TITLE: Twitter sentiment analysis

Problem Statement:

The goal of this project is to identify the sentiment of an entity in a tweet. The project involves analysing a set of tweets related to a particular entity and classifying the sentiment expressed in each tweet as positive, negative, or neutral. This can be useful in understanding the public perception of a particular entity

Data Collection:

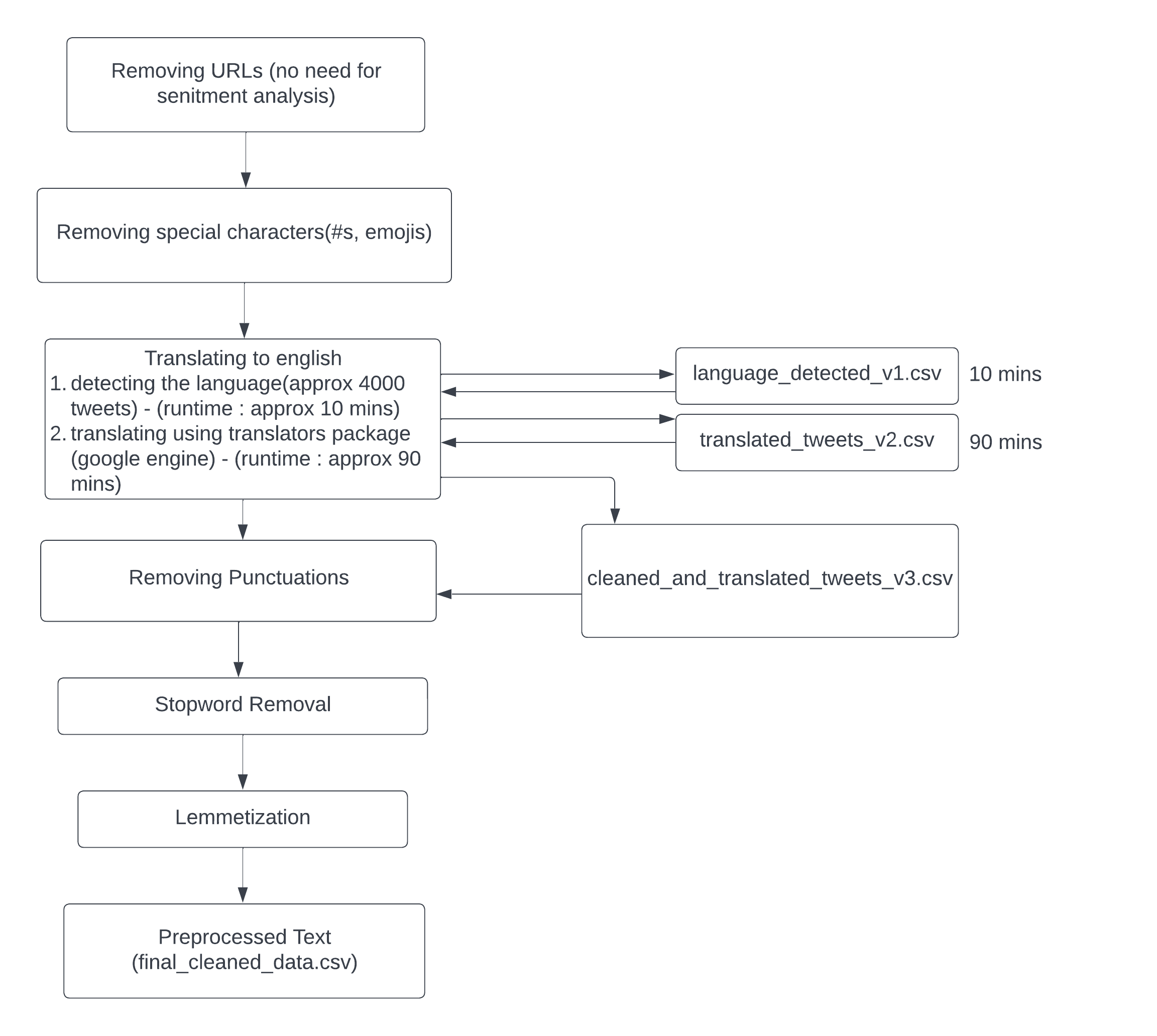
Data is taken from online repository. The data talks about the drugs and medicine, their phases of study, their efficacies, organization that does it, some other health problems as well.

Data CLEANING (Text Pre-processing):

After collecting the data, it is necessary to clean it. This involves removing any irrelevant information such as URLs, emojis, special characters. Note the URLs are first removed then special characters in the given order changing the order will make URLs detection difficult

Some tweets are in other languages( almost 10% of the entire data). In order to translate the tweets, we used langdetect module to detect other languages and translators package to translate the given tweet detected. since it takes around 90 mins for the entire translation to finish, we store the result in a translated\_tweets\_v2.csv

The next step is to pre-process the text data. This involves removing punctuation, removing stop words, lemmatize the sentences. lemmatization to reduce words to their root form as it reduces the words need to analyse haven’t converted the tweets to lower case because some entities are detected are written in abbreviations which can only be identified by its uppercase characteristics



Entity Extraction

The next step is to extract entities from the tweets. since most of the data talks about drug and scientific terms we used en\_core\_sci\_sm spacy Named entity recognition pre-trained model .It is pretrained on the scientific text and it a smaller model compared to full spacy models. Using this we extracted entity from each tweet and stored the output a list for each tweet in the entities column of dataframe.

