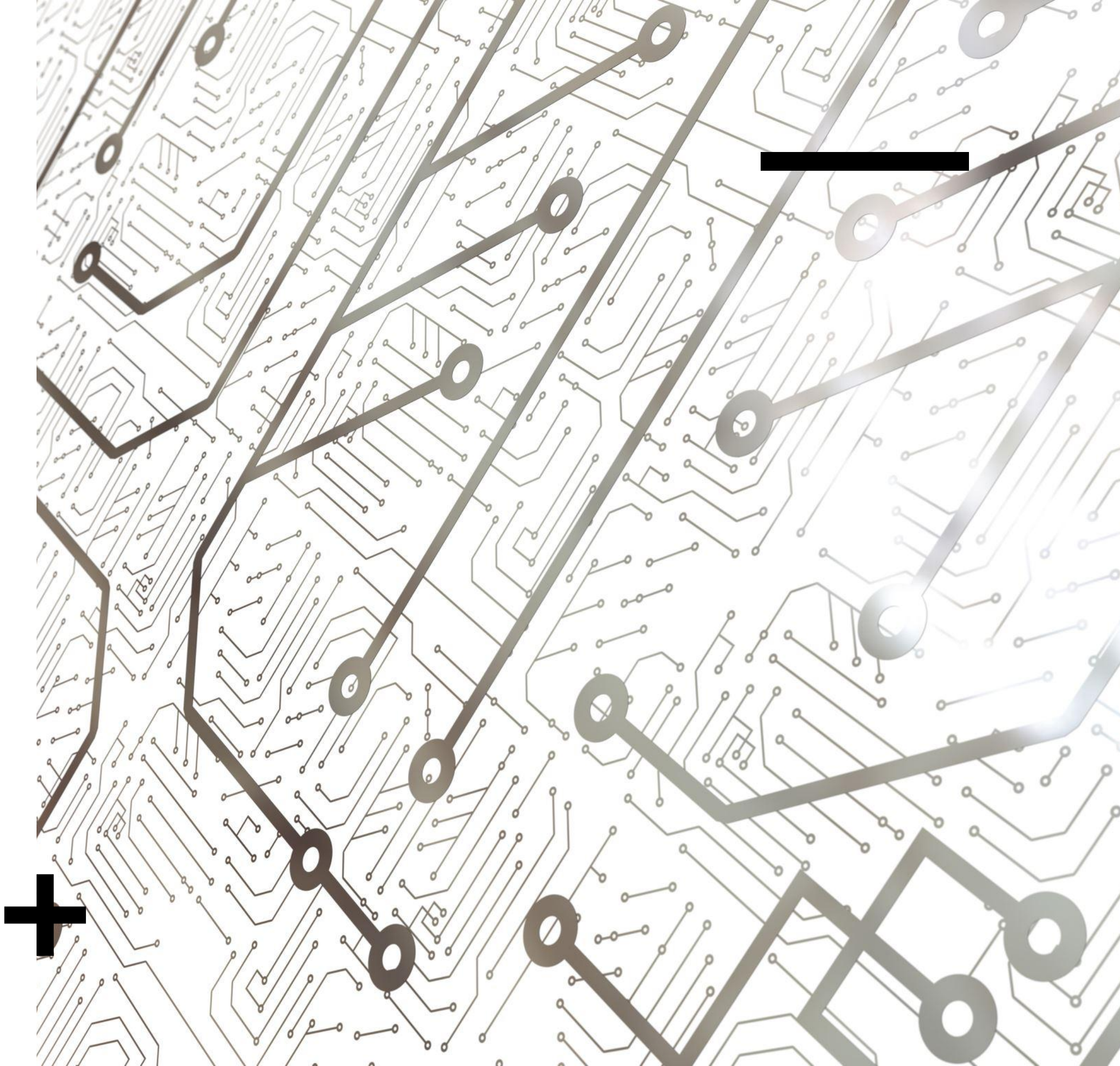
Capstone Project: SIGAL Team

Ash Chetty, Wei Zhu, Po-Chen Su, Harsh Shah, Haochen Wang

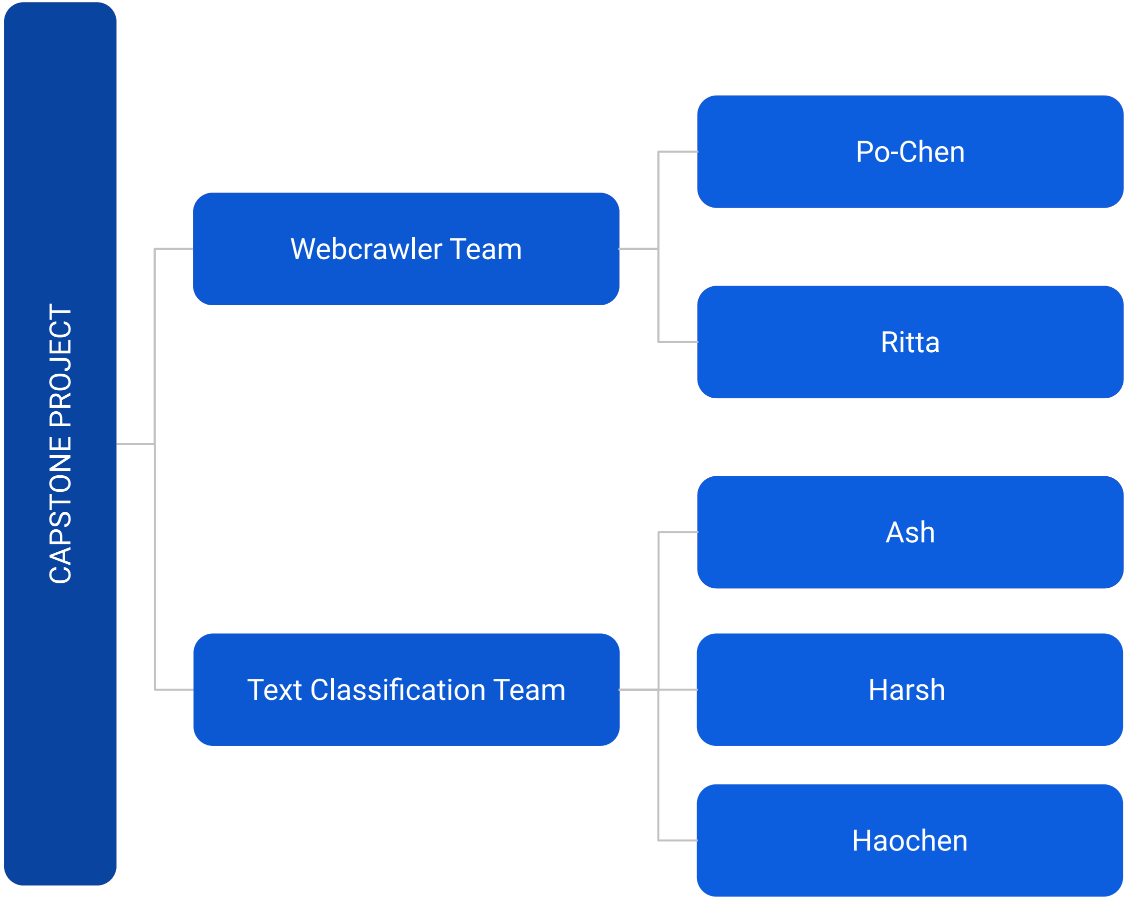


Introduction:

- **QQ Tech Objective**: To provide commonplace legal documents accessible to everyone.

- **SIGAL Team Goal**: To build a website crawler to extract all keyword-based HTML files from two different websites.
  - American Legal Publishing
  - Code of Federal Regulations
- To build a text classification model to extract “good” HTML files
  - Train multiple models to see which has the highest performance

Team Member Contribution:



American Legal Publishing

CODE OF FEDERAL REGULATIONS

Part 1: Website Crawler

Software

System

- MacOS
- 8GB or 16GB Ram
- 256GB, 1TB storage space

- Python - version 3.10
- Jupyter Notebook
- Pycharm - version 3.10
- Browser
  - Google Chrome
  - Safari
  - Firefox
  - IE

Python Packages

copy	Provides functions for creating and manipulating copies of objects.
csv	Provides classes for reading and writing comma-separated values (CSV) files.
os	Provides a way of interacting with the operating system.
re	Provides support for regular expressions (regex).
shutil	Provides a higher level interface for file operations.
requests	Allows sending HTTP/1.1 requests using Python.
BeautifulSoup	Parses HTML and XML documents.
selenium	Provides a way to automate web browsers.
pandas	Provides data structures and tools for data analysis.
time	Provides various time-related functions.
datetime	Provides classes for working with dates and times.
deque	Provides a data structure for a double-ended queue.
logging	Provides a way to log messages from your Python application.
RotatingFileHandler	A handler for logging to a file that automatically rotates the log file when it reaches a certain size.

Web Page Structure

American Legal Publishing

- State(s) (home page)
    - Locality(s) (cities)
        - Titles / Chapters (overview page)
            - Subchapters / Articles / Parts
                - Leaf page (contents)

Code of Federal Regulations

- Title(s) (home page)
    - Chapters
        - Subchapters / Parts
            - Leaf page (contents)

Differences:

Many States.
Each state has one or many localities.
Crawl on the overview page.

Many Titles.
Crawl on the each Title.

Map out the Targeting Result

All contents are written in **Leaf page**.

