The project consists of three main modules that should be run to reproduce the results; these modules are "single_models.py", "ensemble_models.py", "total_score.py". You can find below the requirements, the description for each main module, and the instructions to regenerate the results. Moreover, the description of the files, folders, and the other modules that are called by the main modules are presented.

Requirements:

- Python 3.5.0
- numpy (1.15.0+mkl)
- scipy (0.19.1)
- scikit-learn (0.18.2)
- pandas (0.22.0)
- regex (2018.2.21)
- nltk (3.2.4)
- gensim (2.2.0)
- wordsegment (1.3.0)
- openpyxl (2.5.12)
- JPype1 (0.6.2)
- ekphrasis (0.4.10)

Notes:

- Google word vectors can be downloaded from the following link, and should be saved in the working directory containing the main modules https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g
- Twitter word vectors can be downloaded from the following link, and should be saved in the working directory containing the main modules https://drive.google.com/file/d/1lw5Hr6Xw0G0bMT1ZllrtMqEgCTrM7dzc/view

Description of the three main modules:

- "single_models.py": this module evaluates the performance of each single model for a dataset of your selection; it inputs the extracted features of each single model to SVM and RF classifiers, tunes the classifiers using cross-validation, computes the scores of the single models, and saves the results in excel files in the working directory.
- "ensemble_models.py": this module computes the CV and test scores of the three ensembles (SVM, RF, and SVM RF) for a dataset of your selection. It prints the scores and saves them in excel files in the working directory.
- "total_score.py": this module computes the CV and test scores of the three ensembles for the five targets at once, then computes the total test scores of the three ensembles across the targets. The CV and test scores for each target and each ensemble are printed and saved in excel files in the working directory. Also, the total score of each ensemble across the five targets is printed on the screen.

Instructions:

Here is a list of the files and folders that should be in the working directory containing the modules: ["Stance.db", "dictionary.csv", "resources", "parser", "twise", "climate", "abortion", "atheism", "feminist", "hc"]. The description of these files and folders is provided at the end of the file. Please follow these instructions to run the code:

1) The results of the single models are saved in "RESULTS" folder on Github as "TARGET_svm_pval_0.05_cv5.xlsx" and "TARGET_rf_pval_0.05_cv5.xlsx", where "TARGET" refers to the name of the target. TARGET can be "ab" for Abortion, "hc" for Hillary, "fem" for Feminist, "ath" for Atheism, or "clm" for Climate. To regenerate these results for a specific dataset of your selection, follow these instructions:
    - Run the script "single_models.py"
    - You will be asked to enter 1 to select Hillary dataset, 2 to select Feminist dataset, 3 to select Abortion, 4 to select Atheism, or 5 to select Climate dataset.
    - Enter the number that corresponds to the dataset you are interested in.
    - Finally, the excel files containing the results ("TARGET_svm_pval_0.05_cv5.xlsx" and "TARGET_rf_pval_0.05_cv5.xlsx") will be created in the working directory.

2) The results of the ensemble models are saved in "RESULTS" folder on Github as "TARGET_Ensembles_CV.xlsx" and "TARGET_Ensembles_Test.xlsx", where TARGET refers to the name of the target. To regenerate these results for a specific dataset of your selection, follow the instructions in (a). To regenerate the results for all targets at once, and compute the total score across the five targets, follow the instructions in (b).

    Instructions (a) to regenerate the results of the ensemble models for a specific dataset:
    - "ensemble_models.py" reads the tuned parameters of the single models. So, you should ensure that the files "TARGET_svm_pval_0.05_cv5.xlsx" and "TARGET_rf_pval_0.05_cv5.xlsx" are found in the working directory containing "ensemble_models.py" before running it. You can either download these excel files from "RESULTS" folder and put them in the working directory containing the script or regenerate them by running "single_models.py" before running "ensemble_models.py".
    - Run the script "ensemble_models.py"
    - You will be asked to enter 1 to select Hillary dataset, 2 to select Feminist dataset, 3 to select Abortion, 4 to select Atheism, or 5 to select Climate dataset.
    - Enter the number that corresponds to the dataset you are interested in.
    - Results will be printed on screen.
    - Finally, the excel files containing the results ("TARGET_Ensembles_CV.xlsx" and "TARGET_Ensembles_Test.xlsx") will be created in the working directory.

