**Exploratory Data Analysis and Data Preparation**

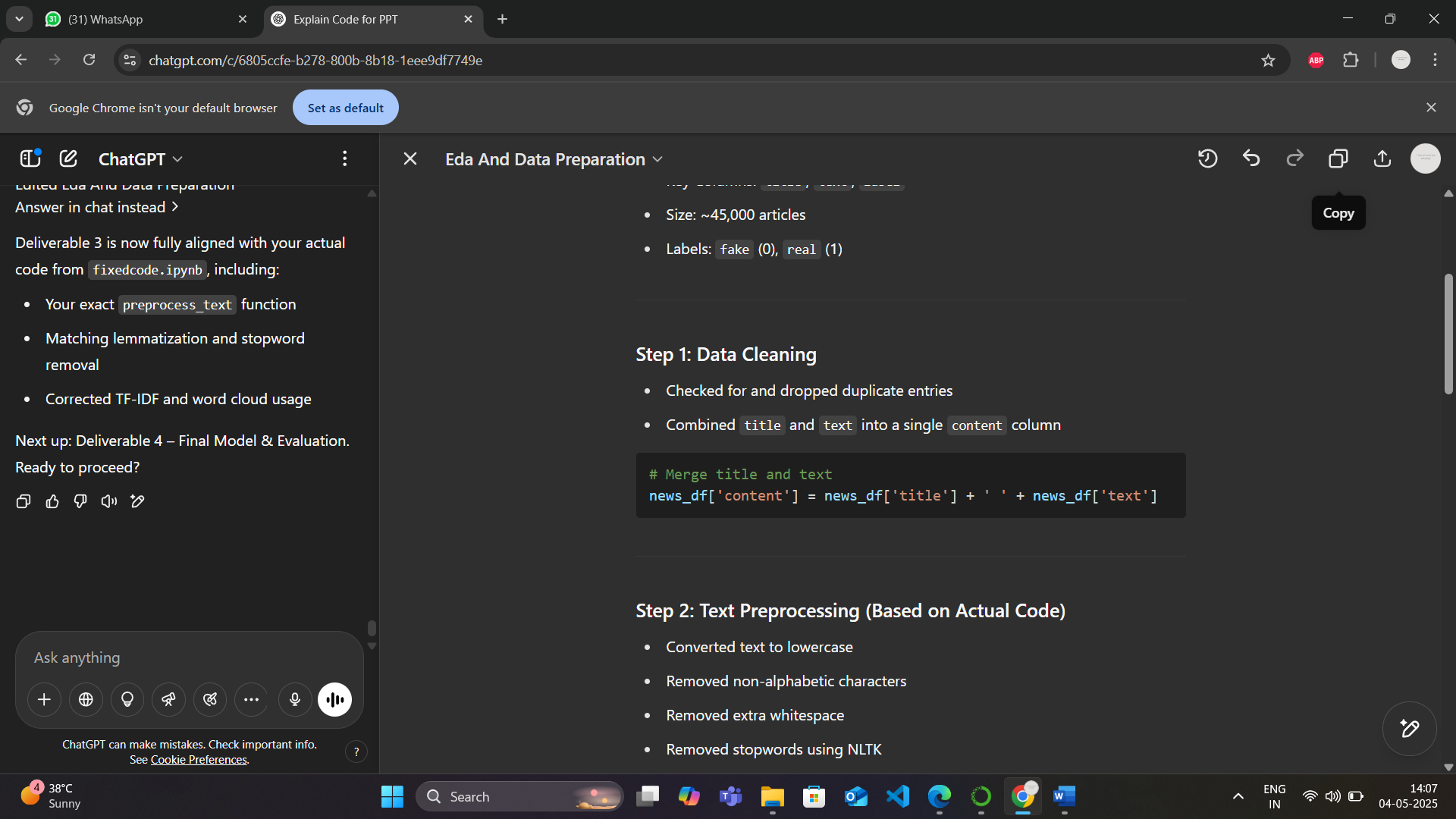
**Project: Fake News Detection Using NLP**

**Dataset Summary:**

* Source: Kaggle Fake and Real News Dataset
* Key Columns: title, text, label
* Size: ~45,000 articles
* Labels: fake (0), real (1)