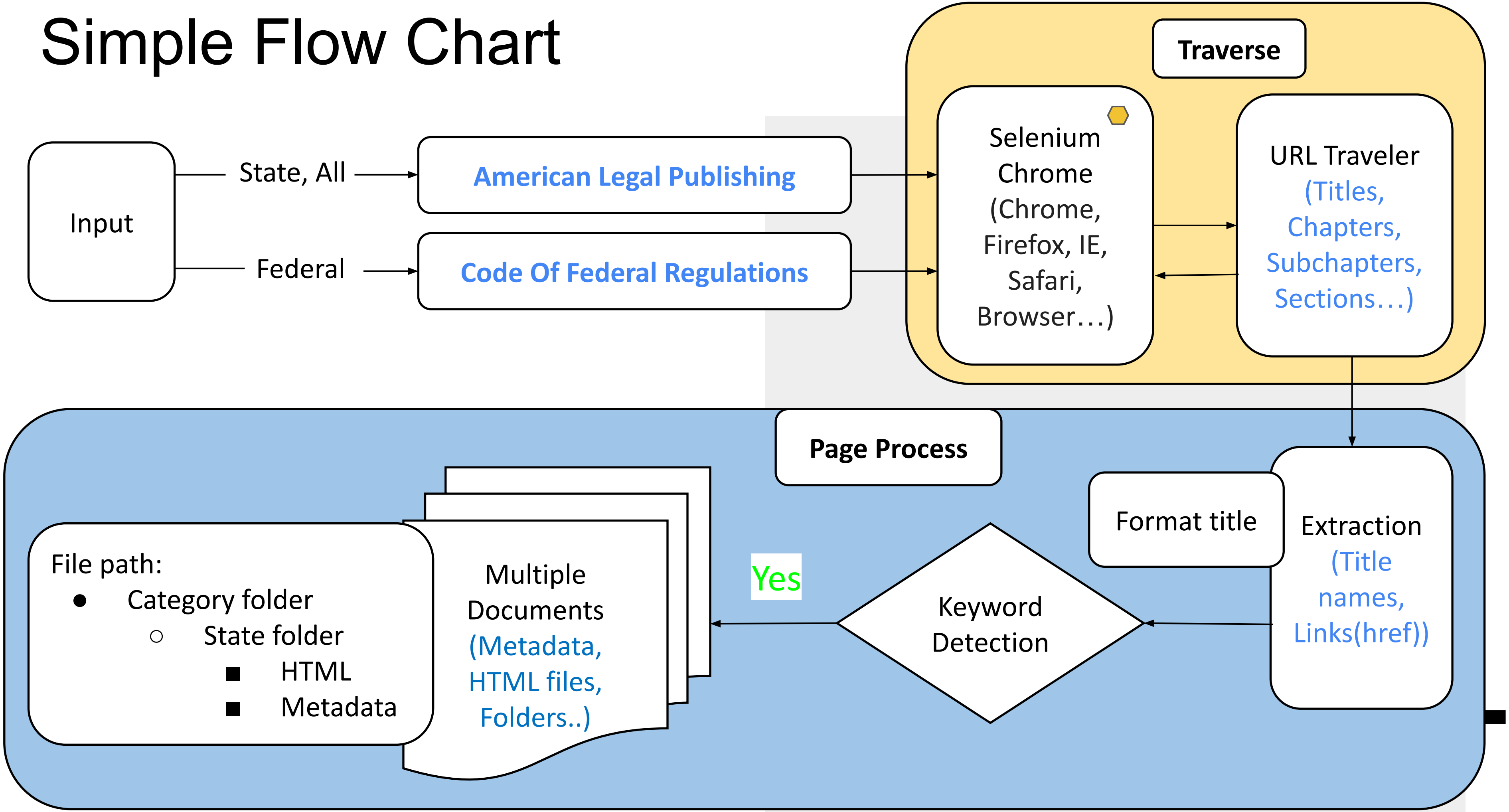
- If a **Chapter Title** is detected with any keywords, all content within that Chapter will be saved.
- If a **Subchapter Title** is detected with any keywords, all content within that Subchapter will be saved.
- If a **Leaf page Title** is detected with any keywords, the content of **that Leaf page** will be saved.

The information will be stored in different **Category Folders** based on the different **Keyword Categories**. (Category / State / HTMLs)

Keyword Category & Keywords

Category	Keywords
Residential Lease & Rental	rent; lease; leasing; tenant; landlord; housing; real property; real properties; real estate; residence; residential; premises; occupancy; eviction
Employment Documents	employ; labor; worker; working; talent; wage; salary; compensation; payroll; benefits; at will; workplace; job; occupation; profession; non disclosure; non competition; vacation time; time off; resign; termination; layoff
Consumer Agreements	consumer; end user; business; merchant; seller; sale; goods; commodity; commodities; retail; commercial; commerce; trade; price; pricing; loan; debt; credit; obligation; liability; indemnity; release; waiver

Simple Flow Chart



Input
State, All
Federal
American Legal Publishing
Code Of Federal Regulations
Traverse
Selenium Chrome (Chrome, Firefox, IE, Safari, Browser…)
URL Traveler (Titles, Chapters, Subchapters, Sections…)
Page Process
Format title
Extraction (Title names, Links(href))
Keyword Detection
Yes
Multiple Documents (Metadata, HTML files, Folders..)

