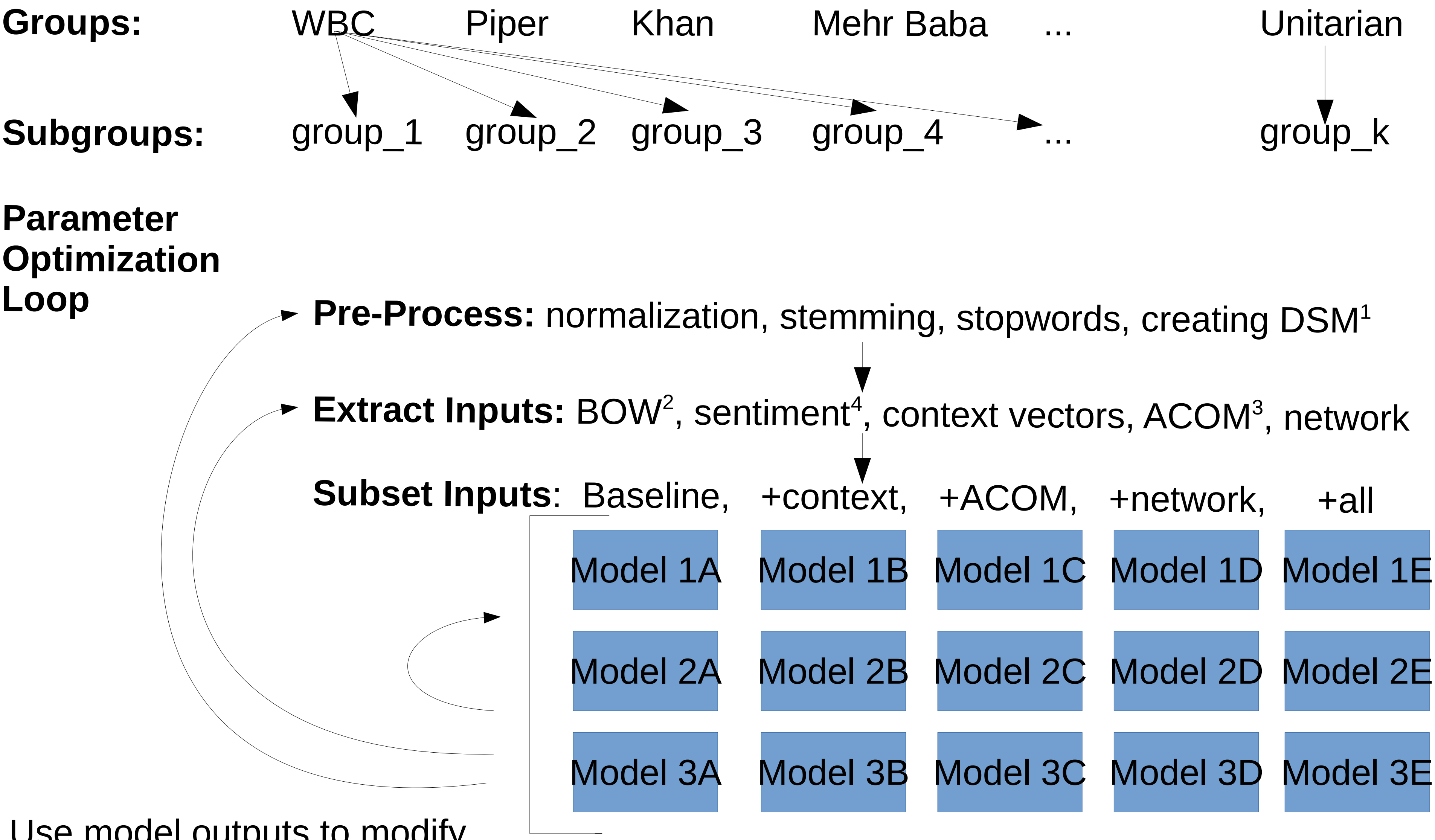


Groups: WBC Piper Khan Mehr Baba ... Unitarian

Subgroups: group_1 group_2 group_3 group_4 ... group_k

Parameter Optimization Loop

Pre-Process: normalization, stemming, stopwords, creating DSM[1]

Extract Inputs: BOW[2], sentiment[4], context vectors, ACOM[3], network

Subset Inputs:

