The steps we took in our project are:

- **Dataset:** We found the dataset from Kaggle.com. The size of the corpus was initially around 10019 tweets, and for the purpose of our project, we randomly chose 472 tweets in English, Russian and Hindi (since we speak the languages)
- **Data cleaning:** From the dataset, we removed numbers and punctuation, emoji, as well as stop words (except for not, cannot, nothing, because they indicate emotions or attitudes of authors). As a result, lemmas were generated
- **Annotation:** We annotated each tweet in our dataset according to Ekman's six emotions. In case a tweet expressed mixed emotions, we chose the one that, in our opinion, was the strongest. Therefore, we got 189 tweets with emotion "happy", 128 with "sad", 16 with "surprise", 47 with "fear", 27 with "disgust", and 65 with "angry".
- **Format transformation:** Initially, our dataset was in .csv format, so, we transformed it into .txt format through a separate python script
- **Document header:** Since our purpose was to use LAF/GrAF to represent our annotations, we created the *Document Header* which contains information about the primary data document and annotation document
- **Resource header:** It provides more detailed information about the resource as a whole
- **Segmentation header:** In our dataset, each tweet is considered to be one segment. Here, we have defined the anchors of each tweet.
- **Annotation document:** In this file, segments are linked to their labels.
- **PostgreSQL:** We use it to connect our data files and deploy the result values over a database. The reason why we need the PostgreSQL tool is to access the values of the XML tags, present in LAF/GrAF. We used psycopg2 and elementTree for connecting to the database, and with the Insert query, we added the values to the database Table. Finally, with the help of the pgAdmin framework interface we executed the queries (screenshots are provided).