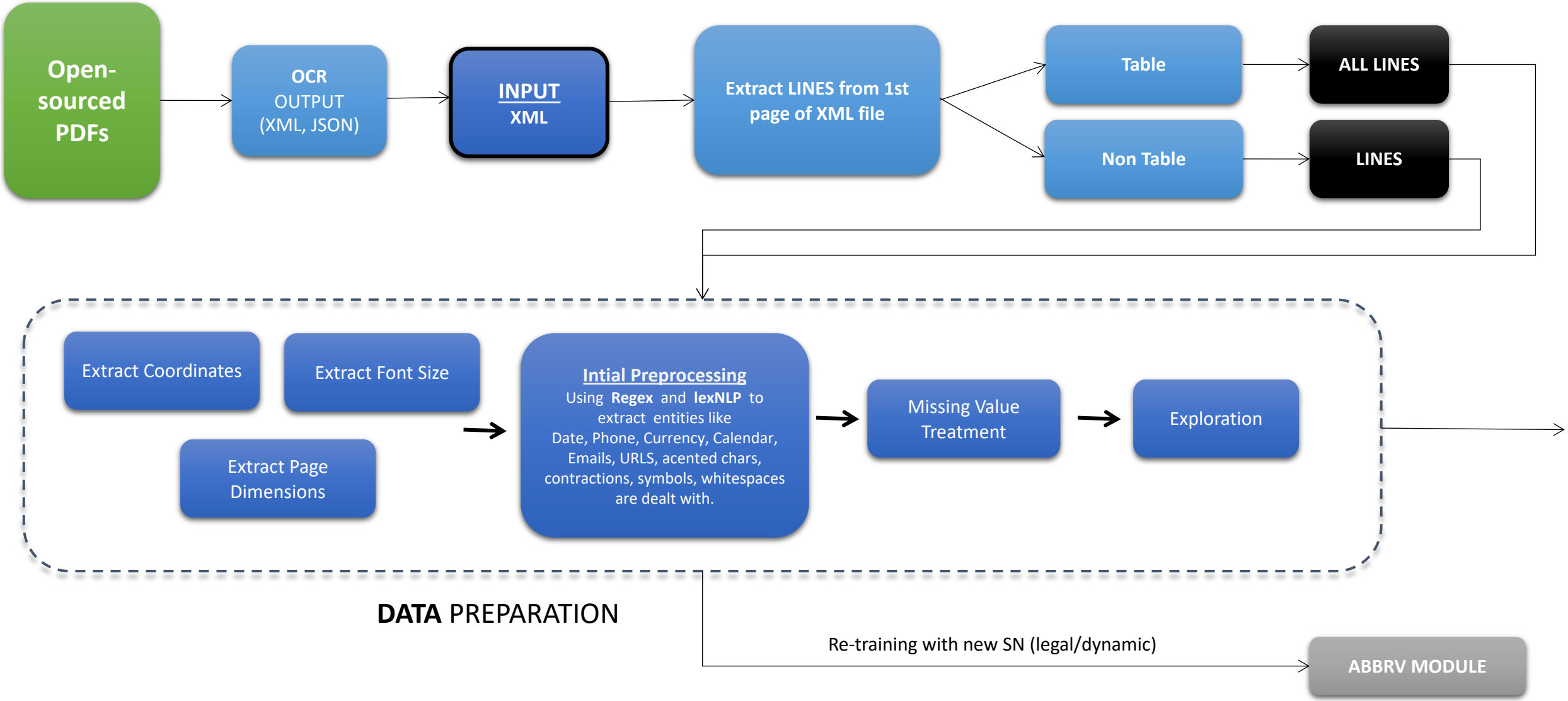
Flowchart

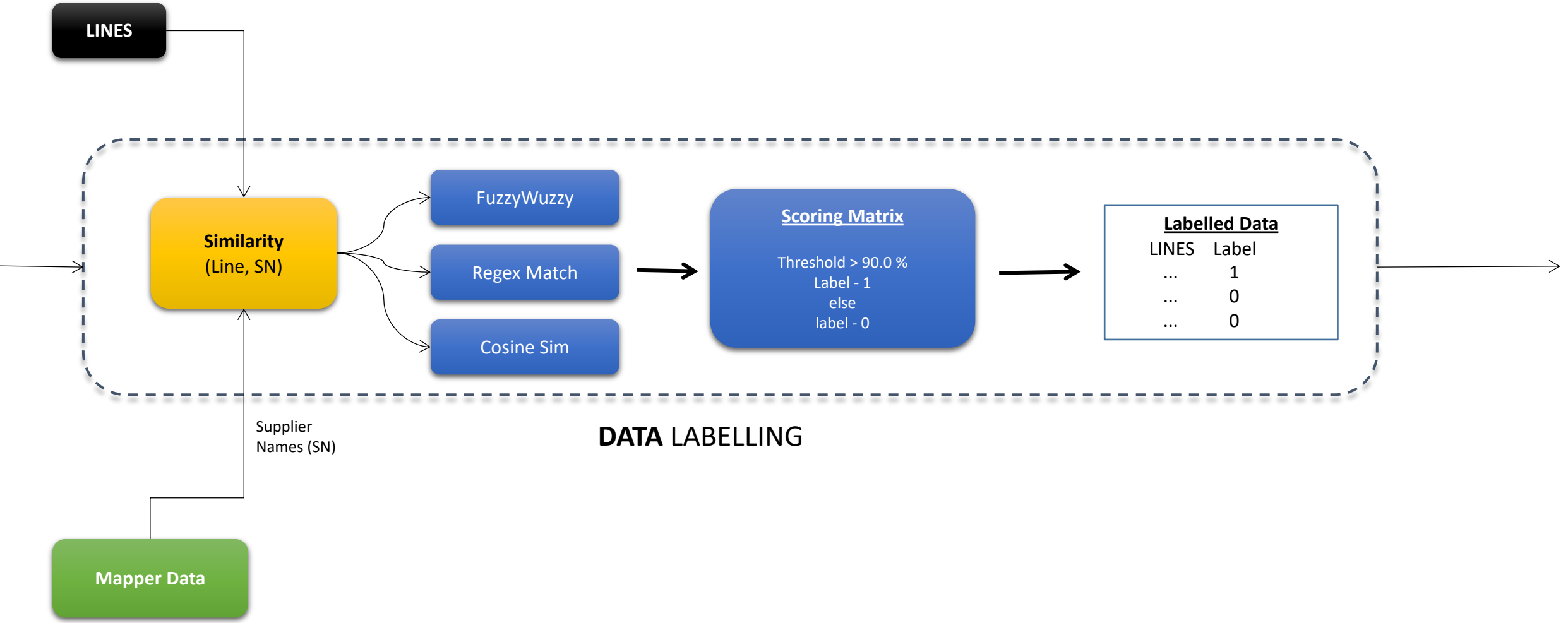

Open-sourced PDFs
OCR
OUTPUT
(XML, JSON)
INPUT
XML
Extract LINES from 1st page of XML file
Table
Non Table
ALL LINES
LINES
Extract Coordinates
Extract Font Size
Extract Page Dimensions
Intial Preprocessing
Using Regex and lexNLP to extract entities like Date, Phone, Currency, Calendar, Emails, URLS, acented chars, contractions, symbols, whitespaces are dealt with.
Missing Value Treatment
Exploration
DATA PREPARATION
Re-training with new SN (legal/dynamic)
ABBRV MODULE

