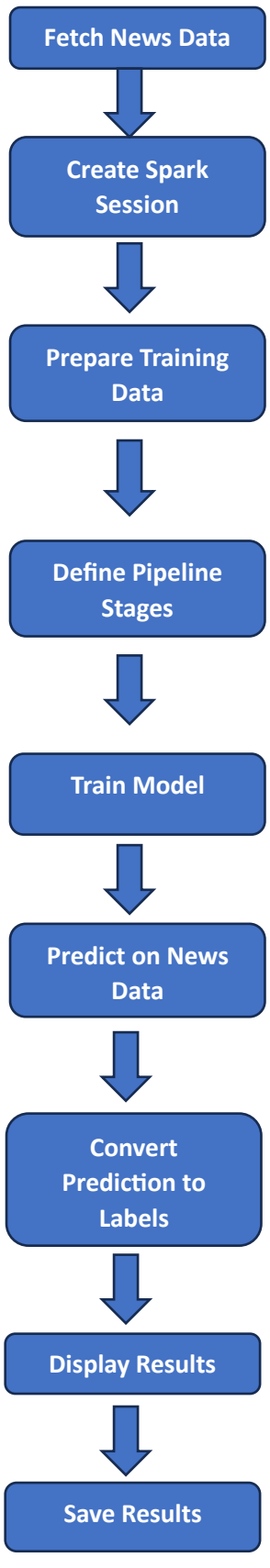
Flowchart



Explanation

Explanation of Steps

1. **Fetch News Data**
    - Use the NewsData API with requests to retrieve the latest English news articles.
    - Extract important fields (title + description) for analysis.
2. **Preprocess Text**
    - Combine title and description into a single string for each article.
    - Store the data in a list of texts.
3. **Create Spark DataFrame**
    - Initialize a SparkSession.
    - Convert the list of texts into a Spark DataFrame (df_news) with one column "text".
4. **Prepare Training Data**
    - Manually define small labeled sentences with "Positive" and "Negative" sentiment.
    - Convert this list into a Spark DataFrame (df_train).
5. **Define ML Pipeline**
    - Tokenizer: split text into words.

- StopWordsRemover: remove common words .
- CountVectorizer: convert words into numeric feature vectors.
- StringIndexer: encode labels "Positive"/"Negative" into numeric form.
- Logistic Regression: machine learning model for classification.

6. **Train Model**
   - Fit the pipeline on the training dataset (df_train).
   - The model learns to distinguish between positive and negative text.
7. Predict on News Data
   - Apply the trained model to the news dataset (df_news).
   - Predictions are generated as numeric codes (0/1).
8. Convert Predictions to Labels
   - Use IndexToString to map numeric predictions back to "Positive" or "Negative".

CODE

```
import requests
from pyspark.sql import SparkSession

API_KEY = 'pub_ad5b62b3db184cba95049d7a94e644a4'
URL = f'https://newsdata.io/api/1/news?apikey={API_KEY}&language=en'

response = requests.get(URL)
data = response.json()
articles = data.get('results', [])

# Combine title + description
texts = [(article['title'] + " " + (article.get('description') or "")) for article in articles]

# Create PySpark DataFrame
spark = SparkSession.builder.appName("NewsSentimentML").getOrCreate()
df_news = spark.createDataFrame([(text,) for text in texts], ["text"])
df_news.show(5, truncate=False)
from pyspark.sql import Row

train_data = [
    ("I love the new product launch", "Positive"),
    ("The stock market crashed today", "Negative"),
    ("The movie was fantastic", "Positive"),
    ("I am very disappointed by the service", "Negative"),
    ("Elections bring uncertainty to the market", "Negative"),
    ("This sports event is amazing", "Positive")
]

df_train = spark.createDataFrame(train_data, ["text", "label"])
df_train.show()
 from pyspark.ml.feature import Tokenizer, StopWordsRemover, CountVectorizer, StringIndexer
from pyspark.ml.classification import LogisticRegression
```

```
from pyspark.ml import Pipeline


# Pipeline stages
tokenizer = Tokenizer(inputCol="text", outputCol="words")
remover = StopWordsRemover(inputCol="words", outputCol="filtered")
vectorizer = CountVectorizer(inputCol="filtered", outputCol="features")
label_indexer = StringIndexer(inputCol="label", outputCol="labelIndex")
lr = LogisticRegression(featuresCol="features", labelCol="labelIndex")

pipeline = Pipeline(stages=[tokenizer, remover, vectorizer, label_indexer, lr])

# Train the model
model = pipeline.fit(df_train)
 from pyspark.ml.feature import IndexToString
# Transform news DataFrame
predictions = model.transform(df_news)

# Convert numeric prediction back to string label
label_converter = IndexToString(inputCol="prediction", outputCol="predicted_label",
                                labels=model.stages[3].labels)
predictions = label_converter.transform(predictions)

# Show results
predictions.select("text", "predicted_label").show(truncate=False)
  # Convert Spark DataFrame to Pandas
final_results = predictions.select("text", "predicted_label")
pandas_df = final_results.toPandas()

# Save with a different filename
pandas_df.to_csv("news_results.csv", index=False)

# In Colab: download the file
from google.colab import files
files.download("news_results.csv")
```

Images

Dashboard:

