Textual Features

The textual features consist of the distinctive formal (particularly stylistic and structural) aspects of an utterance, text, or artwork in any medium (Oxford Dictionary, 2019). They were investigated under six categories, which four of six were adapted from (Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014) and two were introduced for the first time in this study: Text Structure, Text Style, Text Length, Text Sentence Type, Level of Readability and Lexical Variety. All categories are described as follows:

- ***Text Structure*** features are to describe the structure of how well a text is organized. The HTML tags were counted to obtain the number of structural features:
  - Section count
    - Counting tags <h1> to <h6>
  - Paragraph count
    - Counting tag <p>,
  - Bullet count
    - Counting tag <li>
  - Length of shortest section
    - Minimum length of total word count in sections
  - Number of images
    - Counting tag <img>
  - Average section length
    - Length of total word count in sections / Number of sections
  - Average paragraph length
    - Length of total word count in paragraphs / Number of paragraphs

- ***Text Style*** features are used to capture the way authors write their articles through their word usage; they try to capture how the text is written, how author used the language and whether the text has some specific characteristics. Also, professional health literacy guidelines encourage the use of active voice instead of passive voice (Best Practice Guidance on Patient Information Leaflets, 2014; Toolkit for producing patient information, 2003; U.S. Department of Health and Human Services, 2015; MedlinePlus, 2018). Text style features aim to capture such specific usages.
  - Number of uses of verb to be
    - Counting “am, is, are, was, were, being, been, and be” in sentences
  - Number of passive voice sentences
    - Counting ‘nsubjpass’ parser from Stanford dependency parser for English syntax
  - Number of nouns
    - Counting “NN, NNP, NNPS, NNS” POS tags from Stanford POS Tagger toolkit
  - Number of auxiliary verbs
    - Counting “MD” POS tags from Stanford POS Tagger toolkit
  - Number of verbs
    - Counting “VB, VBD, VBG, VBN, VBP, VBZ” POS tags from Stanford POS Tagger toolkit
  - Number of adjectives
    - Counting “JJ, JJR, JJS” POS tags from Stanford POS Tagger toolkit

    - Number of sentences starting with a pronoun
        - Counting if first tag of sentence' is "PRP, PRP$" from POS tags from Stanford POS Tagger toolkit
    - Short sentence rate
        - Calculate the percentage of sentences whose length is less than 15 words

- ***Text Length*** features used by (Rassbach, Pincock, & Mingus, 2007; Blumenstock, 2008; Dalip, Lima, Cristo, & Calado, 2014) such as word count, sentence count, and character count are related to the size of the text in different aspects. These features are useful in assessing whether the information is complete and comprehensive.

    - Character count
        - Total character length of each word
    - Word count
        - Counting by using the method word_tokenize() from nltk to split a sentence into words
    - Sentence count
        - Counting by using the method sent_tokenize() from nltk to split text into sentences

- ***The Level of Readability*** features are used as a way to verify if the text is well written, understandable, and free of unnecessary complexity. Flesch Kincaid Grade Level (FKGL) and Simple Measure of Gobbledygook (SMOG) indices are frequently used. In the calculation of SMOG, at least 30 sentences in a row near the beginning, in the middle and in the end are selected from text. The motivation behind this is health websites are expected to provide understandable health information regardless of age, background or reading level (Toolkit for producing patient information, 2003; Best Practice Guidance on Patient Information Leaflets, 2014; MedlinePlus Trusted Health Information for You, 2018).

$$FKGL = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59 \qquad \text{(Eq. 3)}$$

$$SMOG = 1.0430\sqrt{number\ of\ polysyllables * \frac{30}{number\ of\ sentences}} + 3.1291 \quad \text{(Eq. 4)}$$

Apart from adopted features, the following feature set was developed:

- ***Text Sentence Type*** features are constructed to analyse the types of sentences. Unlike other studies, sentence types were used as a feature in this study. The motivation behind this is health websites generally use imperative and declarative sentences to guide patients. Five different types of sentences in English were considered: Imperative, Interrogative, Exclamatory, Existential and Declarative. Table 1 shows an example for each sentence type.

Table 1: Examples of sentence types

Sentence types	**Examples**
Imperative sentences	"*Monitor* your blood glucose every three to four hours."
Interrogative sentences	"How do insulin pumps work*?*"
Exclamatory sentences	"Initially take it slow you don't want to start off too hard, if you are not used to the exercise you will be sore the next day and this will not make exercising a fun experience*!*"
Existential sentences	"*There* are different types of insulin depending on how quickly they work, when they peak, and how long they last."
Declarative sentences	"Inside the pancreas, beta cells make the hormone insulin."

- If the first word of the sentence is a verb ("VB, VBD, VBG, VBN, VBP, VBZ" POS tags), then the sentence is labelled as imperative.
- The sentence is marked as an interrogative sentence when it has a question mark (?).
- If it has an exclamation mark (!), it is regarded as an exclamatory sentence.
- An existential sentence is a sentence that asserts the existence or nonexistence of something; if a sentence starts with existential "there" ("DT, EX" POS tags), it is considered as an existential sentence.
- The rest of the sentences apart from these in the text are labelled as declarative.

- ***Lexical variety*** is the normalized measure of the unique words used in a text with all words. The motivation behind this is if a website is comprehensive, the lexical variety is expected to be high.

$$Lexical\ variety = \frac{numberof\ unique\ words}{number\ of\ words}$$