**REVIEW OF RELATED LITERATURE**

**Natural Language Processing**

Natural Language Processing (NLP) is a field of study which focuses on discovering ways on how to bridge the gap between interactions involving humans and computers. It aims to provide a method for computers to analyze and comprehend natural languages (a.k.a. human languages) in an intelligent way, or by means of simulating the process of "understanding" - either through Symbolic approach, which utilizes a set of predefined rules, modelling a different language phenomenon, or Statistical approach, which makes use of machine learning algorithms to learn the language phenomena. Concepts in computer science, artificial intelligence (AI) and computational linguistics are what comprises NLP. After all, Natural Language Processing is said to be the main component of AI and that it relies on machine learning as well - in order to enable the system to derive patterns in a given dataset which would help improve its own understanding of speech. It differs from common word processor operations in such a way that NLP possesses the capability to analyze the word for its meaning rather than only for its structure (viewing the word in a symbolic approach).

A great number of current software applications have been incorporated with NLP tasks in order for them to function appropriately. Some of those tasks are as follows:

* Deep Analytics - the extraction of information from large or multi-source datasets with the help of advanced data processing techniques. It is most useful in dealing with precisely targeted or highly complex queries to be done in unstructured or semi-structured data.
* Machine Translation - translating one human language into another.
* Named Entity Extraction - a data mining process which is concerned with the identification of one item in a set from other items having similar attributes based on a named entity definition - a word or phrase that can clearly distinguish an object from the rest.
* Co-reference Resolution - determining which words refer to the same object.
* Automatic Summarization - creating a summarized version of a large chunk of text.
* Sentiment Analysis - identifying subjective information (e.g. judgement, opinion, emotional state, etc.) in text.
* Text Classification - assigning predefined categories to a text.
* Conversational Agents - systems having given the capability to coherently converse with humans.

A system’s skill that could count as an example of a Natural Language Processing capability would be developing a decent conversation in pure human language. Likewise, computer systems that can convert human languages to computer languages and vice-versa are currently existent. Translation programs were also made possible by NLP. Additionally, grammar and spelling checkers were also programmed following the mechanism of implementing text processing techniques under Natural Language Processing. Lastly, a computer that can read human languages (in publications such as books) is also a product of NLP.

Despite the current capability of NLP in terms of Human-Computer Interaction, it still remained limited particularly in producing statements which involved 100% human reasoning and logic. NLP can only refer to a stricter subset of the human languages which means that it cannot allow anomalies which are often occurring in a particular human language.

**Textual Corpus**

Or simply **corpus**, is a collection of large sets of text specifically chosen by the researcher (or linguist, in terms of linguistics) to deduct his/her own linguistic analysis – may either be statistical analysis and hypothesis testing, checking occurrences or validating specific rules - in the text of interest. In the field of Natural Language Processing, the analysis conducted can be used as a basis or test bed for constructing NLP systems. They are typically processed and stored electronically (with the use of databases, etc.) in today’s time.

**Corpus Annotation**

The method of adding linguistic information to the data included in the corpus is known as **corpus annotation**. The most common type of annotation being done to the dataset is by putting labels or tags which indicates the **class** to which a particular word in the body of text belongs to.

**Text Pre-processing Tools**

The performance of the terminology extraction system highly depends on the quality of the linguistic preprocessing. Therefore, a high-performance pre-processing toolkit is crucial in achieving the pre-processing steps.

**LeTs Preprocess Toolkit** is a suite of robust pre-processing modules. It includes Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers for four different languages: Dutch, English, French and German. By using this toolkit, each token will be assigned as a class tag, PoS, named entity, or a lemma tag. It also supports both monolingual (single language) and parallel corpus (two languages). As an initial step, the data in the corpus is separately pre-processed, by splitting it into sentences, tokenized, Part-of-Speech tagged, lemmatized, and finally named entities are extracted.

