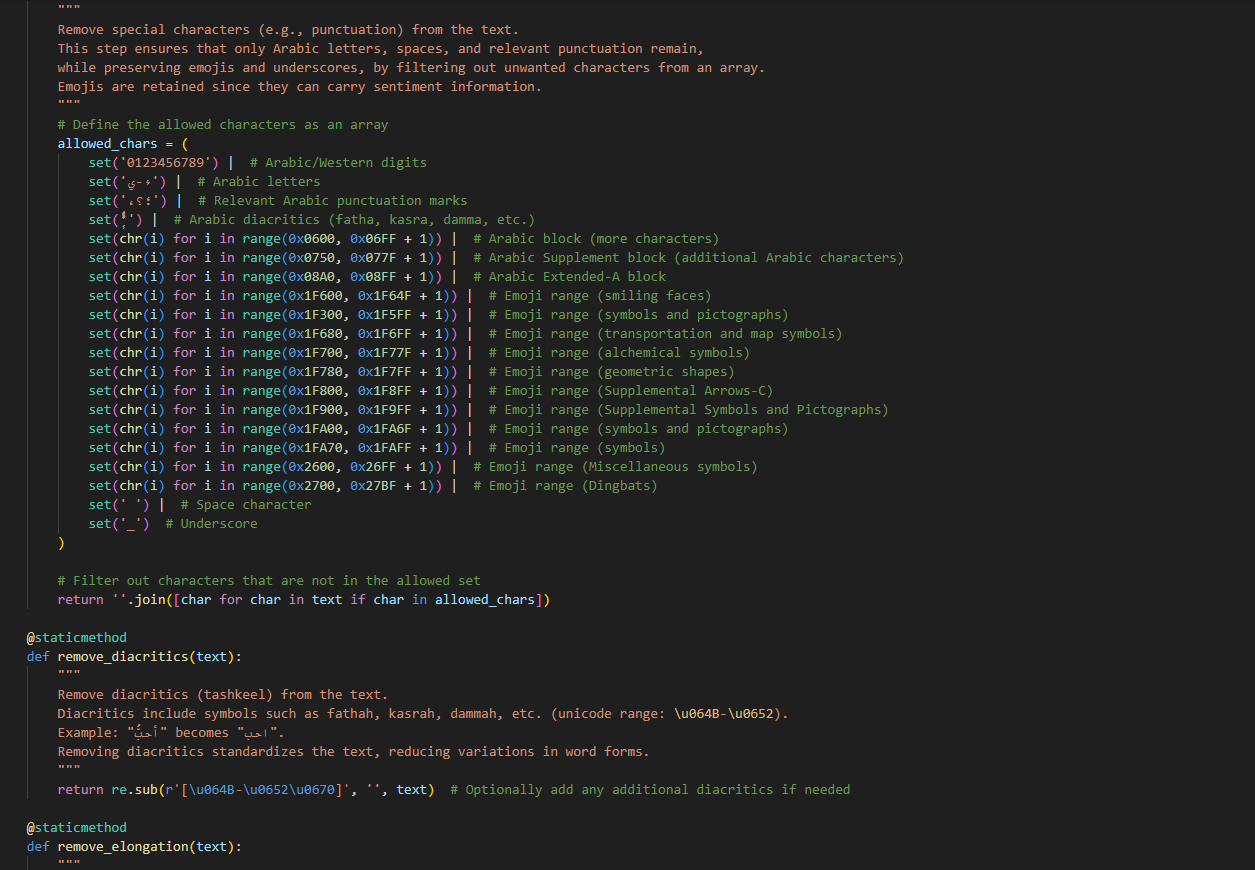


This algorithm aims to clean and preprocess Arabic text for sentiment analysis.

It incorporates several text normalization and filtering steps based on linguistic properties specific to Arabic.