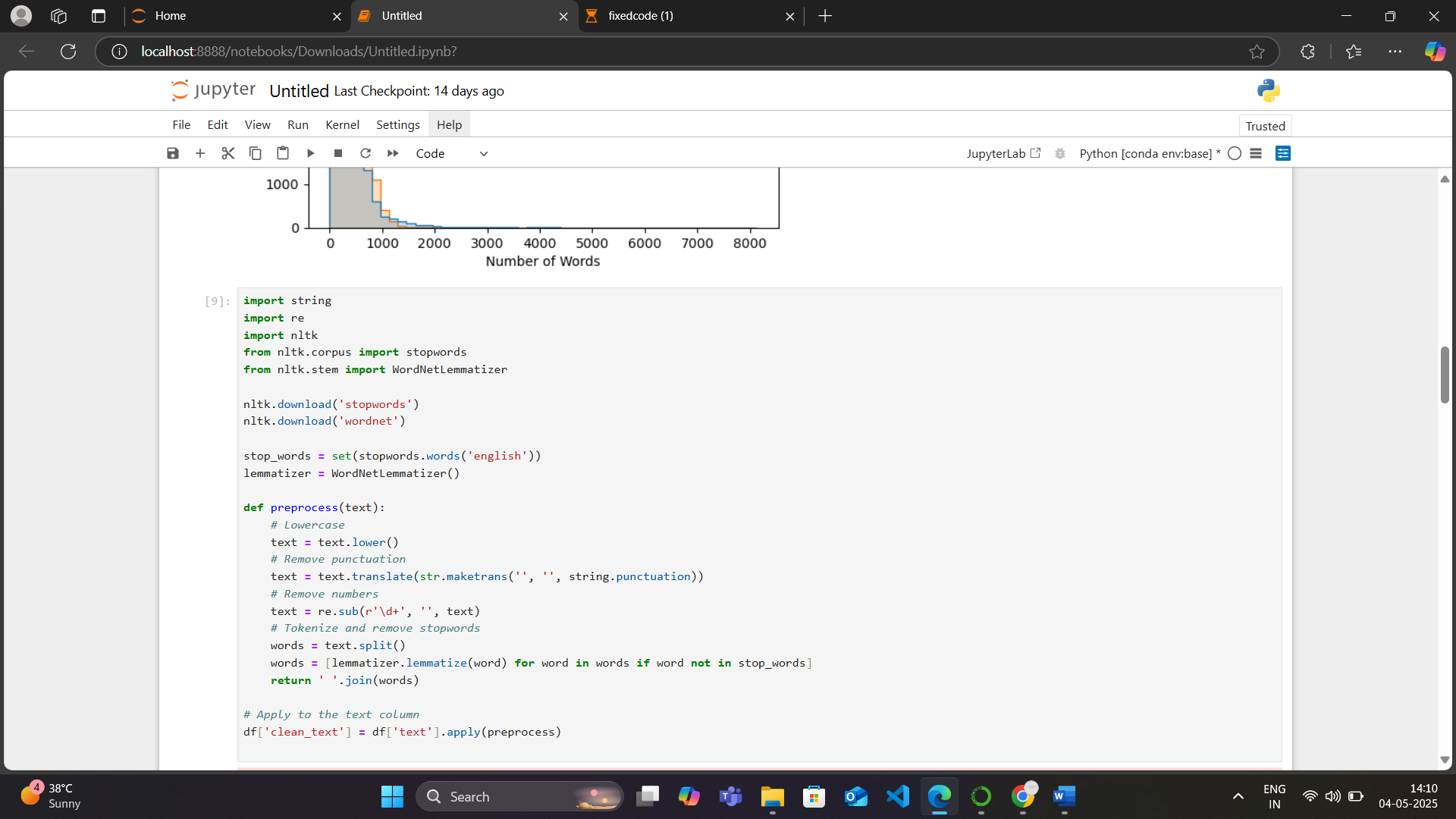
**Step 1: Data Cleaning**

* Checked for and dropped duplicate entries
* Combined title and text into a single content column