Abbreviations

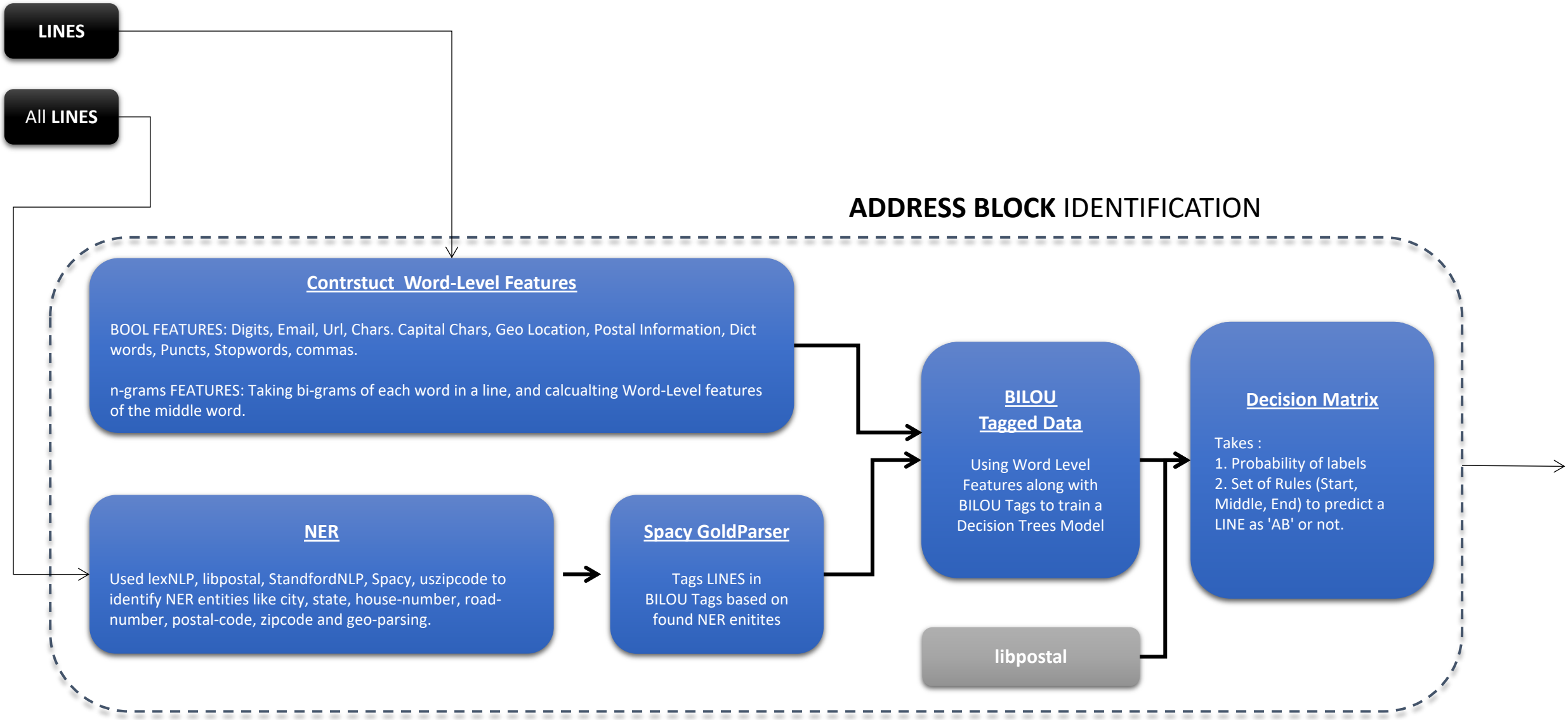
A company name could contain a *legal extension/abbreviation* in its name which was used as a feature in the whole process.

1. **List of Legal abbreviations:** A Wikipedia page containing all known legal abbrvs country-wise was scrapped using beautiful soup to produce a static list of known legal abbrvs. Example: XYZ **Inc.;** ABC **LLC**, etc.
2. **List of Dynamic abbreviations:** The last two token of all given supplier names were observed and using a TF-IDF score of all those tokens having an occurrence in at least 50 unique documents were considered and allowed to enter this list. This list was merged with the list of legal abbrvs.
3. **The combined list** was directly used to find the matching last two tokens in all the LINES across pdf. If a match was found, it was put to regex-based cleaning and was considered as a found abbrv.

Flowchart (Continued..)