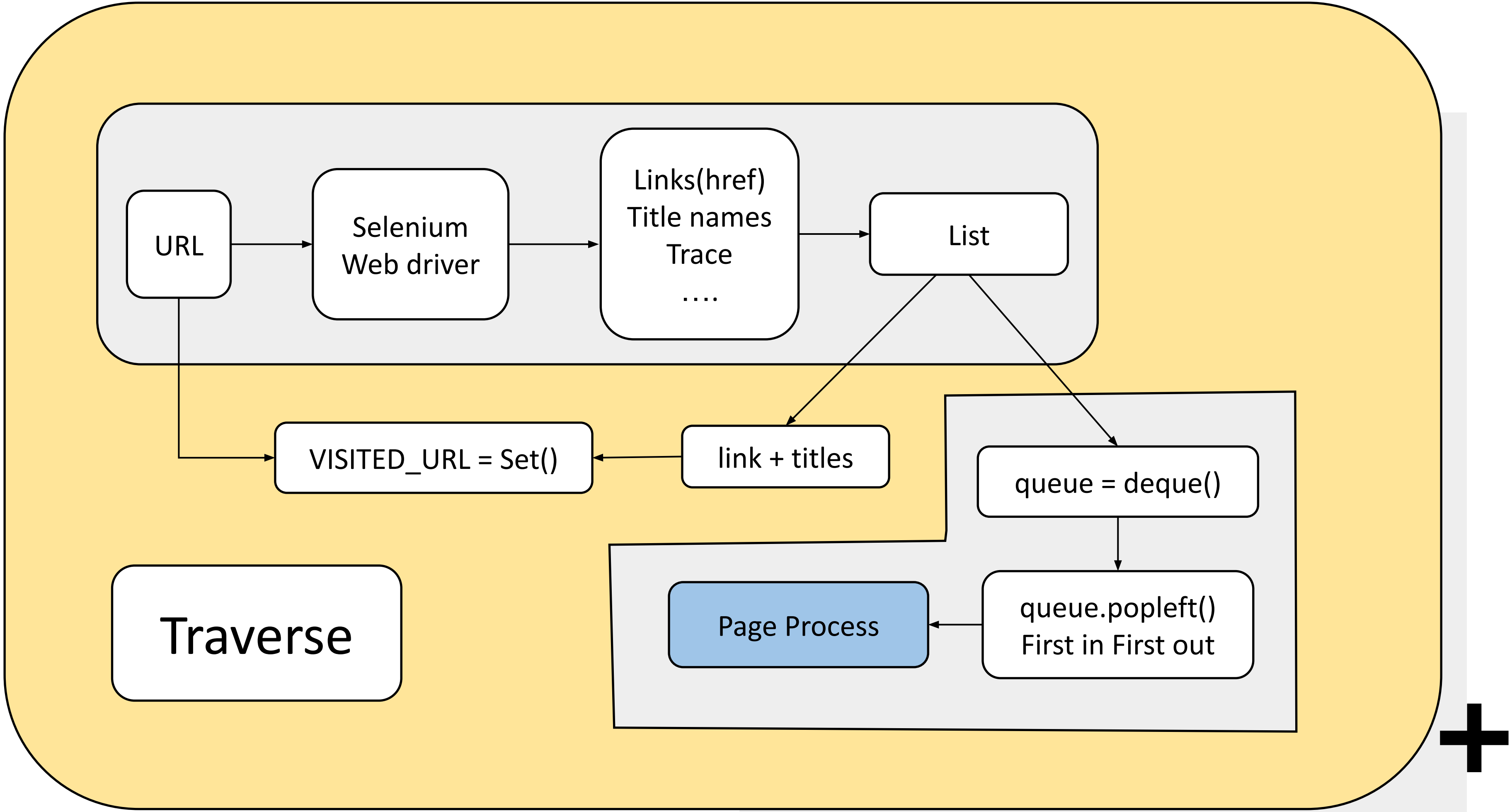
File path:

- Category folder
  - State folder
    - HTML
    - Metadata

Test Browsers

Test different web-drivers

Website	State & Locality	System	Browser	Mode	Run Time
American Legal Publishing	California Los Angeles	MacOS	Google Chrome	“head” And “headless”	12 ~ 13 hours
			Firefox		12 ~ 13 hours
			IE		12 ~ 13 hours
			Safari	“head”	12 ~ 13 hours
Code Of Federal Regulations	Federal	MacOS	Google Chrome	“head”	7 ~ 8 hours



URL

Selenium Web driver

Links(href)
Title names
Trace
….

List

VISITED_URL = Set()

link + titles

queue = deque()