**Step 2: Text Preprocessing (Based on Actual Code)**

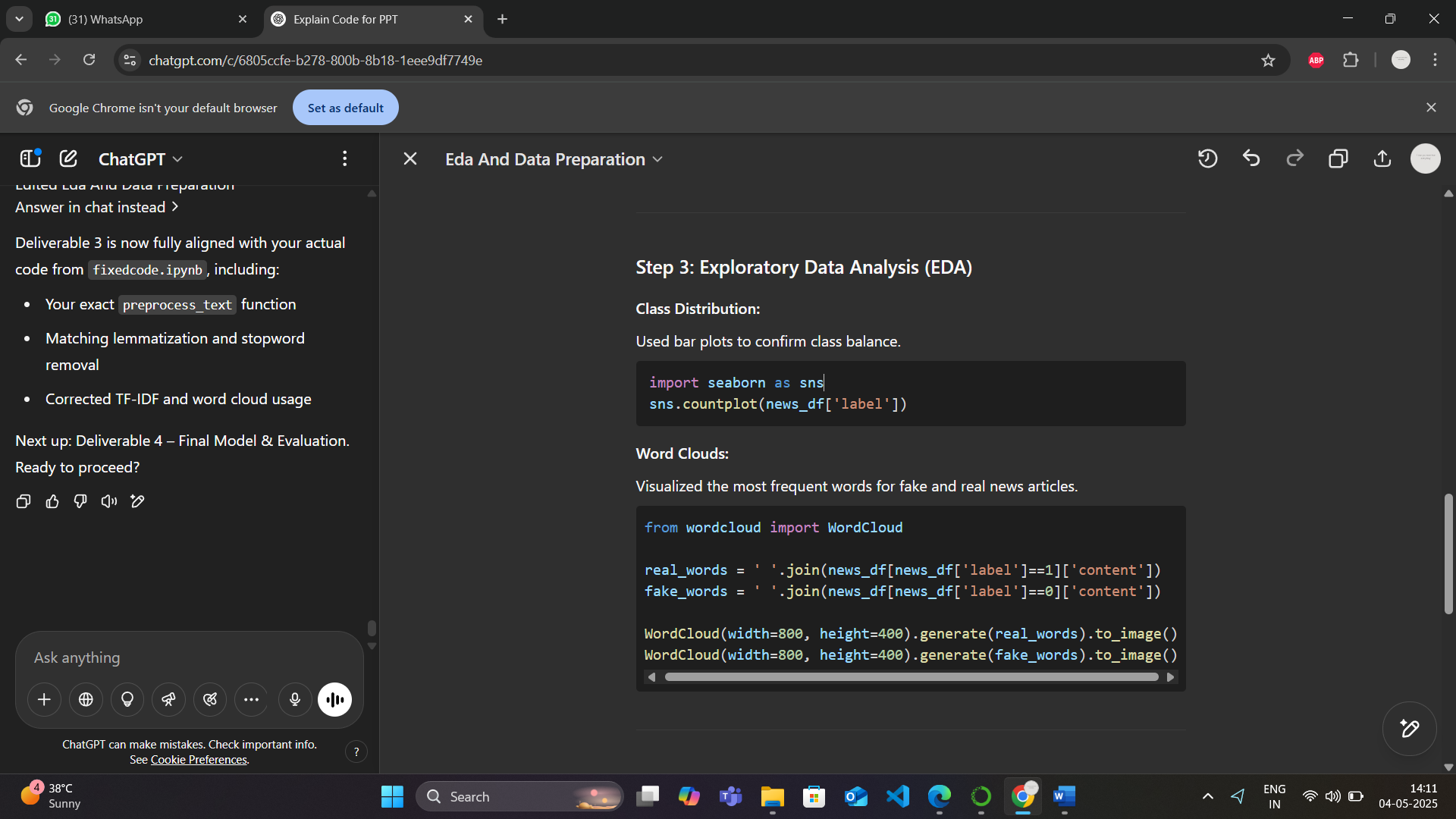
* Converted text to lowercase
* Removed non-alphabetic characters
* Removed extra whitespace
* Removed stopwords using NLTK
* Lemmatized using WordNetLemmatizer