Real-Time News Sentiment Dashboard

Total Headlines

10

Positive

2

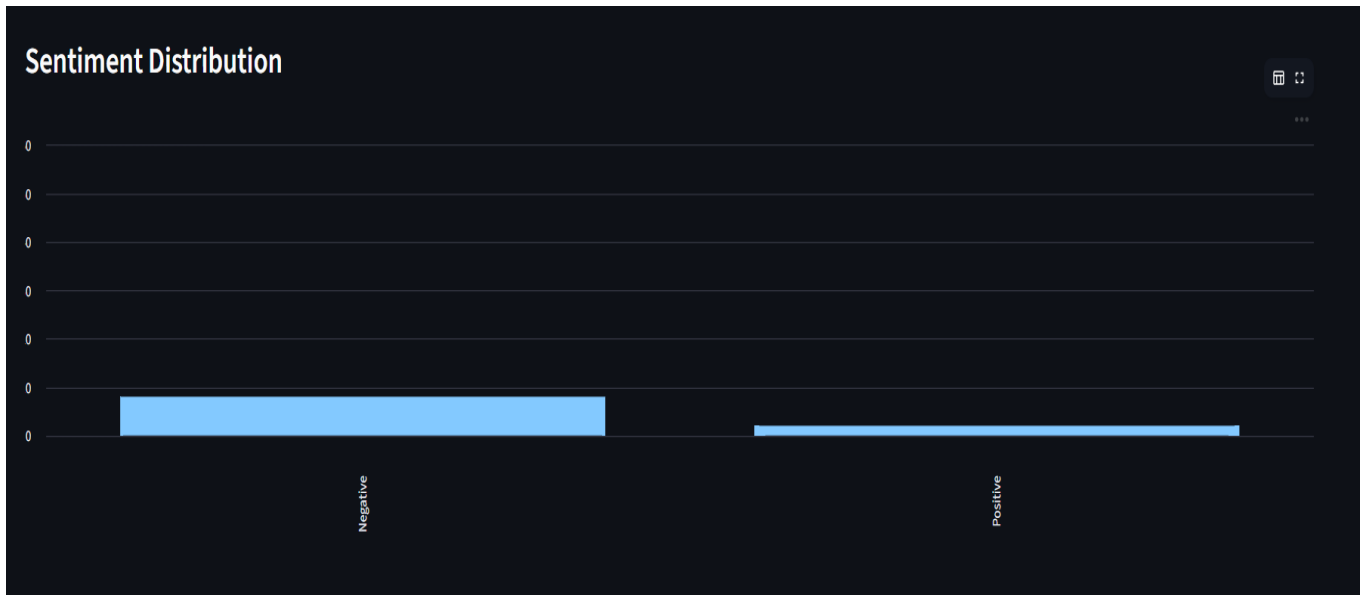
Negative

8

Latest Headlines

	text	Sentiment
0	Wagner Wealth Management LLC Has $326,000 Stake in Costco Wholesale Corporation $COST Wagner Wealth Management LLC increased its stake in	Negative
1	Wagner Wealth Management LLC Has $1.83 Million Stake in Duke Energy Corporation $DUK Wagner Wealth Management LLC boosted its stake in sha	Negative
2	Smith Salley Wealth Management Purchases 2,455 Shares of Duke Energy Corporation $DUK Smith Salley Wealth Management raised its position in I	Negative
3	Stone Summit Wealth LLC Makes New $219,000 Investment in Uber Technologies, Inc. $UBER Stone Summit Wealth LLC acquired a new position in U	Positive
4	Sonora Investment Management Group LLC Increases Position in Blackstone Inc. $BX Sonora Investment Management Group LLC lifted its position ir	Negative
5	Smith Salley Wealth Management Increases Holdings in The Home Depot, Inc. $HD Smith Salley Wealth Management raised its stake in The Home De	Negative
6	San Luis Wealth Advisors LLC Acquires 4,207 Shares of Blackstone Inc. $BX San Luis Wealth Advisors LLC increased its stake in shares of Blackstone In	Negative
7	Stone Summit Wealth LLC Reduces Position in The Home Depot, Inc. $HD Stone Summit Wealth LLC reduced its position in The Home Depot, Inc. (NY	Negative
8	Smith Salley Wealth Management Buys 2,455 Shares of Duke Energy Corporation $DUK Smith Salley Wealth Management increased its position in Du	Negative
9	San Luis Wealth Advisors LLC Makes New Investment in Merck & Co., Inc. $MRK San Luis Wealth Advisors LLC bought a new stake in shares of Merck &	Positive

Sentiment Distribution



Network URL: http://10.10.48.18:8502

Output:

	A	B	C	D
1	text	predicted_label		
2	Wagner Wealth Management LLC Has $	Negative		
3	Wagner Wealth Management LLC Has $	Negative		
4	Smith Salley Wealth Management Purcl	Negative		
5	Stone Summit Wealth LLC Makes New $	Positive		
6	Sonora Investment Management Group	Negative		
7	Smith Salley Wealth Management Incre	Negative		
8	San Luis Wealth Advisors LLC Acquires 4	Negative		
9	Stone Summit Wealth LLC Reduces Posi	Negative		
10	Smith Salley Wealth Management Buys	Negative		
11	San Luis Wealth Advisors LLC Makes Ne	Positive		
12				
13				