queue.popleft()
First in First out

Page Process

Traverse

Website Traversal Logic

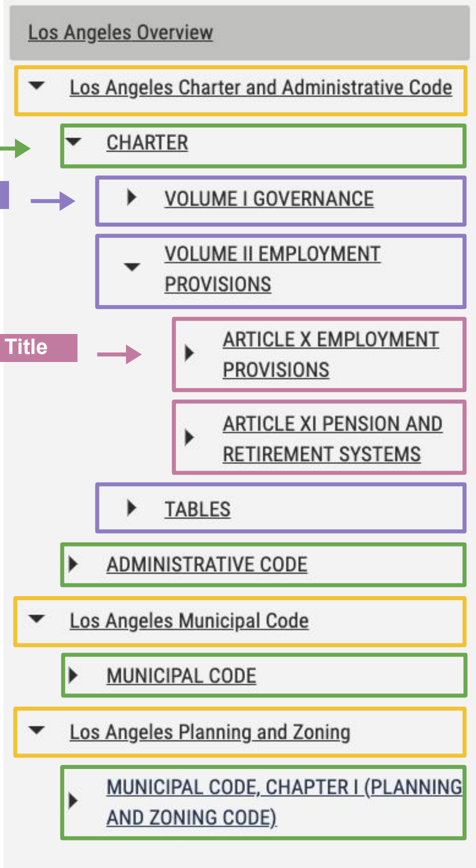
Root

L1 Title

L2 Title

L3 Title

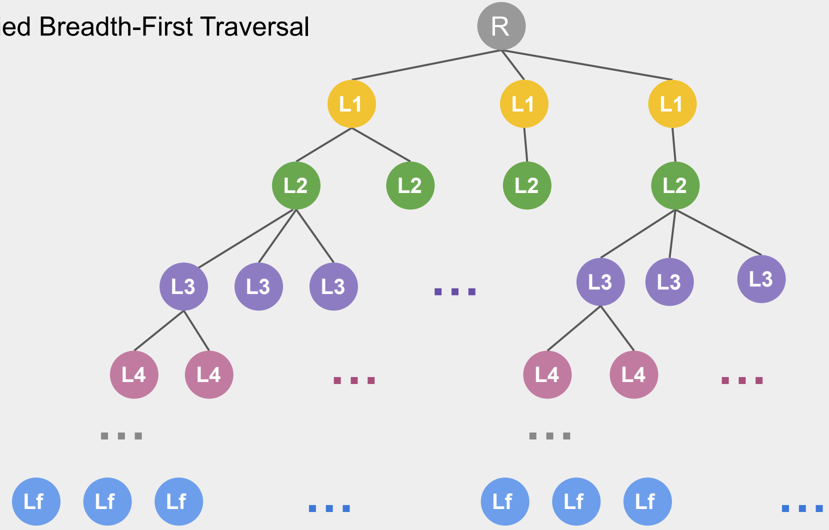
L4 Title



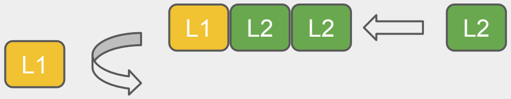
Los Angeles Overview

- Los Angeles Charter and Administrative Code
  - CHARTER
    - VOLUME I GOVERNANCE
    - VOLUME II EMPLOYMENT PROVISIONS
      - ARTICLE X EMPLOYMENT PROVISIONS
      - ARTICLE XI PENSION AND RETIREMENT SYSTEMS
    - TABLES
  - ADMINISTRATIVE CODE
- Los Angeles Municipal Code
  - MUNICIPAL CODE
- Los Angeles Planning and Zoning
  - MUNICIPAL CODE, CHAPTER I (PLANNING AND ZONING CODE)

- Applied Breadth-First Traversal



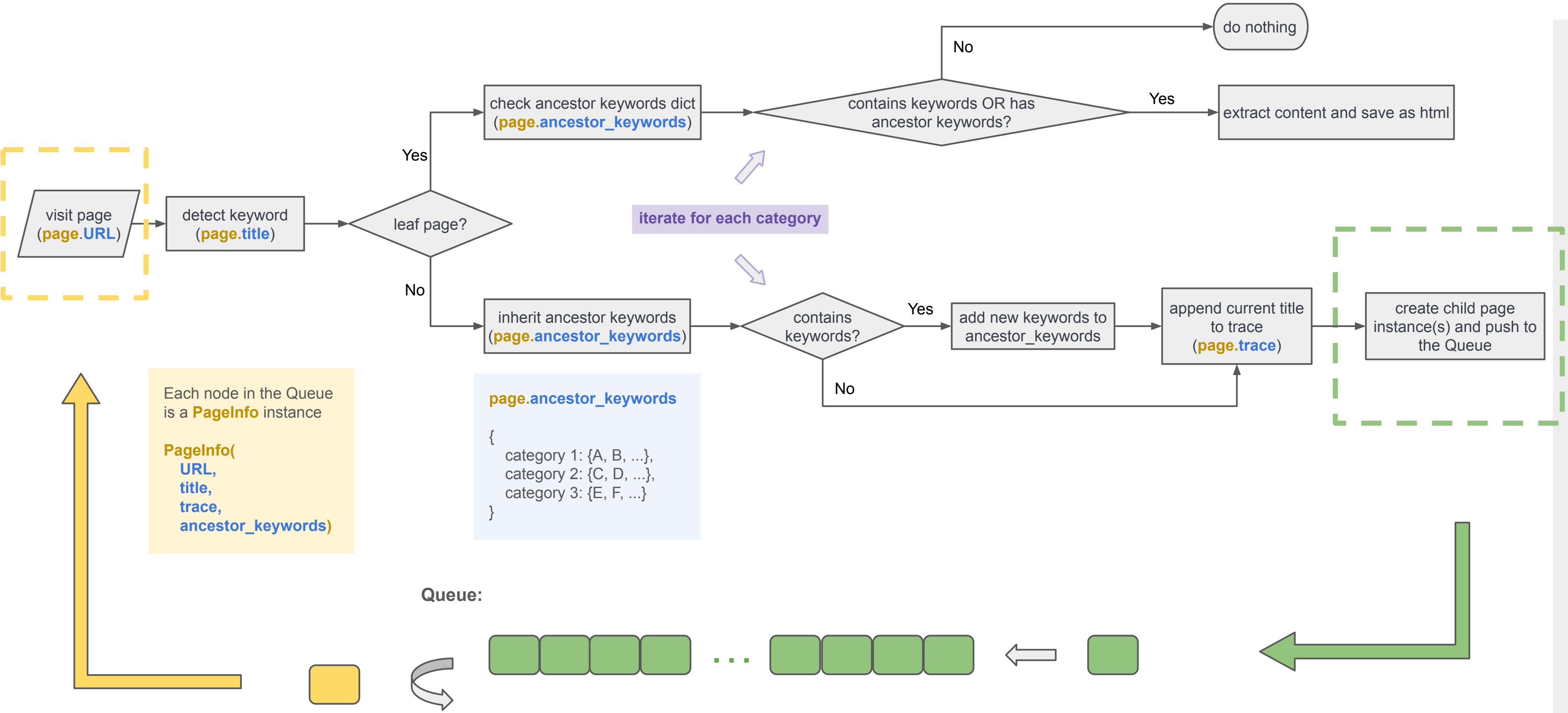
Queue:



- Each Node represent a PageInfo instance with 4 variables

PageInfo(URL, title, trace, ancestor_keywords)

Page Process Logic



visit page (page.URL)

detect keyword (page.title)

leaf page?

Yes

check ancestor keywords dict (page.ancestor_keywords)

contains keywords OR has ancestor keywords?

No

do nothing

Yes

extract content and save as html

No

inherit ancestor keywords (page.ancestor_keywords)

contains keywords?

