Feature Extracted from Text Contents:

**ner\_counts:** NER counts refer to the frequency of named entities (such as people, organizations, and locations) found in a text. Articles with higher NER counts, especially those mentioning popular or trending entities, are more likely to attract attention and be shared due to their relevance and authority. These counts can also improve SEO, increasing article visibility and engagement, thus influencing the popularity of online news.

 People (e.g., "Elon Musk")

 Organizations (e.g., "Tesla")

 Locations (e.g., "New York")

 Dates or times (e.g., "January 2025")

Articles with high **NER counts** featuring well-known entities may attract more attention.

**neg, neu, pos: from NER Counts**: reflect the emotional tone toward these named entities (positive or negative)

**compound ( Sentiments Scores( based on NER)-** combine sentiment scores, indicating overall sentiment toward the entities, with values ranging from -1 (negative) to +1 (positive).

**Readability Scores: Assess how easy or difficult the text is to read and understand, Types of**

**Readability Metrics:**

 **flesch\_kincaid** (**Flesch-Kincaid Grade Level**)

* Measures the text's readability based on U.S. school grade levels.
* Higher values indicate more complex text that requires higher education to understand.
* A lower score suggests easier readability for a wider audience.

 **gunning\_fog** (**Gunning Fog Index**)

* Estimates the years of formal education needed to understand the text on first reading.
* Considers the average sentence length and the percentage of complex words (3 or more syllables).
* A higher score indicates more complex and harder-to-read text.

 **smog\_index (SMOG Index)-** measures readability based on the number of complex words (three or more syllables) in a text. It estimates the years of education needed to understand the content, with higher values indicating more complex language

**TTR (Type-Token Ratio)**:

* Measures vocabulary diversity in a text.
* It is calculated by dividing the number of unique words (types) by the total number of words (tokens).
* A higher TTR indicates greater vocabulary diversity, while a lower value suggests more repetition of words.

**Hapax\_Ratio (Hapax Legomena Ratio**)

* Refers to the proportion of words that appear only once in a text.
* It is calculated by dividing the number of hapax legomena (words used only once) by the total number of words (tokens).
* A higher hapax ratio indicates more unique, less repetitive vocabulary, and can suggest a more varied or complex text.