Flowchart (Continued..)



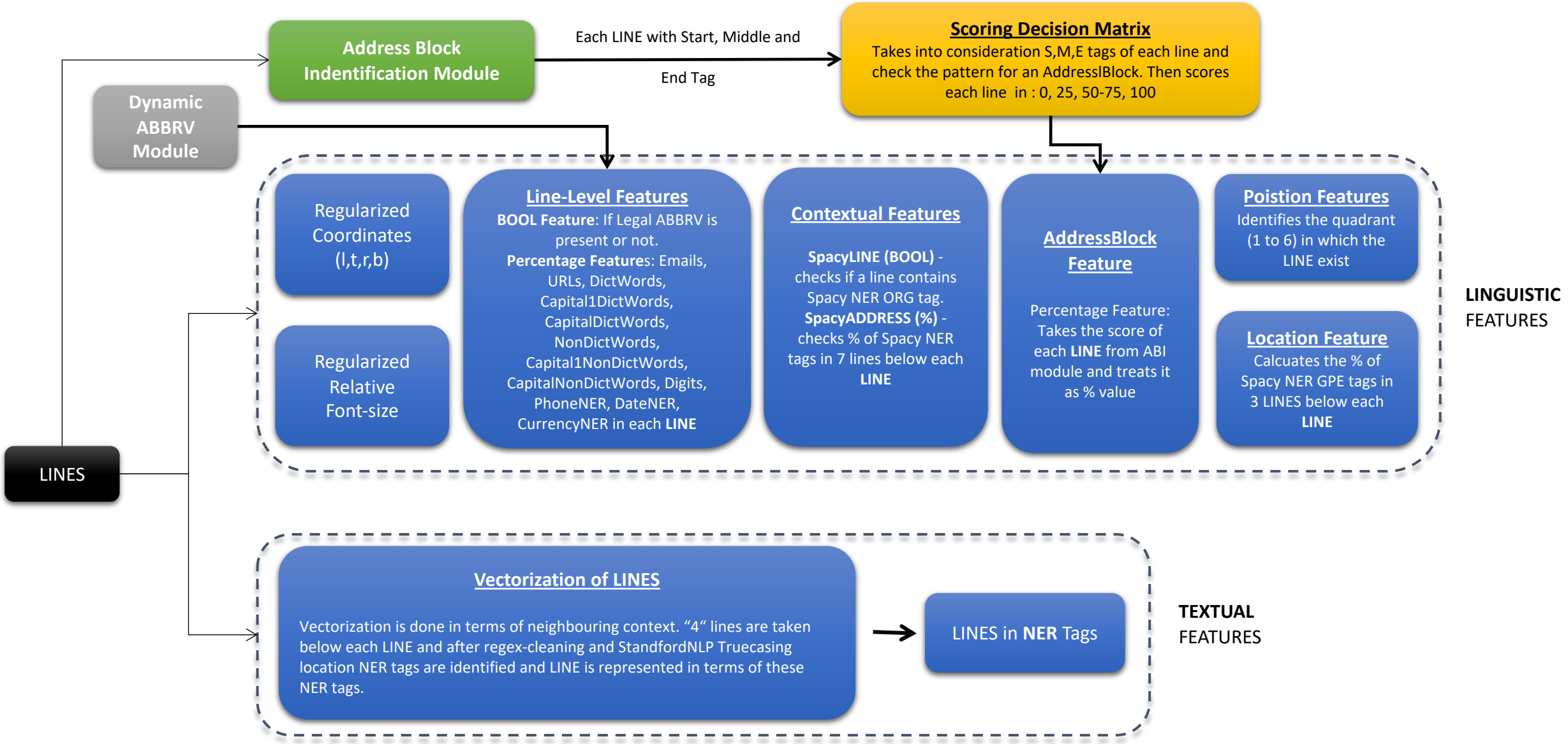
Address Block Identification

Each PDF would contain an Address Block following some common rules of address convention. These rules are identified, trianed upon and utilised to create a decision matrix that would help in extracting address details.