Yes

add new keywords to ancestor_keywords

No

append current title to trace (page.trace)

create child page instance(s) and push to the Queue

iterate for each category

Each node in the Queue is a PageInfo instance

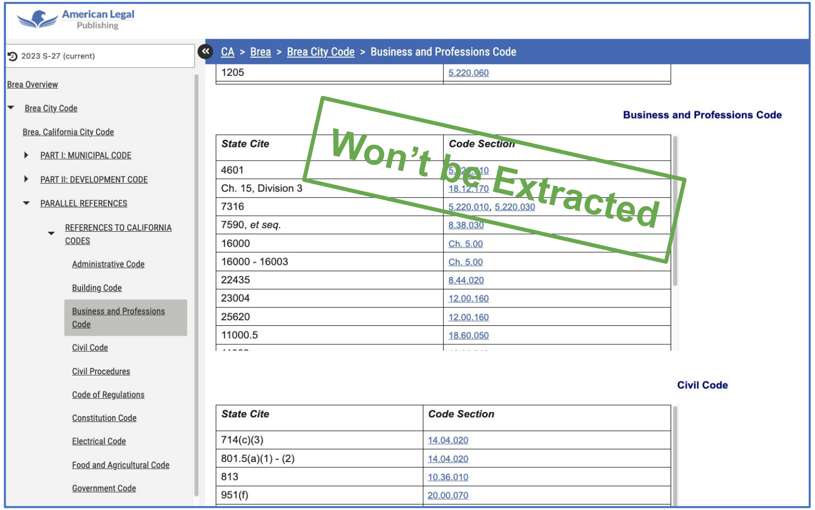
```
PageInfo(
    URL,
    title,
    trace,
    ancestor_keywords)
```

page.ancestor_keywords

```
{
    category 1: {A, B, ...},
    category 2: {C, D, ...},
    category 3: {E, F, ...}
}
```

Queue:

Corner Case Handling



CA > Brea > Brea City Code > Business and Professions Code

Brea Overview

- Brea City Code
  - Brea, California City Code
  - PART I: MUNICIPAL CODE
  - PART II: DEVELOPMENT CODE
  - PARALLEL REFERENCES
    - REFERENCES TO CALIFORNIA CODES
      - Administrative Code
      - Building Code
      - Business and Professions Code
      - Civil Code
      - Civil Procedures
      - Code of Regulations
      - Constitution Code
      - Electrical Code
      - Food and Agricultural Code
      - Government Code

1205	5.220.060

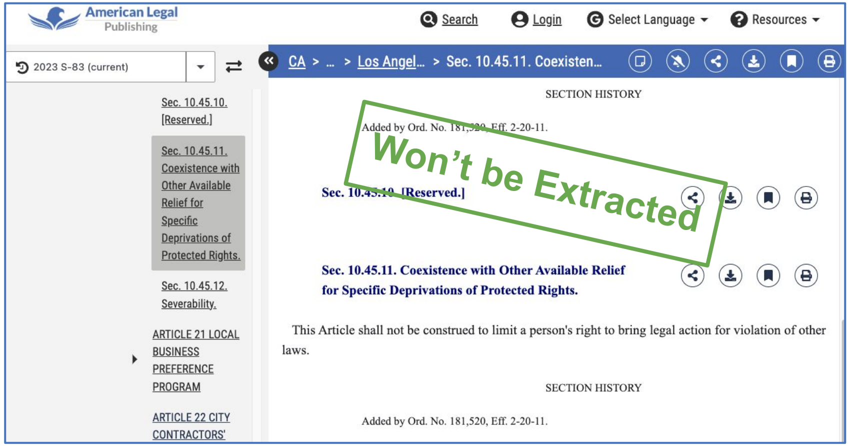
Business and Professions Code

State Cite	*Code Section*
4601	5.2[illegible]10
Ch. 15, Division 3	18.12.170
7316	5.220.010, 5.220.030
7590, *et seq.*	8.38.030
16000	Ch. 5.00
16000 - 16003	Ch. 5.00
22435	8.44.020
23004	12.00.160
25620	12.00.160
11000.5	18.60.050

Civil Code

State Cite	*Code Section*
714(c)(3)	14.04.020
801.5(a)(1) - (2)	14.04.020
813	10.36.010
951(f)	20.00.070

Won't be Extracted



CA > ... > Los Angel... > Sec. 10.45.11. Coexisten...

- Sec. 10.45.10. [Reserved.]
- Sec. 10.45.11. Coexistence with Other Available Relief for Specific Deprivations of Protected Rights.
- Sec. 10.45.12. Severability.
- ARTICLE 21 LOCAL BUSINESS PREFERENCE PROGRAM
- ARTICLE 22 CITY CONTRACTORS'

SECTION HISTORY

Added by Ord. No. 181,520, Eff. 2-20-11.

Sec. 10.45.10. [Reserved.]

Sec. 10.45.11. Coexistence with Other Available Relief for Specific Deprivations of Protected Rights.

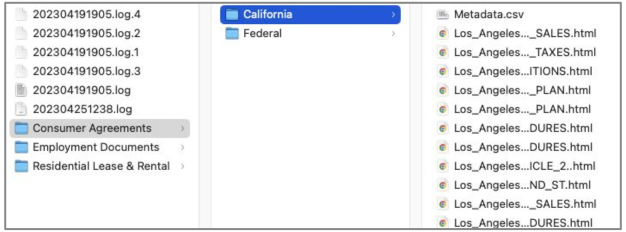
This Article shall not be construed to limit a person's right to bring legal action for violation of other laws.

SECTION HISTORY

Added by Ord. No. 181,520, Eff. 2-20-11.

