OMDENA SCHOOL: HATE_SPEECH_DETECTION

PROFESSOR: DR. MOHAMMED ZUHAIR AI TAIE

NAME: ARMIELYN C. OBINGUAR

DATE ASSIGNED: November 21,2021

DATE SUBMITTED: November 25,2021

1. **How is unstructured data different from structued and semi-structured?**

It is suggested that the unstructured data is not arranged in a way that anybody who would want to do anything with it would smoothly and simply utilize. Although we also know that some unstructured texts may be available on the platforms such as the text papers in word, PDF and even in any media logs that have no format nor any specified instructions. On the other hand, semi-structured is described as the Semi-structured data is information that does not live in a relational database but that has certain organizational qualities that make it simpler to examine. With certain procedures, you can store them in the related database (it may be quite hard for some sort of semi-structured data), but Semi-structured exist to ease

2. **Name one or more NLP libraries that support your local language.**

Considering that English is officially recognized as the primary language of communication in our nation. In order to do this, we make use of the following NLP libraries:

- **spaCy**
- **polyglot**
- **scikit−learn**
- **Pattern**

3. **What are the main issues within the binary encoding and BoW methods for feature extraction from text?**

A. **By taking into consideration all potential consecutive word pairings, there might be a large number of possible bigrams. Furthermore, if the vocabulary is enormous, employing N-grams might result in a massive sparse(containing a significant number of 0's) matrix, which makes the calculation very difficult.**
B. **If the prospect of words increasing and continually becoming bigger exists, one of the conceivable consequences would be that the vector would most likely rise as well, as previously stated.**