**Step 3: Exploratory Data Analysis (EDA)**

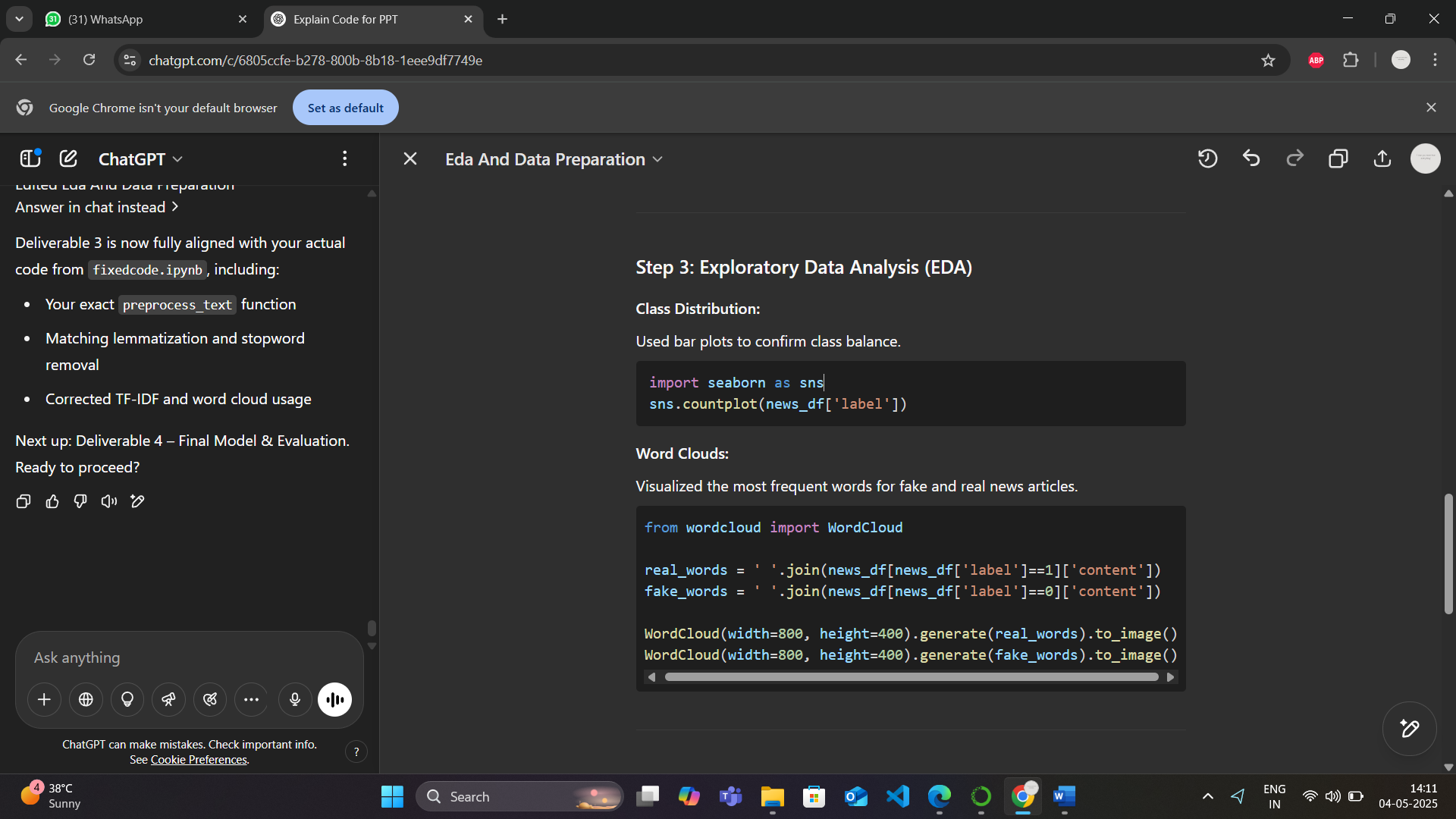
**Class Distribution:**

Used bar plots to confirm class balance.

