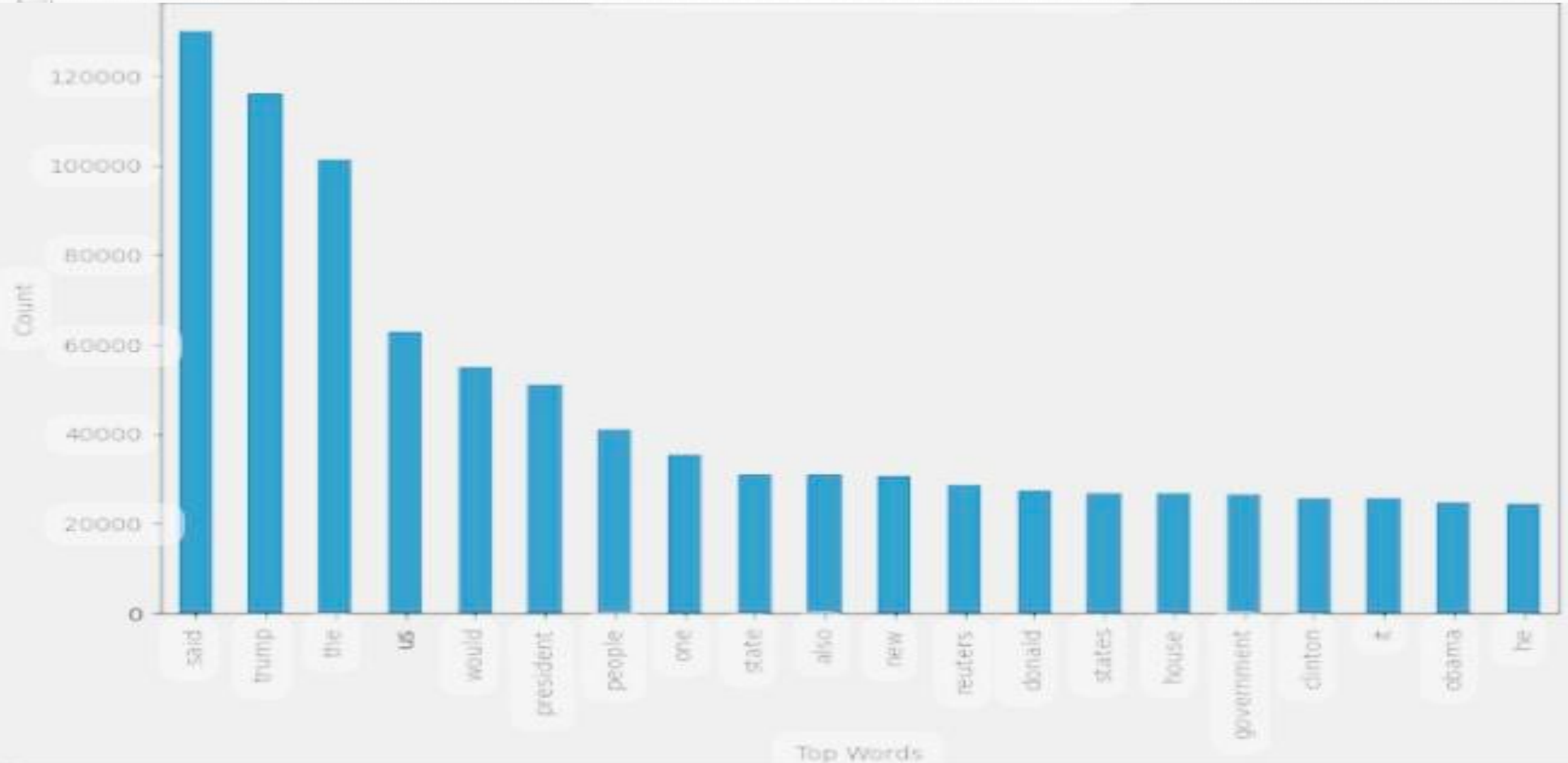
```
[2]:
```

	Unnamed: 0		title	text	label
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg LinkedIn Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy		U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...		FAKE
4	875	The Battle of New York: Why This Primary Matters		It's primary day in New York and front-runners...	REAL

