

Text Analysis

SEC 10-K and its Amendment filings

Objective of this document is to explain methodology adopted to perform text analysis to drive sentimental opinion, sentiment scores, readability, passive words, personal pronouns and etc. from SEC filings and reports.

Table of Contents

1	Sentimental Analysis	2
1.1	Cleaning using Stop Words Lists	2
1.2	Creating dictionary of Positive and Negative words	2
1.3	Extracting Derived variables	2
1.4	Sentiment score categorization	3
2	Analysis of Readability	3
3	Average Number of Words Per Sentence	3
4	Complex Word Count	4
5	Word Count	4
6	Syllable Count Per Word	4
7	Personal Pronouns	4
8	Passive Words	4
9	Average Word Length	5

1 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 on Financial Texts. It consists of steps:

1.1 Cleaning using Stop Words Lists

The Stop Words Lists (found [here](#)) are used to clean the text so that Sentiment Analysis can be performed by excluding the words found in Stop Words List. Use this url if above does not work <https://sraf.nd.edu/textual-analysis/resources/>

1.2 Creating dictionary of Positive and Negative words

The Master Dictionary (found [here](#)) 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. Use this url if above does not work <https://sraf.nd.edu/textual-analysis/resources/>

1.3 Extracting Derived variables

We convert the text into a list of tokens using the nltk tokenize 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 so that the score is 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:

$$\text{Polarity Score} = (\text{Positive Score} - \text{Negative Score}) / ((\text{Positive Score} + \text{Negative Score}) + 0.000001)$$

Range is from -1 to +1

Subjectivity Score: This is the score that determines if 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)

Range is from 0 to +1

1.4 Sentiment score categorization

This is determined by grouping the Polarity score values in the following groups.

Most Negative: Polarity Score below -0.5

Negative: Polarity Score between -0.5 and 0

Neutral: Polarity Score equal to 0

Positive: Polarity Score between 0 and 0.5

Very Positive: Polarity Score above 0.5

2 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 = $0.4 * (\text{Average Sentence Length} + \text{Percentage of Complex words})$

3 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

4 Complex Word Count

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

5 Word Count

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

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

6 Syllable Count Per Word

We count the number of Syllables in each word of the text by counting the vowels present in each word. We also handle some exceptions like words ending with "es", "ed" by not counting them as a syllable.

7 Personal Pronouns

To calculate Personal Pronouns mentioned in the text, we use regex to find the counts of the words - "I," "we," "my," "ours," and "us". Special care is taken so that the country name US is not included in the list.

8 Passive Words

Passive Words are the Auxiliary verbs followed by a word ending in "ed" or one of 200 irregular verbs present in the Past form column in this [link](#).

Auxiliary verbs are - auxiliary verb variants of "to be" including: "to be", "to have", "will be", "has been", "have been", "had been", "will have been", "being", "am", "are", "is", "was", and "were".

9 Average Word Length

Average Word Length is calculated by the formula:

Sum of the total number of characters in each word/Total number of words