**Text Analysis**

The objective of this document is to explain the methodology adopted to perform text analysis to drive sentimental opinion, sentiment scores, readability, passive words, personal pronouns, etc.

Table of Contents

[**1**](#_heading=h.35nkun2) **Sentimental Analysis** 2

[**1.1**](#_heading=h.1ksv4uv) **Cleaning using Stop Words Lists** 2

[**1.2**](#_heading=h.44sinio) **Creating a dictionary of Positive and Negative words** 2

[**1.3**](#_heading=h.2jxsxqh) **Extracting Derived variables** 2

[**2**](#_heading=h.z337ya) **Analysis of Readability** 3

[**3**](#_heading=h.3j2qqm3) **Average Number of Words Per Sentence** 3

[**4**](#_heading=h.1y810tw) **Complex Word Count** 3

[**5**](#_heading=h.4i7ojhp) **Word Count** 3

# **Sentimental Analysis**

Sentimental analysis is the process of determining whether a piece of writing is positive, negative, or neutral. The below Algorithm is designed for use in Financial Texts. It consists of steps:

## **Cleaning using Stop Words Lists**

The Stop Words Lists (found in the folder StopWords) are used to clean the text so that Sentiment Analysis can be performed by excluding the words found in Stop Words List.

## **Creating a dictionary of Positive and Negative words**

The Master Dictionary (found in the folder MasterDictionary) is used for creating a dictionary of Positive and Negative words. We add only those words in the dictionary if they are not found in the Stop Words Lists.

## **Extracting Derived variables**

We convert the text into a list of tokens using the nltk tokenise module and use these tokens to calculate the 4 variables described below:

**Positive Score**: This score is calculated by assigning the value of +1 for each word if found in the Positive Dictionary and then adding up all the values.

**Negative Score**: This score is calculated by assigning the value of -1 for each word if found in the Negative Dictionary and then adding up all the values. We multiply the score with -1 to be a positive number.

**Polarity Score**: This is the score that determines if a given text is positive or negative in nature. It is calculated by using the formula:

Polarity Score = (Positive Score – Negative Score)/ ((Positive Score + Negative Score) + 0.000001)

The range is from -1 to +1

**Subjectivity Score**: This is the score determining whether a given text is objective or subjective. It is calculated by using the formula:

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

The range is from 0 to +1

# **Analysis of Readability**

Analysis of Readability is calculated using the Gunning Fox index formula described below.

**Average Sentence Length** = the number of words / the number of sentences

**Percentage of Complex words** = the number of complex words / the number of words

**Fog Index** = 0.4 \* (Average Sentence Length + Percentage of Complex words)

# **Average Number of Words Per Sentence**

The formula for calculating is:

**Average Number of Words Per Sentence =** the total number of words / the total number of sentences

# **Complex Word Count**

Complex words are words in the text that contain more than two syllables.

# **Word Count**

We count the total **cleaned** words present in the text by

1. removing the stop words (using the stopwords class of the nltk package).
2. removing any punctuations like ? ! , . from the word before counting.