**The Natural Language Toolkit (NLTK)** is a platform used for creating Python programs to work with human language data in the application of natural language processing (NLP). It contains a suite of text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. The NLTK’s basic functionalities include: sentence tokenization, word tokenization, wordnet, synsets, and synonyms, stemming words and lemmas

**Methods in Pre-processing Text**

**Chunking**. It is used for recovering phrases constructed by the part-of-speech tags, finding noun phrases, finding verb groups, and others. It is commonly used for information extraction, keywords extraction, entity recognition, and even relation extraction. Like tokenization, it also omits whitespaces, however, chunking usually selects a subset of the tokens.

**Lemmatization.** It is the process of grouping together the different inflected forms of a word in order to determine the lemma in the given context. Lemma is defined as the canonical form, dictionary form, or citation form of a set of [words](https://en.wikipedia.org/wiki/Word). The lemmatization process involves understanding the context and determining the [part of speech](https://en.wikipedia.org/wiki/Part_of_speech) of a word in a sentence.

**Stemming**. It involves the process of reducing inflected words to their [word stem](https://en.wikipedia.org/wiki/Word_stem), base or [root](https://en.wikipedia.org/wiki/Root_(linguistics)) form by removing its affixes. The stem does not need to be identical to the [morphological root](https://en.wikipedia.org/wiki/Morphological_root) of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

**Part-of-Speech Tagging.** It is the process of classifying words into their parts of speech and labeling them accordingly. Parts of speech are also known as lexical categories. Moreover, the collection of tags used for this process is known as a tagset.

**Tokenization.** Naturally, text needs to be segmented into linguistic units such as words, punctuation, numbers, and alpha-numeric characters before text processing is to be done. This process is called tokenization. It is the process of breaking a set of text into words, phrases, symbols or tokens. At the same time, it omits certain characters such as punctuations and white spaces. Tokens are used as inputs in parsing or text mining.

**Related Studies**

Van Hee et al. (2015) conducted a research on Automatic Detection and Prevention of Cyberbullying. The team presented the construction and annotation of a corpus of Dutch social media posts annotated with fine-grained text categories, such as insults, threats, sexual talk, defamation, defense, and curse. The participants in a cyberbullying context were also identified in order to enhance the analysis of human interactions involving cyberbullying. Initially, the researchers had decided to use this particular research paper as their main basis for creating the project. However, the process of manually annotating the statements within the dataset, according to the aforementioned fine-grained text categories, proved to be difficult as some of the categories were closely related to each other. Additionally, the succeeding methods after the data annotation process proved to be difficult to comprehend given the current knowledge the researchers possess under the NLP field.

Dinakar et al. conducted a research at Massachusetts Institute of Technology. The research was aimed towards detecting cyberbullying content in Youtube comments. The first level of classification involves determining if the comment is in a range of sensitive cyberbullying topics such as sexuality, race, intelligence and physical attributes. The second level is to determine what topic. The experiment achieved 66.7% accuracy for detecting instances of cyberbullying. This project also used a support vector machine learner. The researchers based their pre-defined classification of cyberbullying types on the sensitive topics enumerated above. Hinduja and Patchin included harmfulness as one of the constituents of cyberbullying. Similarly, in terms of sensitive topics, there will always be that particular group of people which would have been offended by it (which is why it was deemed controversial). Therefore, the researchers thought that by recalling well-known controversial issues in the Philippines, they will be able to determine the borderline between cyberbullying and non-cyberbullying statements, making it easier to tell them apart and classify them.

In 2016, Cheng and Ng conducted a research at De La Salle University. The research aimed towards detecting cyberbullying roles through textual context in Facebook and Twitter. First, the researchers identified six roles in a cyberbullying context: the bully, victim, assistants of the bully, reinforcers, outsiders, and defenders. Among the three algorithms used by the researchers such as Naïve Bayes classifiers, decision trees and Support Vector Machine (SVM), the SVM had the highest accuracy. The optimal model produced an accuracy of 59.7% in detecting the bullying roles; while detecting the bully role produced an accuracy of 80.9%. The researchers are currently using this study as their basis in the creation of their proposed cyberbullying detection model because unlike the other study (as mentioned before), they found this paper easier to comprehend. It gave them a clear picture of what they should do in order to achieve their desired output. Additionally, since SVM has been proven to be the most accurate model, the researchers were also planning to use SVM in automating the detection of cyberbullying occurrences.