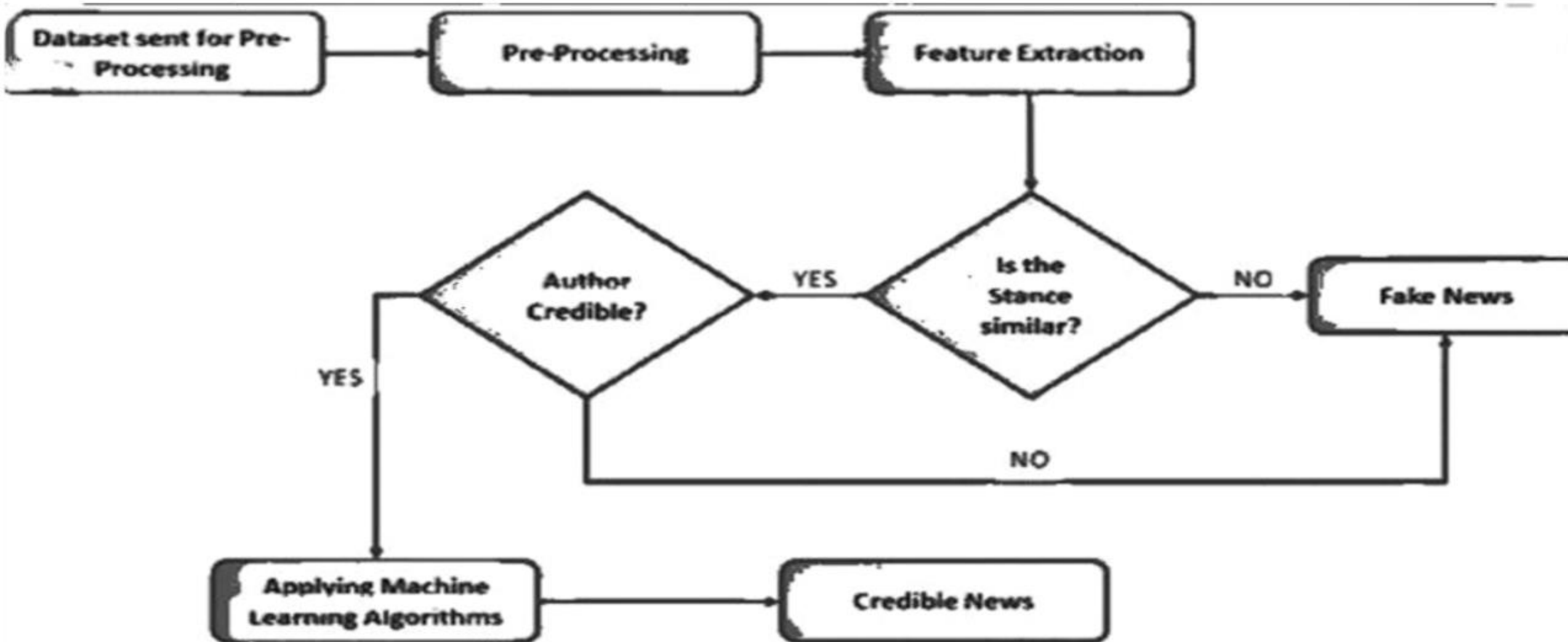
## PREDICT ON TEST SET:



# BAR CHAR FOR TOP WORD FREQUENCY



# FLOW CHART FOR DETECTING FAKE NEWS





# RESULT OF THE FAKE NEWS DETECTION

- THE RESULT OF THE FAKE NEWS DETECTION USING NLP IS TYPICALLY EVALUATED BASED ON METRICS LIKE ACCURACY, PRECISION, RECALL, AND F1 SCORE. THESE METRICS HELP ASSESS HOW WELL THE MODEL PERFORMS IN CLASSIFYING NEWS ARTICLES AS FAKE OR REAL. THE SPECIFIC RESULTS WOULD DEPEND ON THE DATASET USED, THE MODEL ARCHITECTURE, AND THE EVALUATION METHODOLOGY.

# CONCLUSION:

- IN CONCLUSION, FAKE NEWS DETECTION USING NLP IS A COMPLEX TASK THAT REQUIRES A COMBINATION OF TECHNIQUES AND APPROACHES. BY LEVERAGING NATURAL LANGUAGE PROCESSING TECHNIQUES SUCH AS TEXT CLEANING, TOKENIZATION, STOPWORD REMOVAL, AND FEATURE EXTRACTION METHODS LIKE TF-IDF AND N-GRAMS, WE CAN PREPROCESS THE DATA AND EXTRACT MEANINGFUL FEATURES.
- THESE FEATURES ARE THEN USED TO TRAIN MACHINE LEARNING MODELS THAT CAN EFFECTIVELY DIFFERENTIATE BETWEEN REAL AND FAKE NEWS ARTICLES. HOWEVER, IT'S IMPORTANT TO CONTINUOUSLY UPDATE AND ADAPT THE SYSTEM TO KEEP UP WITH EVOLVING FAKE NEWS TACTICS. OVERALL, WITH THE RIGHT TECHNIQUES AND APPROACHES, NLP CAN PLAY A CRUCIAL ROLE IN COMBATING THE SPREAD OF FAKE NEWS AND PROMOTING INFORMATION ACCURACY.