For a pdf, n-grams (tri) were considered and various word-level festures (like presence of puncts, presence of digits, presence of citynames, etc) were calcuated for the middle term. These features are then used along with each BILUO labelled line (using NER and spacy's goldparser) to train a decision tree model. A biLSTM-CRF model was also trained on this annotated data but the DT model showed a better performance. Along with this a stored list of city-names, postal-address were also used to help in this identification process. A address parser library postal was also implemented to generate NER tags.

LINE	Tag	Label	Pattern	AB
Dated 10th March	None	O	O	O
ABC Pvt. Ltd.	None	O	O	O
14th road, A Block, F1 highway	Road number, street number	SS	SME	AB
Seattle, Washington 98101	city, state, postal code	ME	SME	AB
This invoice tells about....	None	O	O	O
31st , PA- 98101, Seattle, USA	Road number, state, postal code, city country	SMEMM	SMEMM	O

Flowchart (Continued..)



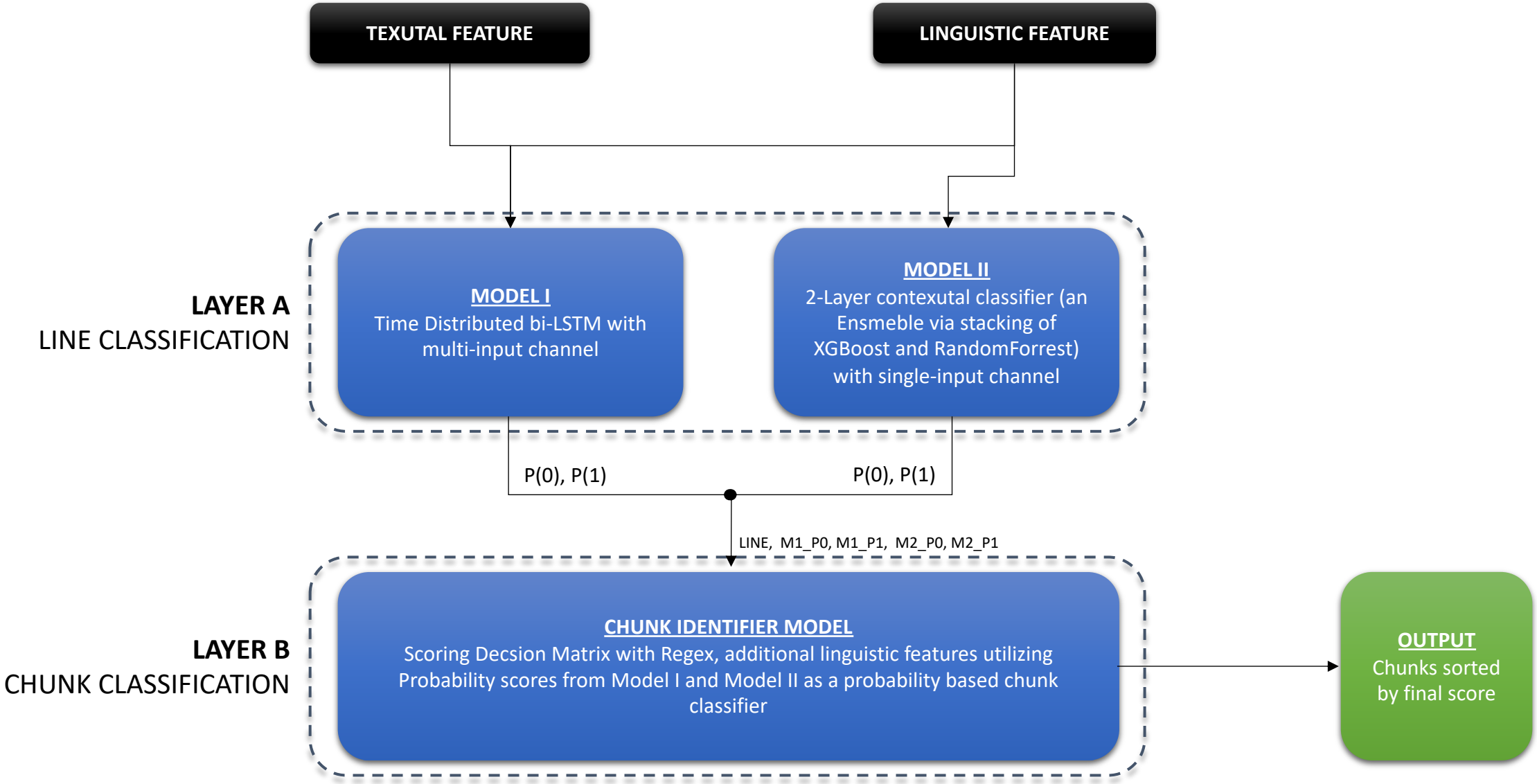
Flowchart: Experimented on few sample PDFs