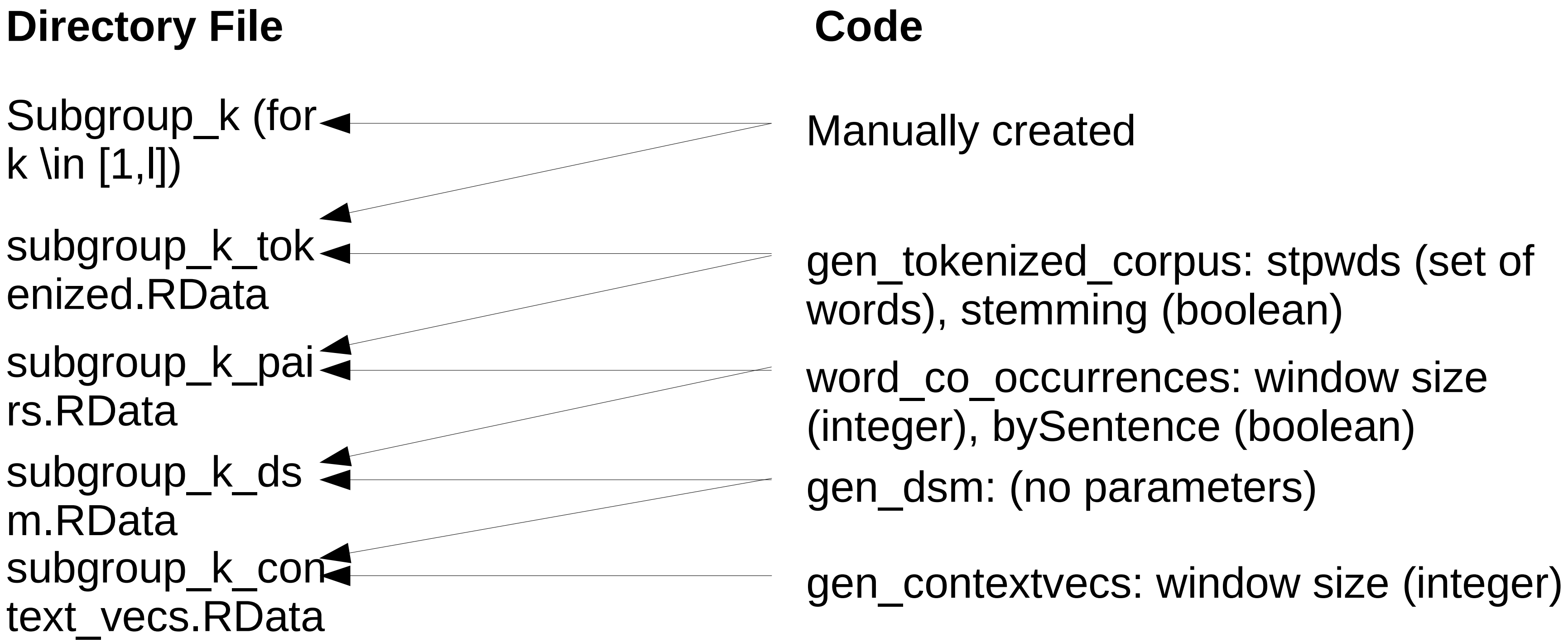
Baseline,	+context,	+ACOM,	+network,	+all
Model 1A	Model 1B	Model 1C	Model 1D	Model 1E
Model 2A	Model 2B	Model 2C	Model 2D	Model 2E
Model 3A	Model 3B	Model 3C	Model 3D	Model 3E

Use model outputs to modify hyperparameters governing pre-processing, input extraction, model parameter for each model variant independently

Output: For each subset of inputs and each one of 3 model types, we will have a preprocessing and input-extraction process that yields the best results.

With these ~15 'best models', we can calculate the impact of context vectors, ACOM, and network metrics. *Note: This can be paired down if it's too demanding*

Script/directory architecture
(how to generate context vectors)



Directory File	**Code**
Subgroup_k (for k \in [1,l])	Manually created
subgroup_k_tokenized.RData	gen_tokenized_corpus: stpwds (set of words), stemming (boolean)
subgroup_k_pairs.RData	word_co_occurrences: window size (integer), bySentence (boolean)
subgroup_k_dsm.RData	gen_dsm: (no parameters)
subgroup_k_context_vecs.RData	gen_contextvecs: window size (integer)

Standards for final version of functions

- All code should be encapsulated within separate scripts with functions to be called in by a master script

- The scripts should
--- have only one or a handful of closely-relate functions
--- not execute any code or modify the environment (outside of declaring necessary functions)
--- easily be called with the “source” function in R

- The functions should
--- Accept as inputs all parameters
--- Have no default values
--- Accept as input the file containing the data on which to operate
--- Accept as input the file to which to save the output