# Applications (I): Project Proposals

**Experimental:** experiment in a Text Classification or Sequence Labelling (or any other) task of your choice (it can be any of the tasks included in the module or some others): (i) Identify relevant datasets for training; (ii) identify relevant related work to your chosen trask; (iii) train, evaluate, report and compare with the current state-of-the-art results (iv) extract the most common false positive and false negative prediction errors with respect to the gold standard and perform a qualitative error analysis.

**Visualization:** Use a visualization tool such as Scattertext (Kessler 2017) to perform corpus-based data analysis. Scattertext is intended for visualizing what words and phrases are more characteristic of a category than others. This may help, for example, to compare discourses by different political parties, or by different publishers about the same topic. We can also use Scattertext to visually compare the words and phrases most common in stance detection corpora. For example, by comparing the FAVOR and the AGAINST categories in the SemEval 2016 data. Alternatively, you could use other corpora such as the Hyperpartisan News Detection corpus, or the fake news and celebrity data, to characterize via Scattertext the hyperpartisan and objective categories.
More specifically:
(i) Check the tutorial
https://github.com/JasonKessler/scattertext#using-scattertext-as-a-text-analysis-library-finding-characteristic-terms-and-their-associations to learn about Scattertext;
(ii) apply the tutorial to your data (either SemEval 2016, Hyperpartisan News or fake/celebrity data);
(iii) make a qualitative analysis based on the visualization obtained.
**Data can be found here:**
https://drive.google.com/drive/folders/1eENf2b0ArnA25h2E_znBZYNU17HFpPt9?usp=sharing

**Chatterbot:** While state-of-the-art chatbot systems use Transformers such as GPT-2 (https://arxiv.org/pdf/2004.12752.pdf), for this project you can use Chatterbot (https://chatterbot.readthedocs.io/en/stable/), a very simple machine learning-based chatbot. While easy to use, Chatterbot provides only a tiny little amount of training data. Thus, it would be interesting to leverage larger corpora to train Chatterbot for a language(s) of your choice. Summarizing, this may involve the following: (i) create a virtual environment in your personal machine to install chatterbot and chatterbot-corpus; (ii) Download your **monolingual untokenized raw files** for your chosen language (https://opus.nlpl.eu/OpenSubtitles-v2018.php); (iii) Format the corpus to the YAML format used in Chatterbot and include it in the appropriate language directory (https://github.com/gunthercox/chatterbot-corpus#create-your-own-corpus-training-data); **NOTE:** formatting the data requires saving it into YAML format, using a YAML parser, otherwise Chatterbot will output many parsing errors; (iv) train the model; (v) chat with the Chatterbot performing a qualitative analysis of the dialogue, for example, by comparing the dialogues performed by the models trained with the tiny corpus available in Chatterbot or with the OpenSubtitles corpus; (vi) **OPTIONAL**: Split the training data into train and test sets in the format required by the evaluation toolkit https://github.com/ricsinaruto/dialog-eval, re-train and evaluate the obtained model on the newly created testset.

**VaxxStance@IberLEF 2021:** Participate in the VaxxStance shared task. (i) Choose one or more participation tracks (ii) Pick a system to train stance detection models, (iii) experiment with various system configurations (type of embeddings, local features, etc.), (iv) describe your final systems and final results. https://vaxxstance.github.io/

**MultiCoNER@SemEval 2022**: Participate in the https://multiconer.github.io/ shared task. (i) Choose one or more participation tracks (ii) Pick a system to train sequence labelling models, (iii) experiment with various system configurations (type of embeddings, local features, etc.), (iv) describe your final systems and final results.
**Data can be found here:**
https://drive.google.com/drive/folders/1HJvR4_3heysJsUR7Ccgr9OtpMV9N2QJd?usp=sharing