P - Precision; R - Recall

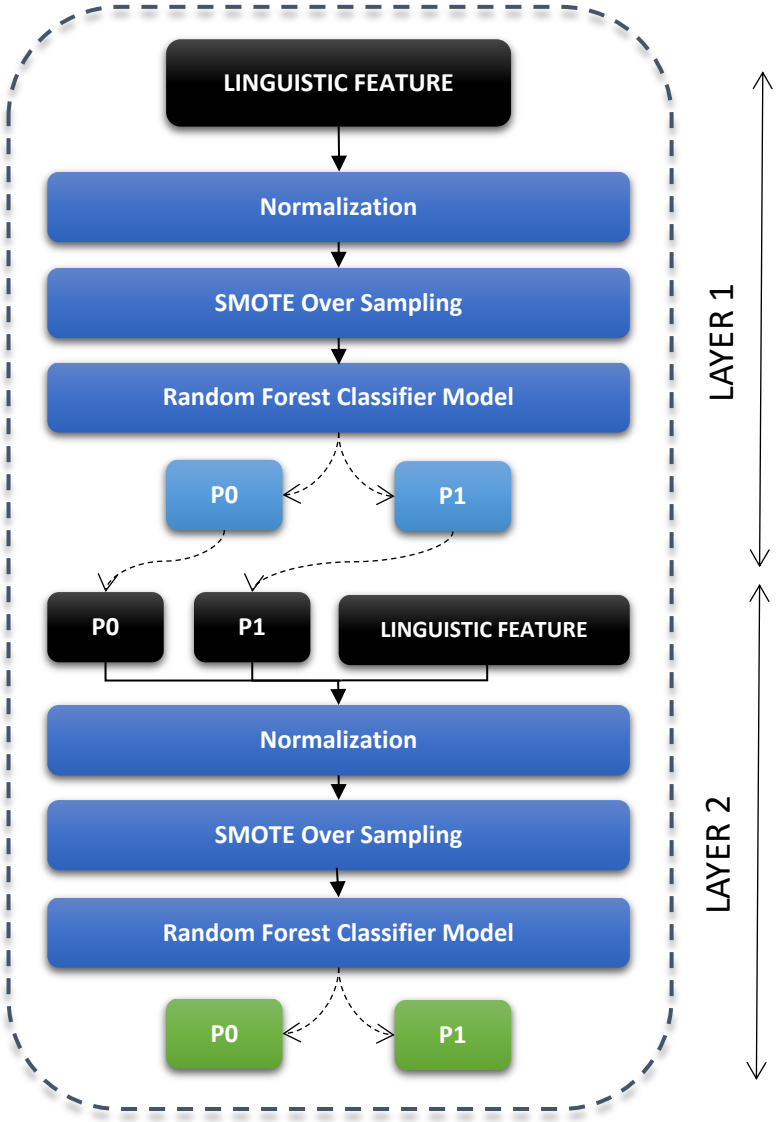
Flowchart: Experimented on few sample PDFs