    Instructions (b) to regenerate the results of the three ensembles for all targets at once and compute the total scores across the five targets:
    - Make sure that the files "TARGET_svm_pval_0.05_cv5.xlsx" and "TARGET_rf_pval_0.05_cv5.xlsx" are found in the working directory containing "total_score.py" before running it since it reads the tuned parameters of the single models. You can either download these excel files from "RESULTS" folder and put them in the working directory containing the script or regenerate them by running "single_models.py" before running "total_score.py".
    - Run the script "total_score.py"
    - CV and Test scores of the three ensembles for all targets will be printed on the screen and saved in excel files in the working directory.
    - Finally, the final test scores across the five targets will be printed on the screen.

Description of the other modules that are called by the main modules described above:

- "Features_manager_modified.py": it performs feature extraction and saves the extracted features of the twelve single models inside the folders called "DATA", where "DATA" refers to the name of the target or the sub-dataset. "DATA" can be "hc", "climate", "abortion", "feminist", or "atheism".
- "tweet_preprocess.py" and "Tweet.py": they perform tweet preprocessing. They read unstructured data from the database "Stance.db" and save the pre-processed tweets for each dataset in two directories "DATA/nodic/" and "DATA/newdic/", where "DATA" refers to the name of the target or the sub-dataset.
- "majority_cv.py": it computes the CV scores of the three ensembles.
- "majority_test.py": it computes the test scores of the three ensembles.
- "stanford_parser.py": it is called by the feature extraction module to generate dependency features. It needs some files that are saved in "parser" folder; this folder should be in the working directory.
- "twise.py" and "twitterTokenizer.py": they are called by the feature extraction module to compute the sentiment features inspired from (Balikas and Amini, 2016). They need some files that are saved in "twise" folder; this folder should be in the working directory.
- "Linguistic_resource_GI.py": it generates sentiment features based on General Inquirer (GI) lexicon.
- "Linguistic_resource_DAL.py": it generates sentiment features based on Dictionary of Affect in Language (DAL).
- "Linguistic_resource_HL.py": it generates some sentiment features based on Bing Liu lexicon.
- "Linguistic_resource_AFINN.py": it generates sentiment features based on AFINN lexicon.

Description of folders and files that are called by the modules and should be in the working directory

- "Stance.db": this file contains the unstructured datasets.
- "dictionary.csv": this file contains the manual dictionary.
- "resources": this folder contains the lexicons and resources used in feature extraction and pre-processing stages.
- "parser": this folder contains some files that are called by stanford_parser.py.
- "twise": this folder contains some files that are called by twise.py.
- "climate": this folder contains the pre-processed tweets for Climate dataset and the extracted features for its single models. The pre-processed tweets are found in two subfolders "nodic" and "newdic". The folder "nodic" contains the pre-processed tweets without using the dictionary, while "newdic" contains the pre-processed tweets using the dictionary.
- "abortion": This folder contains the pre-processed tweets for Abortion dataset and the features extracted for its single models.
- "atheism": This folder contains the pre-processed tweets for Atheism dataset and the features extracted for its single models.
- "feminist": This folder contains the pre-processed tweets for Feminist dataset and the features extracted for its single models.
- "hc": This folder contains the pre-processed tweets for Hillary dataset and the features extracted for its single models.

References

G. Balikas and M.-R. Amini, Twise at semeval-2016 task 4: Twitter sentiment classification, ArXiv Preprint ArXiv1606.04351 (2016).