Data Inspection Findings

1. Tokenization, Stop Words Removal, Stemming, and Lemmatization

Tokenization breaks the text into individual words or tokens, which is essential for nearly all text processing tasks.

Text: "I'm going to give my 2 months notice to my employer today."

Tokenization Result: This process would break the text down into individual words or tokens. "I'm", "going", "to", "give", "my", "2", "months", "notice", "to", "my", "employer", "today."

Stop Words Removal eliminates commonly used words that generally do not carry significant meaning, focusing the analysis on more meaningful content.

Text: "I'm going to give my 2 months notice to my employer today."

Resulting Text: I'm going give 2 months notice employer today.

Stemming and Lemmatization reduce words to their base or root form, aiding in the consistency and comparability of textual data.

Resulting Text: i go give 2 month notic employ today

2. Handling NaN Values

Text data often contains missing (NaN) values. Processing such entries without prior handling can lead to errors. Removing or replacing these with a blank string ensures the preprocessing pipeline runs smoothly.

Example: In the dataset, the row with the index of 3, titled "How's everyone feeling today?", is an instance where the selftext column is missing or null (NaN). This has been replaced with a blank string.

3. Removing URLs

URLs in text data often do not contribute meaningful information for text analysis tasks and can skew the analysis.

Example: A post contains a link within the text: "betterhelp.me\n\nThanks to the people we have already talk to us, soo manny good words and motivation to keep on.\n\nWe want to cheer up the world :D". The URL https://betterhelp.me/ would be removed, leaving the meaningful content intact.

4. Removing Special Sequences

Special sequences like are HTML or encoding artifacts that do not contribute to the meaning of the text and can interfere with text processing and analysis. Removing these sequences cleans the text for more accurate analysis.

Example: A text excerpt containing "It's happened so much to me in the past that it's just the standard I have now for relationships. I never feel safe. I'm always wondering when they're going to reject me, and like clockwork it always happens eventually. I've never been in a relationship for more than a month.\n\n\n\nEvery time, I think about how crushed and depressed I will be after the rejection, making me depressed during the relationship as well, just thinking about how it will be. And then it happens. Every. Time. \n\n\n\nI just want the pain to end. The only cure is death." Here, would be removed to clean the text.

5. Removing or Replacing Unusual Characters

This step aims to remove characters that might introduce noise into the dataset, such as symbols or emojis, which are not relevant to many text analysis tasks.

Example: This “ ## Physical? In 2014 ….” becomes “physic emoji in 2014..”

6. Lowercasing:

This step converts all letters in the text to lowercase.

Example: " I do everything very fast, I don't know what. I try to slow down sometimes but I just can't.” becomes " i do everything very fast, i don't know what. i try to slow down sometimes but i just can't."

7. Handling Short Texts:

Texts shorter than a certain length (e.g., 5 words) might not provide enough context or information for meaningful analysis. Removing or flagging such texts can help focus the analysis on content-rich entries.

Example: “Anyone else experience this” gets replaced by “Insufficient context”