AI Sentiment Analysis Project

Overview

The goal of this project is to determine public sentiment towards artificial intelligence (AI) on social media platforms, specifically Reddit. The project involves collecting, preprocessing, and analyzing Reddit posts to classify them into three sentiment categories: positive, negative, and neutral. The final objective is to build a supervised learning model to predict sentiment based on these classifications.

1. Data Collection

Libraries Used:

- `praw`: Python Reddit API Wrapper for interacting with Reddit's API.
- `pandas`: Data manipulation and analysis.
- `numpy`: Numerical operations.

Authentication:

python
Copy code

```
reddit = praw.Reddit(
    client_id='AcMQ2MCSOfbrnH_eytdelw',
    client_secret='DLBxbdIj_9RX9anTTodhRV5-QJeoVg',
    user_agent='my-app by Abhivanth',
    username='Effective_Insect9336',
    password=''
)
```

Functions:

- **`search_posts(query)`**: Searches for Reddit posts based on a query and returns a dictionary with post features including titles, selftexts, scores, upvote ratios, comment counts, link flair texts, creation dates, and top comments.
- **`accumulate_data(list_queries, label)`**: Collects data for a list of queries, labels the data, and creates a DataFrame. It saves data for positive, negative, and neutral sentiments into CSV files.

Queries for Sentiment Analysis:

- **AI_IS_GOOD_QUERIES**: Queries related to positive sentiments about AI.
- **AI_IS_BAD_QUERIES**: Queries related to negative sentiments about AI.
- **NEUTRAL_QUERIES_ON_AI**: Queries related to neutral sentiments about AI.

2. Data Preprocessing

Libraries Used:

- `pandas`
- `numpy`
- `re`: Regular expressions for text cleaning.
- `nltk`: Natural Language Toolkit for stopwords and stemming.
- `sklearn`: Machine learning library for feature extraction and model building.

Functions:

- **`clean_column(col_name)`**: Cleans text data in specified columns by removing non-alphabetic characters, converting to lowercase, removing stopwords, and applying stemming.

Text Preprocessing Steps:

1. **Remove Non-Alphabetic Characters**: Retain only alphabetic characters.
2. **Convert to Lowercase**: Standardize text to lowercase.
3. **Remove Stopwords**: Filter out common words that do not contribute to sentiment.
4. **Apply Stemming**: Reduce words to their root form.

Feature Extraction:

- **`CountVectorizer`**: Converts text data into numerical vectors for machine learning.

Data Splitting:

- Splits data into training and testing sets using `train_test_split`.

3. Model Building

Models Tested:

- **Naive Bayes**: Accuracy = TBD
- **Random Forest**: Accuracy = TBD
- **Decision Tree**: Accuracy = TBD
- **Logistic Regression**: Accuracy = TBD

Metrics:

- **Confusion Matrix**: Evaluates model performance.
- **F1 Score**: Measures the balance between precision and recall.

4. Results and Evaluation

Evaluation Metrics:

- **Confusion Matrix**: To understand the model's performance across different classes.
- **F1 Score**: To measure the model's accuracy in classification.

Final Data Outputs:

- `AI_GOOD_TRAINING.csv`
- `AI_BAD_TRAINING.csv`
- `AI_NEUTRAL_TRAINING.csv`
- `TRAINING.csv`

5. Tools and Technologies

APIs:

- **Reddit API**: For data collection from Reddit.

JSON Files:

- **JSON**: Lightweight data format used for data interchange.

API Endpoints:

- Specific URLs for accessing Reddit data and features.

6. Considerations

Source Choice:

- **Reddit**: Provides anonymous and candid opinions, but lacks verified users and relies on upvote/downvote systems.

Cleaning and Preprocessing:

- Handling missing values and ensuring balanced data for training.

7. Key Terms and Definitions

- **API (Application Programming Interface)**: Set of functions and protocols for interacting with software applications.
- **JSON (JavaScript Object Notation)**: Format for transmitting data in a readable and writable manner.
- **API Endpoints**: Specific URLs for accessing particular functionalities of an API.

Conclusion

This project focuses on understanding public sentiment towards AI through Reddit posts. By collecting, preprocessing, and analyzing data, a supervised learning model is built to classify sentiments effectively. The choice of Reddit as a data source and the use of various machine learning tools provide a comprehensive approach to sentiment analysis.