****

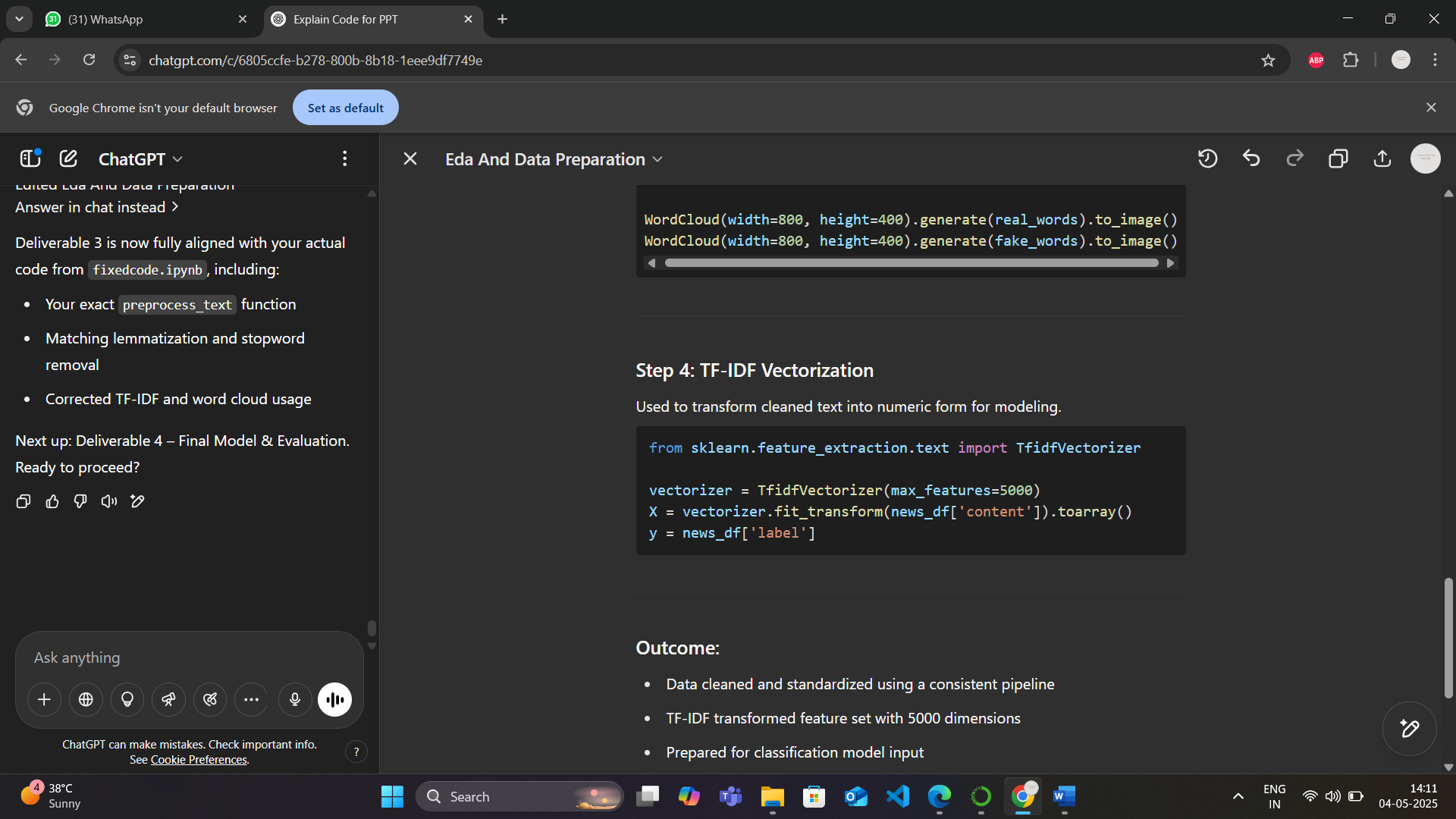
**Word Clouds:**

Visualized the most frequent words for fake and real news articles.



**Step 4: TF-IDF Vectorization**

Used to transform cleaned text into numeric form for modeling.



**Outcome:**

* Data cleaned and standardized using a consistent pipeline
* TF-IDF transformed feature set with 5000 dimensions
* Prepared for classification model input