Junaid Asif
BSAI-144

Include a report explaining your implementation, challenges faced, and observations.

Implementation:

The implementation demonstrates a series of basic text-processing tasks typically used in Natural Language Processing (NLP). The provided text (referred to as `pg`) is processed using several NLP techniques, including:

1. **Tokenization**: Splitting the text into individual words (tokens).
2. **Removing Punctuation**: Eliminating punctuation from the tokenized words.
3. **Cleaning the Text**: Removing non-alphabetic characters.
4. **Profanity Check**: Checking the text for profanity.
5. **Stemming**: Reducing words to their root form.

Challenges Faced**:**

Challenge Faced in Profanity Check

Initially, I faced a challenge with the "**`profanity-check`**" library for profanity detection. While I was testing the library, I found that it didn't perform as expected due to its reliance on older packages and models. This led to inaccurate results in some cases, and it was clear that this library wasn't well-maintained or suited for my needs.

Solution: Finding an Alternative

After researching the issue, I discovered that the "**`profanity-check`**" library was using outdated packages, which contributed to its limitations. To address this, I decided to switch to a more reliable and actively maintained library, "**`profanity`**", which provides accurate profanity detection by using a simple, yet effective approach.

Observations:

- **Tokenization and punctuation removal** are effective for preparing raw text for further processing.
- The **text-cleaning** step to remove non-alphabetic characters works well for filtering out numbers and special characters, making the text more uniform.
- The **profanity filter** method is useful for checking text but may need improvements for more complex or context-specific applications.
- **Stemming** demonstrates the power of reducing words to their root form, but it's important to note that this can sometimes lead to words that are difficult to understand or incorrect in the context.