Won't be Extracted

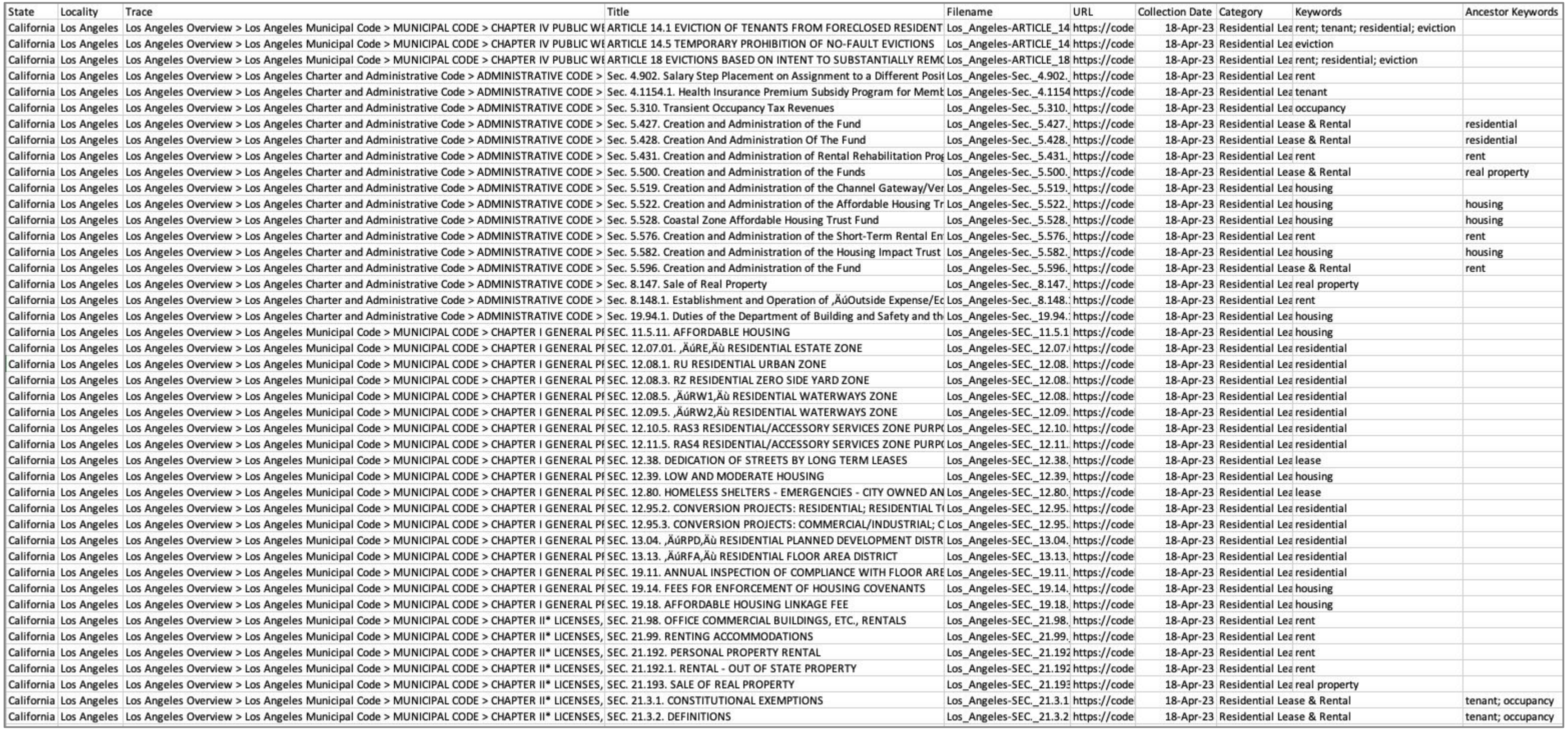
Crawling Result



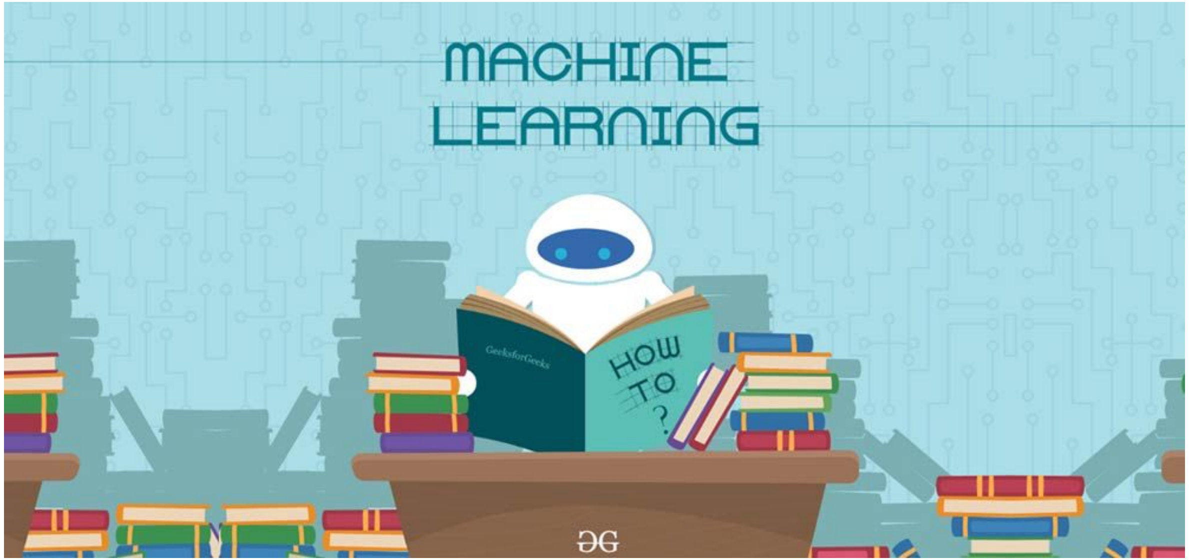
202304191905.log.4
202304191905.log.2
202304191905.log.1
202304191905.log.3
202304191905.log
202304251238.log
Consumer Agreements
Employment Documents
Residential Lease & Rental

