Code can be found here:
https://github.com/lucianistrati/Machine-Translation-Romanian-Dialects
Models can be found here: https://huggingface.co/fmi-unibuc

Structure of the code file inside the repository

- src/
    - Analysis.py
        - Plotting wordclouds of the words used in every dialect;
    - load_data.py
        - Can load three types of data:
            - Transcriptions with 3 different formats (custom format of our annotations, sonix-format and vatis-tech-format);
            - Books by extracting their content from pdfs;
    - scrap_dexonline.py
        - Work-in-progress - should be used to scrap/crawl data from dexonline;
    - Train_word2vec.py
        - Trains a word2vec CBOW model over all the books in every dialect;
    - compare_anns.py
        - Script used to compare how accurate were the Speech-To-Text transcriptions between Sonix-Ai & Vatis-Tech solutions;
    - mat.py
        - Script used to obtain a similarity matrix between the overlaps of each dialect;
    - Oltenizator/Tense_changer.py
        - Used for changing the tense from passe-compose to passe-simple and vice-versa;
        - Can be called from the command line:
        - python tense_changer.py -s "A aranjat camera."
        - python tense_changer.py -s "Aranjai camera." -r
    - Translation.py
        - Has all the translation functionality from one dialect to another;
    - detect_dialect.py
        - Inference of the dialect model
    - oltenizator
        - Conjugari.json
            - Json containing conjugations of all the verbs in Romanian
        - Verbe.json
            - Json containing all the verbs in Romanian
        - Tense_changer.py
            - Script used to change the tense
    - tests
        - Test_tense_changer.py
            - Unit test code file used to test the functionality of src/oltenizator/tense_changer.py
    - Utils.py
        - Has several dictionaries that contains useful rules for translation as well as mapping of the videos and the books to dialect labels

- Train_model.py
    - Trains a model able to classify in what dialect a text is in;