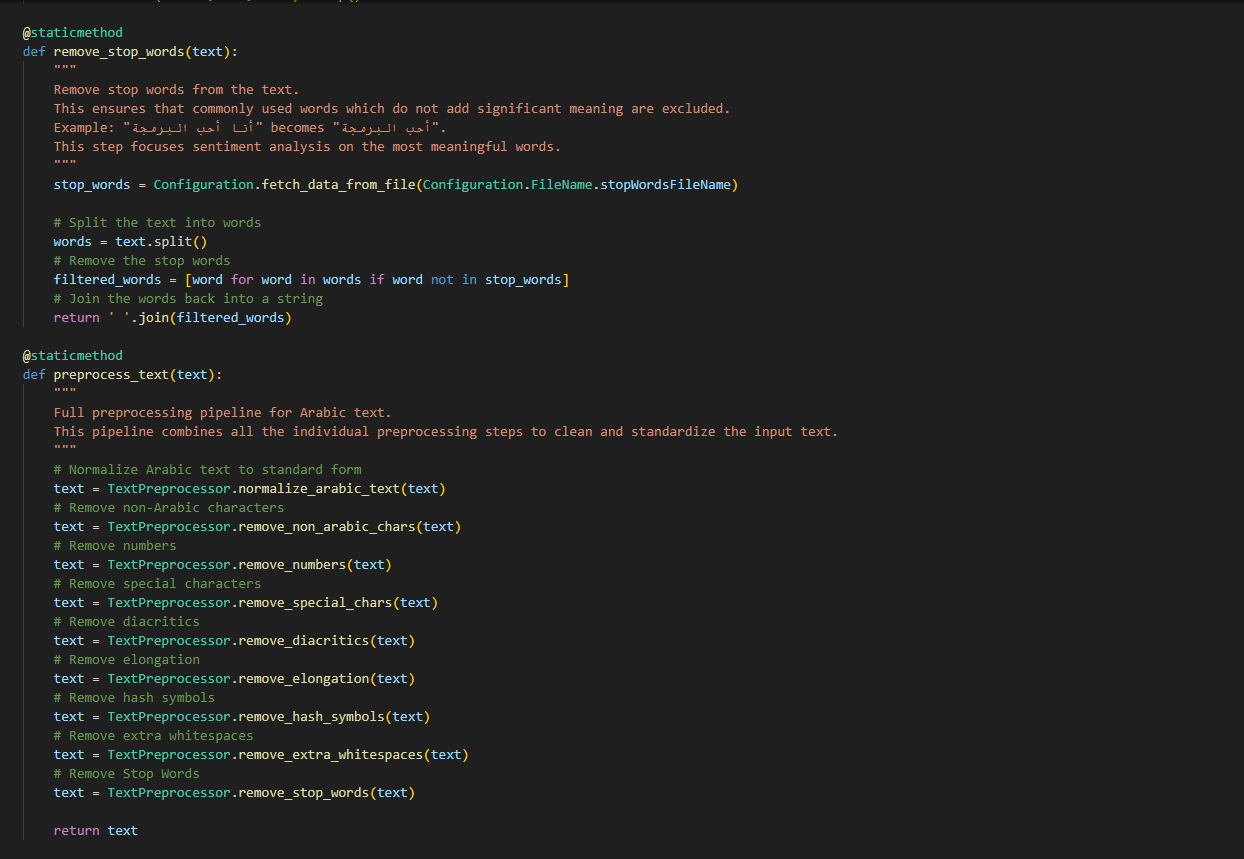


Remove special characters (e.g., punctuation) from the text.

This step ensures that only Arabic letters, spaces, and relevant punctuation remain,

while preserving emojis and underscores, by filtering out unwanted characters from an array.

Emojis are retained since they can carry sentiment information.



Full preprocessing pipeline for Arabic text.

This pipeline combines all the individual preprocessing steps to clean and standardize the input text.

**Summary of the Code**

The code implements a text preprocessing algorithm for Arabic tweets in sentiment analysis. It provides methods to clean, normalize, and structure text data effectively. The preprocessing steps are tailored to handle the unique linguistic features of the Arabic language, making it suitable for machine learning applications focused on sentiment detection.

**Description in Points**

1. **Normalization of Arabic Characters:**
   * Converts Arabic letter variations to standard forms.
   * Example: 'إ', 'أ', and 'آ' are normalized to 'ا'.
2. **Removal of Non-Arabic Characters:**
   * Removes English letters while preserving Arabic text, spaces, punctuation, and emojis.
3. **Removal of Numbers:**
   * Removes both Arabic and Western numeric digits from the text to reduce noise.
4. **Special Character Filtering:**
   * Retains Arabic letters, spaces, underscores, and emojis while filtering out irrelevant punctuation.
5. **Diacritics Removal:**
   * Removes Arabic diacritics such as fathah, kasrah, and dammah, which do not affect the word meaning in sentiment analysis.
6. **Elongation Removal:**
   * Eliminates kashida (ـ), often used for word stretching in Arabic, to standardize text.
7. **Hash Symbol Removal:**
   * Removes hash (#) and underscores (\_) typically used in social media hashtags.
8. **Whitespace Removal:**
   * Collapses multiple spaces into a single space and trims leading/trailing spaces.
9. **Stop Word Removal:**
   * Excludes common Arabic stop words that do not contribute meaningfully to sentiment analysis.
10. **Preprocessing Pipeline (preprocess\_text method):**

* Combines all the individual cleaning functions sequentially to process input text efficiently.

**Key Benefits of the Algorithm**

* Handles Arabic-specific linguistic properties.
* Preserves relevant textual elements like emojis that may contribute to sentiment.
* Reduces noise and enhances text consistency for better sentiment analysis results.
* Provides a structured and reusable approach to preprocessing Arabic text data for machine learning tasks.

Let me know if you want a refined version for documentation or presentation.