P - Precision; R - Recall

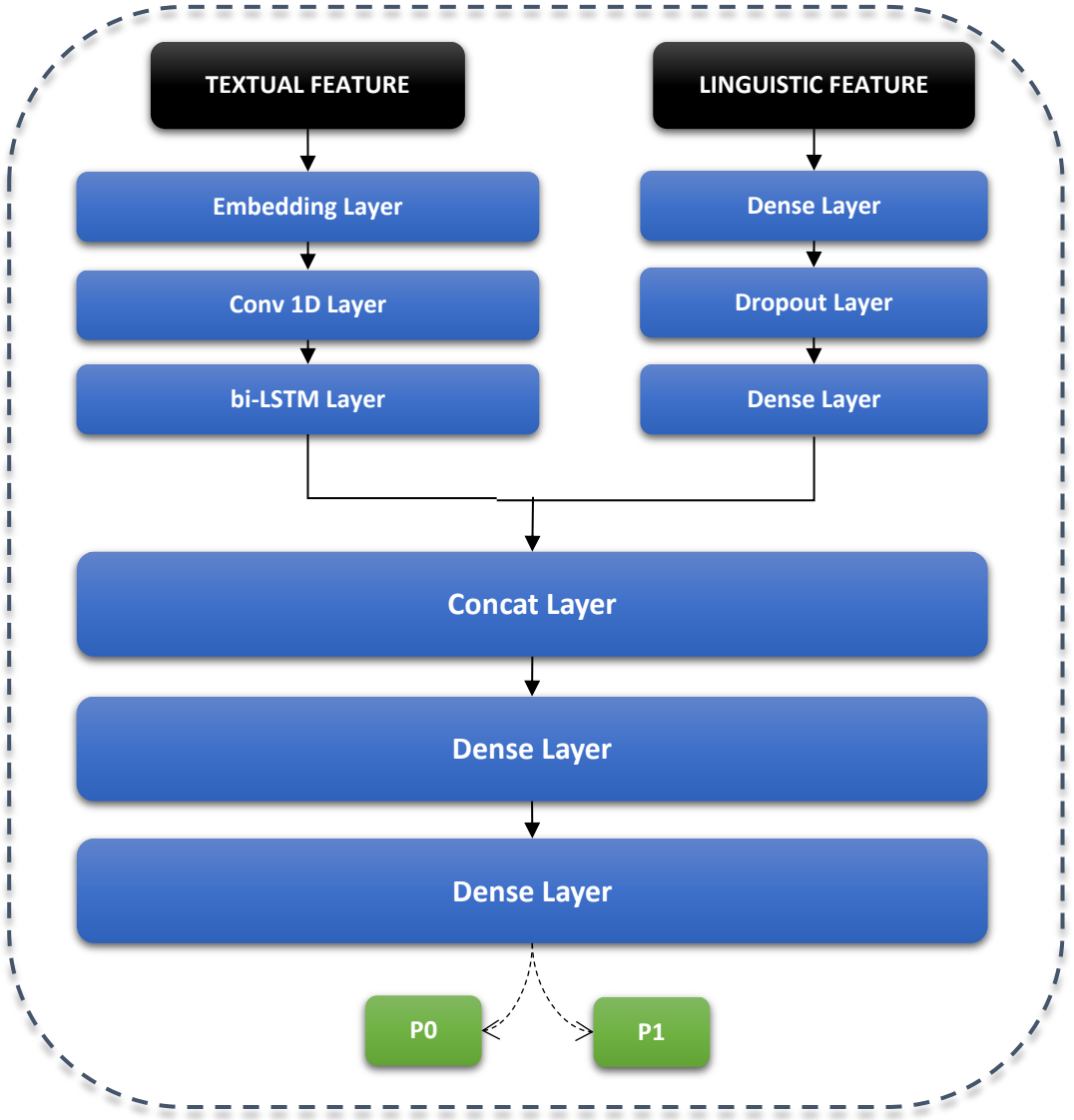
Final Architecture



Layer A - Line Classification



Model 1: Generic Ensemble RF Model



Model 2: Neural Network biLSTM Model

Layer B - Chunk Classification