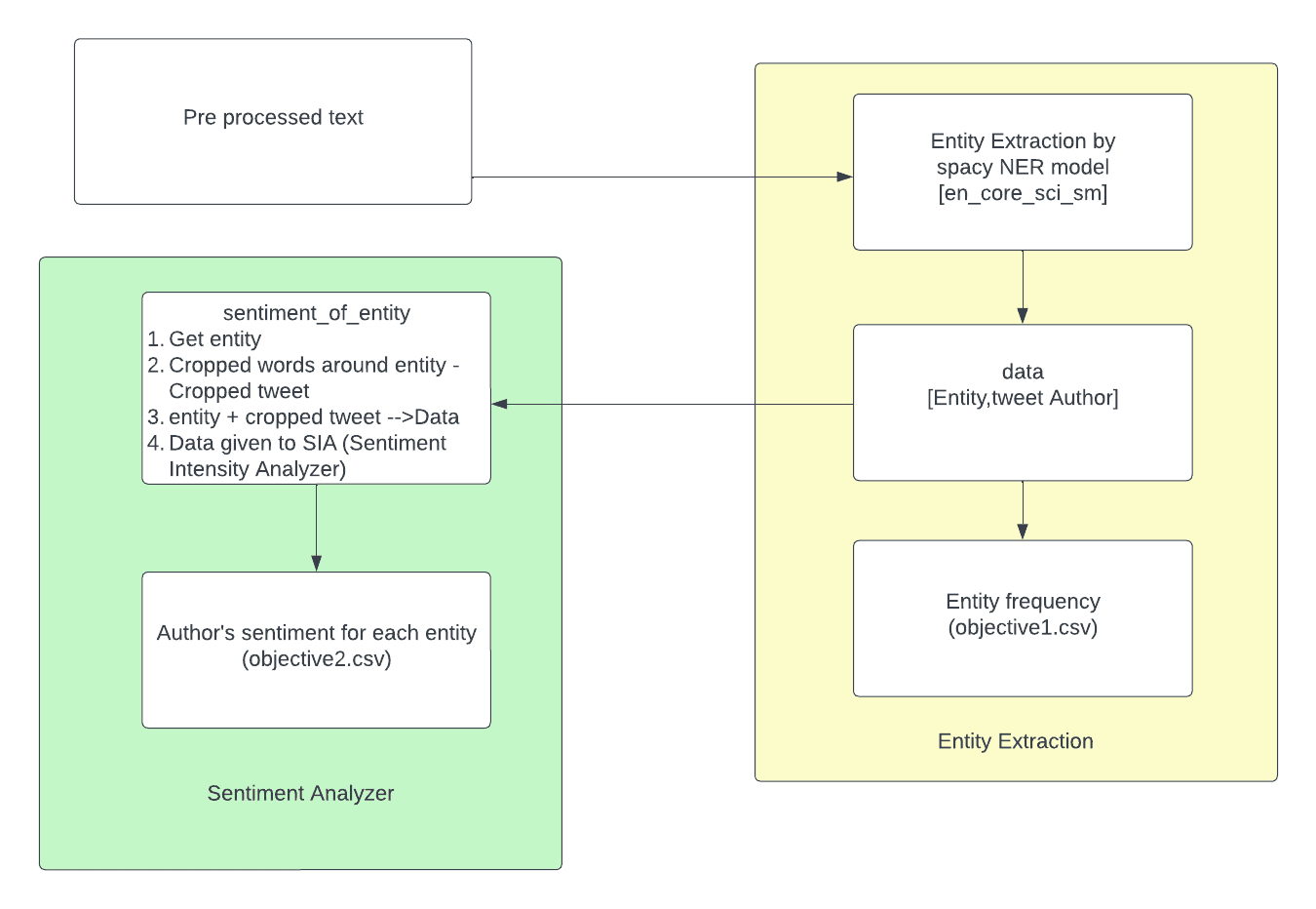
The resultant data (entities and its frequency) is stored in objective1.csv

Sentiment Analysis:

In order to extract sentiment of the entity from the given text I used pretrained sentiment analysis tool data frame from nltk library. The sentiment score generated by SentimentIntensityAnalyzer ranges from -1 (very negative) to +1 (very positive). This tool is relatively simple to use and provides a quick way to perform sentiment analysis on a given text.

Entity and the the 10 words around the entity is selected to perform the sentiment analysis. If the compound score is zero it is taken as neutral if it is positive number it is taken as positive and vice versa

The resultant data is stored in objective2.csv



Reference

<https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d>

<https://towardsdatascience.com/using-scispacy-for-named-entity-recognition-785389e7918d>

<https://www.youtube.com/watch?v=R-AG4-qZs1A&list=PLeo1K3hjS3uuvuAXhYjV2lMEShq2UYSwX>

<https://allenai.github.io/scispacy/>