**Text Analysis Methodology**

**Objective:**

The goal of this document is to explain the advanced methodology utilized to perform a comprehensive analysis of text data. This includes deriving sentiment analysis, readability metrics, identifying passive constructions, counting personal pronouns, and other key text features. This approach is designed for robust processing of textual data to provide actionable insights.

**Table of Contents**

**1) Sentiment Analysis – Extracting Emotion from Text**

1.1 Data Cleaning using Stop Words

1.2 Building a Positive and Negative Lexicon

1.3 Sentiment Score Calculation

**2) Readability Analysis – Assessing Text Complexity**

2.1 Average Sentence Length

2.2 Percentage of Complex Words

2.3 Fog Index

**3) Complex Word Identification**

**4) Word Count – Counting Relevant Words**

**5) Syllable Counting per Word**

**6) Personal Pronouns Usage**

**7) Average Word Length Calculation**

**1. Sentiment Analysis**

Sentiment analysis gauges the emotional tone behind a series of words, determining whether the text leans positive, negative, or neutral. This method is tailored for analyzing text in a financial context but can be adapted to various fields.

**1.1 Data Cleaning Using Stop Words**

Before performing sentiment analysis, the text undergoes a cleaning process where common words that do not contribute to the meaning (e.g., "the," "a," "in") are removed. This is achieved using a pre-defined stop word list, which ensures that the analysis focuses on meaningful terms.

**1.2 Building a Positive and Negative Lexicon**

A master dictionary containing words categorized as positive or negative is used. Only words not found in the stop word list are included in the analysis to ensure that common, irrelevant words do not skew results.

**1.3 Sentiment Score Calculation**

The sentiment analysis is broken down into the following key metrics:

* **Positive Score:** A score of +1 is assigned to each positive word found in the lexicon. These are summed to give the overall positive score.
* **Negative Score:** Similarly, a score of -1 is assigned to negative words. We multiply by -1 for positive output.
* **Polarity Score**: This score determines whether the overall text sentiment is more positive or negative. It is calculated using the formula:

**Polarity Score**: This score determines whether the overall text sentiment is more positive or negative. It is calculated using the formula:

**Polarity Score =** (Positive Score − Negative Score) / (Positive Score + Negative Score) + 0.000001​

The range is from -1 (most negative) to +1 (most positive).

**Subjectivity Score:** This reflects the degree of personal opinion versus factual information in the text. The formula is:

**Subjectivity Score =** (Positive Score + Negative Score) / (Total Words after Cleaning) + 0.000001

​ This score ranges from 0 (objective) to 1 (subjective).

**2. Readability Analysis**

The readability of a text is essential for understanding how accessible the content is to the target audience. We employ several metrics to gauge text complexity.

**2.1 Average Sentence Length**

This metric is derived by dividing the total number of words by the total number of sentences. It gives an idea of sentence complexity.

**2.2 Percentage of Complex Words**

Complex words are those with more than two syllables. This metric is calculated as:

**Percentage of Complex Words =** Number of Complex Words / Total Words × 100

**2.3 Fog Index**

The Gunning Fog Index is a popular readability formula that estimates the years of formal education required to understand the text. It is calculated as:

**Fog Index =** 0.4 ×( Average Sentence Length + Percentage of Complex Words)

**3. Complex Word Identification**

Complex words are those that contain more than two syllables. The identification of such words is important for assessing the overall complexity of the text.

**4. Word Count**

Word count is a fundamental metric in text analysis. The process includes:

**Cleaning the text:** Removing stop words and punctuations.

**Counting:** After cleaning, we count all the words that remain to assess the article's length.

**5. Syllable Counting per Word**

Each word is analyzed for its syllable count. We count vowels in each word and account for exceptions such as words ending in “es” or “ed,” which are not considered as additional syllables.

**6. Personal Pronouns Usage**

Personal pronouns such as "I," "we," "my," and "us" are identified using regular expressions (regex). This analysis helps in understanding the level of personal involvement or subjectivity in the text.

**7. Average Word Length Calculation**

This metric is derived by calculating the total number of characters in each word and dividing by the total number of words. A higher average word length suggests more advanced vocabulary usage.

**Conclusion**

This methodology employs a combination of text cleaning, tokenization, and various natural language processing (NLP) techniques to extract meaningful insights from textual data. By performing sentiment analysis, readability assessments, and identifying key linguistic features, we provide a detailed and actionable overview of the content. This analysis can help businesses, writers, and researchers understand the tone, complexity, and impact of their written material.