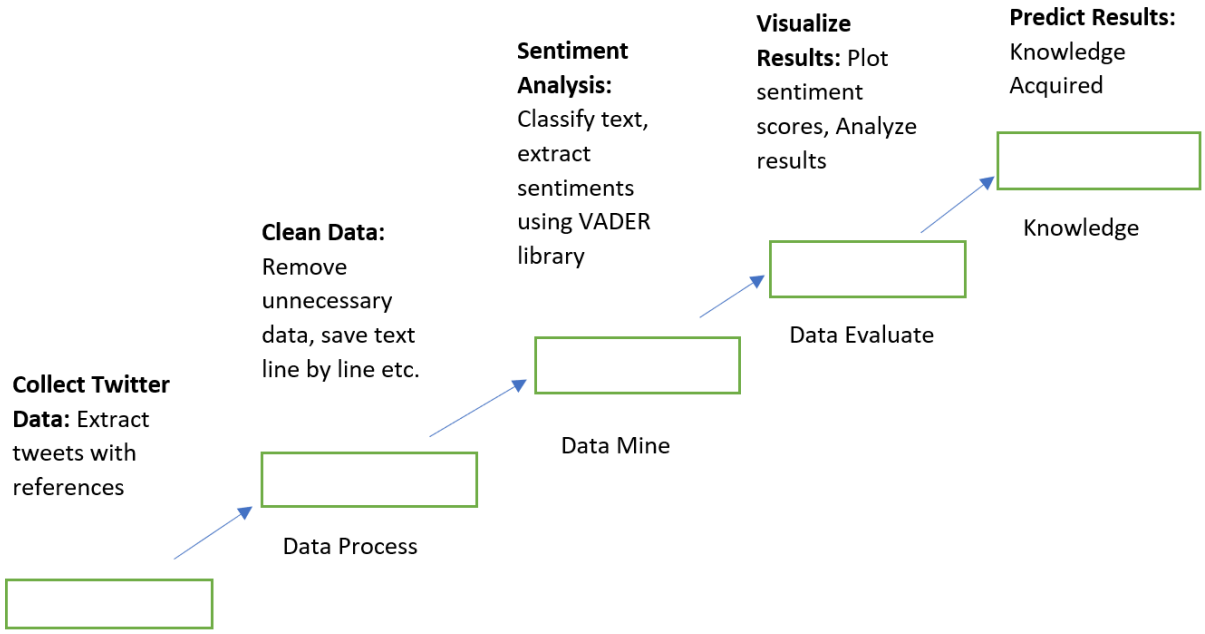
Public opinion and sentiment analysis on the COVID-19 vaccine using Twitter data

- **Team**: Chirps
- **Team Members**: Ashika Anand Babu, Subarna Chowdhury Soma, Vishnu Vardhan Reddy Yeruva

KNOWLEDGE DATA DISCOVERY



Selection:

Selection of raw data from kaggle

Sources:

1) https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets
2) https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Preprocessing:

Preprocessing the data by removing irrelevant data and dropping the null, duplicate values and clearing off the data which is not in proper format or converting the format is taken care of.

Transformation:

Transforming the data by picking unique values and extracting features and setting up dimensions to make the data efficient.

Data Mining:

Reducing the dimensions of the data and using NLP algorithms like Sentiment Analyzer

Interpretation/Evaluation:

Using XGBClassifier for evaluation and putting Subjectivity forward would like to push this forward.

Knowledge Representation:

Visualizing under Tweet Platform-wise Distribution and Hashtag, Date and Time feature analysis and representing the knowledge.

FEATURE ENGINEERING

Features:

Feature Engineering is needed to prepare the input data properly so that our algorithm to determine a favourable candidate can run properly. In the data pre processing stage of our project, we conducted the following:

1. **Covid Vaccination Progress:**
    a. Progress of covid vaccines per day, per week, month, to an year visualizing using plotly for every country
2. **Vaccines Success rate:**
    a. Sentiment analysis of predictive as well as subjective data on vaccines such that people couldn't be affected by the powerful toxins that they are allergic to.
3. **Date Time Features: these are components of the time step itself for each observation:**
    a. 'Series' refers to each time step(you will see that in the next cell)
    b. 'Target' refers to the target value at the current time step.
4. **Change emojis to either positive or negative:**
    a. This will add value to sentiment analysis appropriately

5. **Lower all the text or tweets:**
    a. This will ensure tweets with capitalization and non-capitalization are tested as the same.
6. **Delete numbers, quotes, links, user information, one characters and punctuation:**
    a. This only adds noise or is unhelpful to NLP analysis.
7. **Clean double spaces and multiple treats:**
    a. This is just to make sure non characters and consecutive repeated words aren't used for NLP analysis.
8. **Replace common words:**
    a. This is to update contraction words with their appropriate expansion which can be used in NLP analysis.