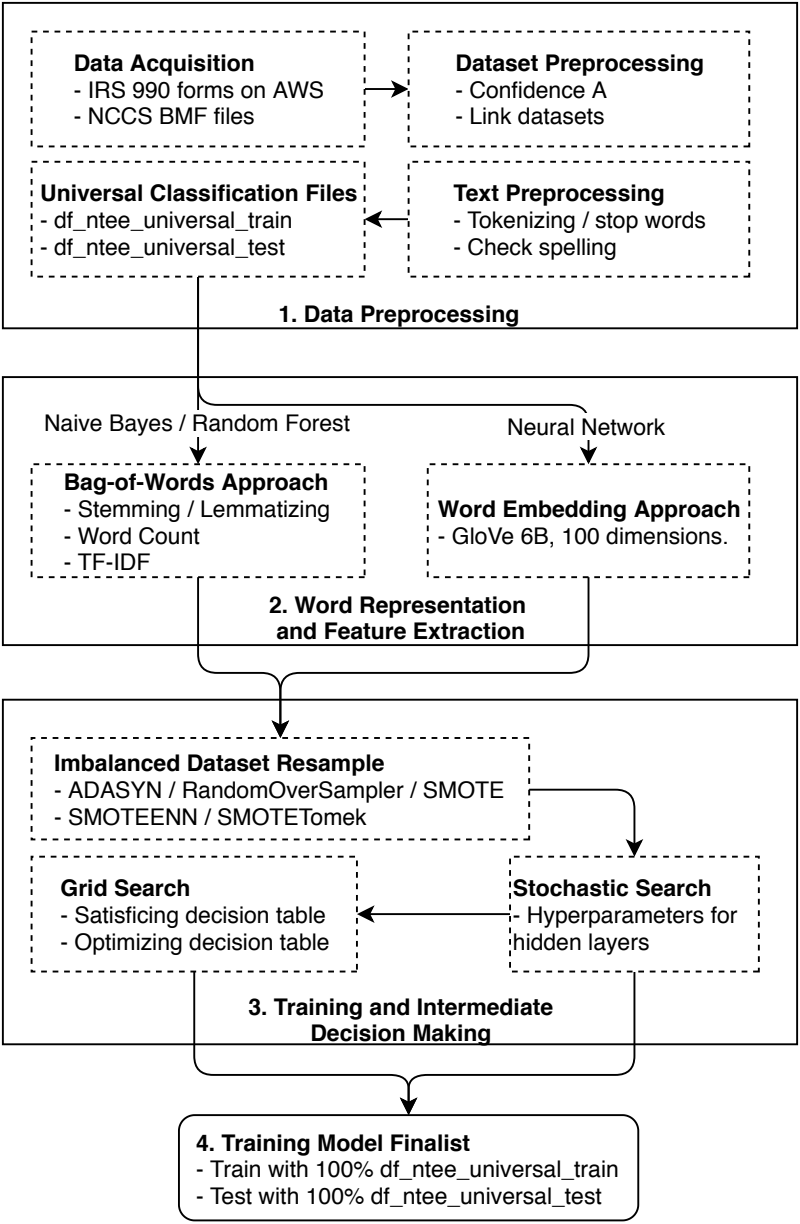
Data Acquisition
- IRS 990 forms on AWS
- NCCS BMF files
Dataset Preprocessing
- Confidence A
- Link datasets
Universal Classification Files
- df_ntee_universal_train
- df_ntee_universal_test
Text Preprocessing
- Tokenizing / stop words
- Check spelling
1. Data Preprocessing
Naive Bayes / Random Forest
Neural Network
Bag-of-Words Approach
- Stemming / Lemmatizing
- Word Count
- TF-IDF
Word Embedding Approach
- GloVe 6B, 100 dimensions.
2. Word Representation and Feature Extraction
Imbalanced Dataset Resample
- ADASYN / RandomOverSampler / SMOTE
- SMOTEENN / SMOTETomek
Grid Search
- Satisficing decision table
- Optimizing decision table
Stochastic Search
- Hyperparameters for hidden layers
3. Training and Intermediate Decision Making
4. Training Model Finalist
- Train with 100% df_ntee_universal_train
- Test with 100% df_ntee_universal_test