Université Chouaib Doukkali
Faculté des Sciences
Département d'Informatique
EL Jadida

AU : 2022/2023

Programmation Avancée : Python
Devoir à rendre avant le 14/05/2023

Cyberbullying refers to the use of digital technologies, such as social media, text messaging, and online forums, to deliberately harass, intimidate, or humiliate someone. Cyberbullying can take many forms, including sending threatening or insulting messages, sharing embarrassing photos or videos, spreading rumors, and impersonating someone online.

Cyberbullying can have serious consequences for the victim, including emotional distress, depression, anxiety, and in extreme cases, suicide. It can also affect the victim's relationships with others, their academic and professional opportunities, and their overall quality of life.

It's important to recognize the signs of cyberbullying and take steps to prevent it. This can include reporting abusive behavior to the appropriate authorities, blocking or unfriending the perpetrator, and seeking support from friends, family, or a mental health professional.

Cyberbullying is a serious issue and using machine learning can be a powerful tool to help detect and prevent it.

In what follows are some steps to follow to detect cyberbullying messages from others using some machine learning algorithms.

1. **Data collection**: The first step in any machine learning project is to collect data. For our project on cyberbullying, you will use a dataset we got from Kaggle.
2. **Data cleaning**: The process of data cleaning in text can include tasks such as:

    a. Removing punctuation marks, stopwords, and special characters
    b. Correcting spelling errors and typos
    c. Standardizing text by converting all characters to lowercase or uppercase
    d. Removing irrelevant information such as URLs, HTML tags
    e. Stemming or lemmatization

3. **Feature engineering**: Once you have collected your data and clean, you will need to extract features from it that can be used as inputs to your machine learning model. Some common features for text-based data include term frequency, TfIdf, n-grams, and embedding.
4. **Model selection**: There are many machine learning algorithms that can be used for cyberbullying detection, including Naive Bayes, Support Vector Machines (SVM), Random Forests, and Neural Networks.
5. **Evaluation**: To evaluate the performance of your machine learning model, you will need to use appropriate evaluation metrics. For a binary classification task (e.g., detecting whether a text contains cyberbullying or not), you can use metrics such as accuracy, precision, recall, and F1-score.

To do:

- Form groups of 2, 3 or 4 students
- Each group must choose one of these 3 algorithms: SVM, NB and Random Forrest. At least 2 groups per algorithm
- Follow the steps above
- Provide a report containing the theoretical part and the source code

I will send you the dataset