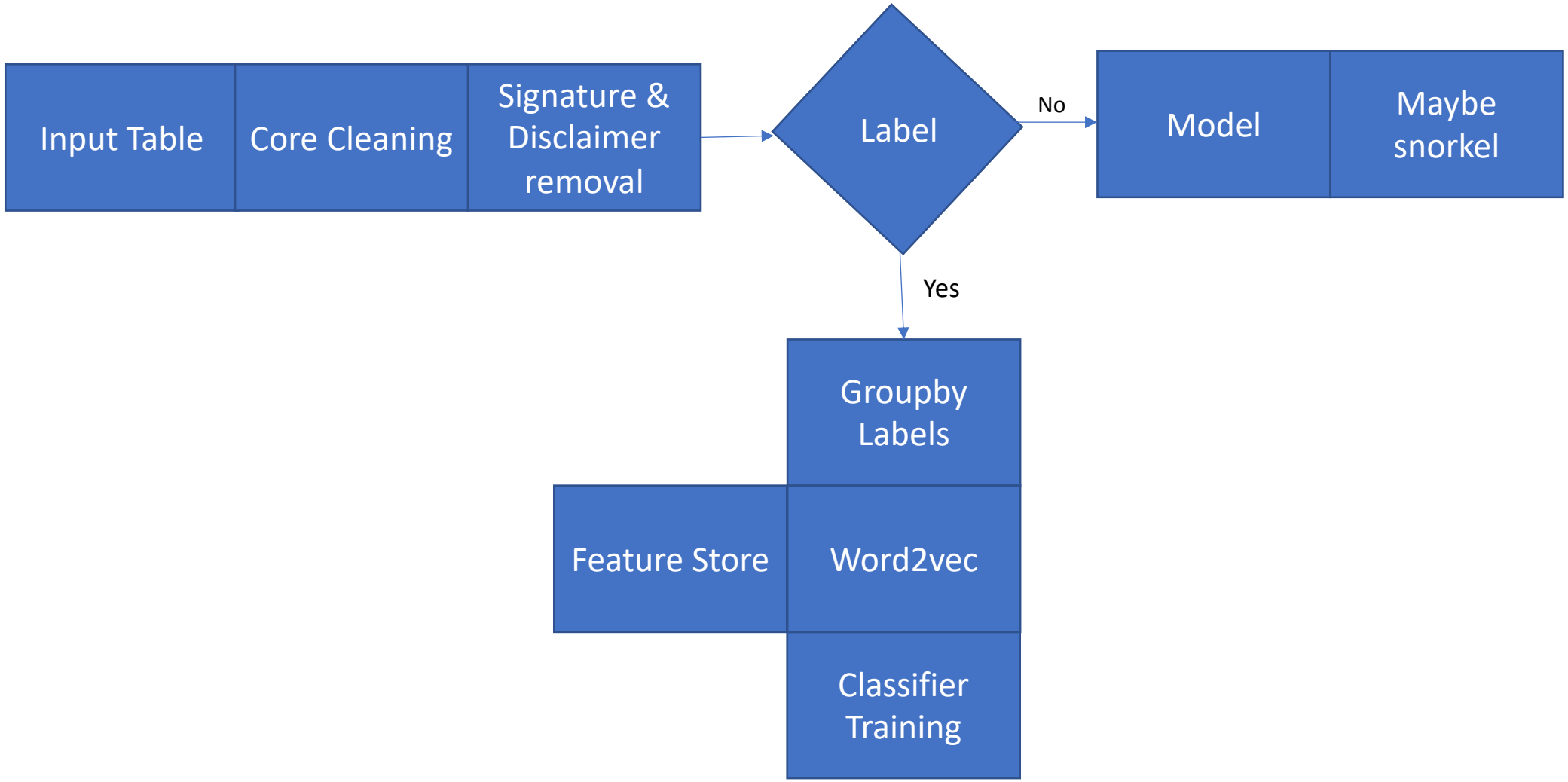
Data Pipeline

Pipeline


Input Table
Core Cleaning
Signature & Disclaimer removal
Label
No
Model
Maybe snorkel
Yes
Groupby Labels
Feature Store
Word2vec
Classifier Training

- Scikit Learn pipeline has CountVectoriser, TfIdfTransformer, MachineLearning Classifier. This is good because it is very optimized.
- Inhouse model can be constructed using nltk,pyspark it has ne_tree,pos_tag etc., allows customization.
- Snorkel can also be used because it has various member functions that are helpful like lemma, NER, POS. We can use re tags to write label functions and still obtain better accuracy, probably we can use a model as LF
- Snorkel is interesting and should be easy to try out on the text data.
- Can try VAE after tfidf or word2vec for autolabelling and feed to snorkels generative model.