Submit a pdf document that should contain the following details:

Team member's details :

Group Name (give a name to your group): Data Junkies

Name: Farheen Fatima
Email: Farheen.fatima07@gmail.com
Country: USA
College/Company: California State University, Sacramento.
Specialization (Data Science, NLP, Data Analyst): Data Science

Name: Mohamed Derbeli
Email: derbelimohamed1@gmail.com
Country: Spain
College/Company: Data Glacier
Specialization: Data Science

Name : Mohamed Akrem Ben Jemia
Email: akrem008.benjamie@gmail.com
Country: Spain
College/Company : École National d'ingénieur de Carthage
Specialization: NLP

Problem description: The term hate speech is understood as any type of verbal, written, or behavioral communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are. In other words, based on their religion, ethnicity, nationality, race, color, ancestry, sex, or another identity factor.

Data understanding: Hate Speech detection is generally a task of sentiment classification. So, for training, a model that can classify hat speech from a certain piece of text can be achieved by training it on data that is generally used to classify sentiments. So, for the task of the hate speech detection model, we will use Twitter tweets to identify tweets containing Hate Speech.

What type of data you have got for analysis:
The dataset contains 3 features. Id, Label, and the tweets.

What are the problems in the data (number of NA values, outliers, skewed, etc):
This dataset doesn't have any null values but it contains stop words and symbols.

What approaches you are trying to apply to your data set to overcome problems like NA value, outlier, etc and why?
To remove the stop words, we have used nltk's stopwords library.
And we have used Regular expressions to remove the symbols.

GitHub Repo link : s