California
Federal

Metadata.csv
Los_Angeles..._SALES.html
Los_Angeles..._TAXES.html
Los_Angeles...ITIONS.html
Los_Angeles..._PLAN.html
Los_Angeles..._PLAN.html
Los_Angeles...DURES.html
Los_Angeles...DURES.html
Los_Angeles...ICLE_2..html
Los_Angeles...ND_ST.html
Los_Angeles..._SALES.html
Los_Angeles...DURES.html

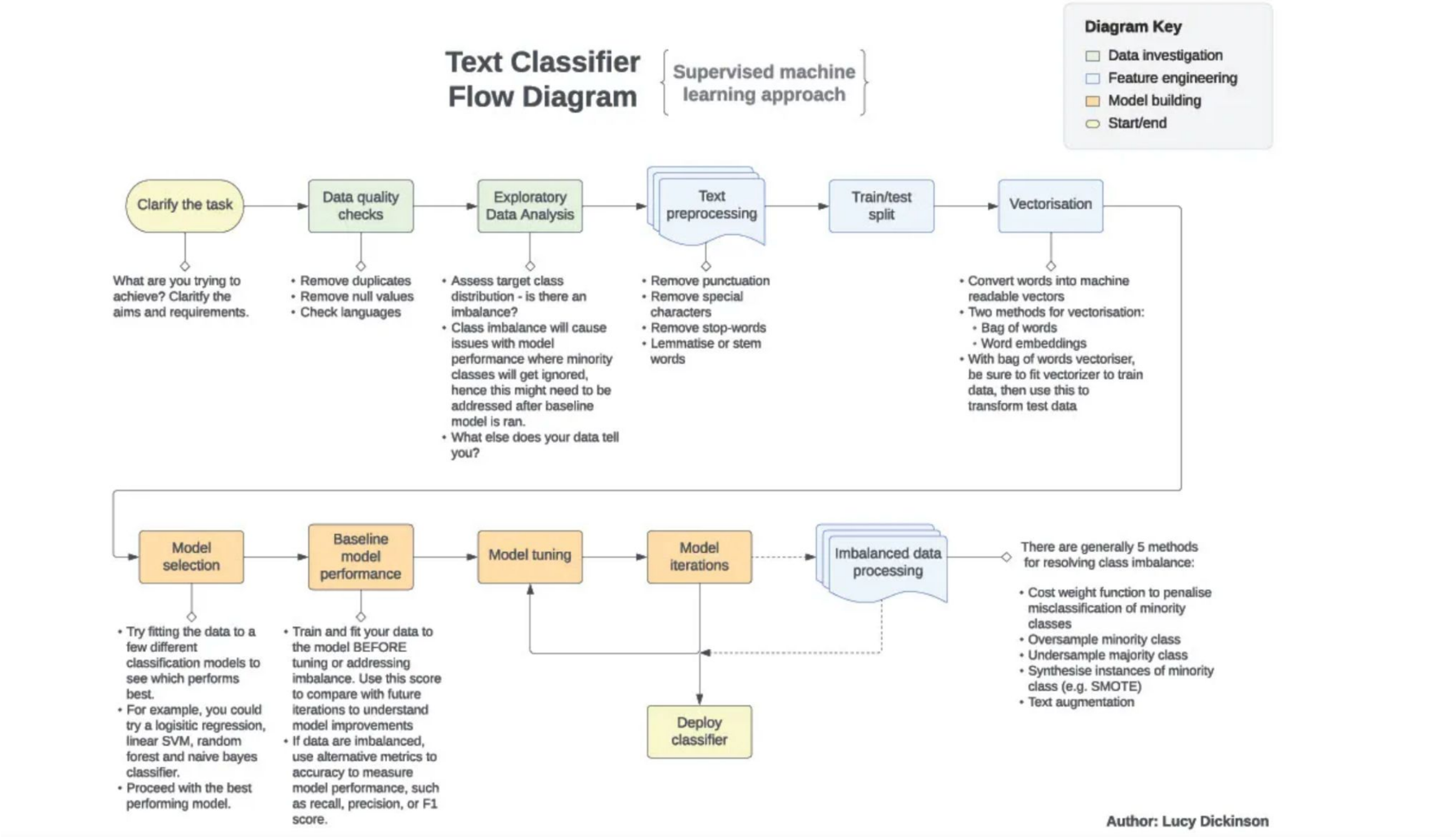


State	Locality	Trace	Title	Filename	URL	Collection Date	Category	Keywords	Ancestor Keywords
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER IV PUBLIC WE	ARTICLE 14.1 EVICTION OF TENANTS FROM FORECLOSED RESIDENT	Los_Angeles-ARTICLE_14	https://code	18-Apr-23	Residential Lea	rent; tenant; residential; eviction	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER IV PUBLIC WE	ARTICLE 14.5 TEMPORARY PROHIBITION OF NO-FAULT EVICTIONS	Los_Angeles-ARTICLE_14	https://code	18-Apr-23	Residential Lea	eviction	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER IV PUBLIC WE	ARTICLE 18 EVICTIONS BASED ON INTENT TO SUBSTANTIALLY REMC	Los_Angeles-ARTICLE_18	https://code	18-Apr-23	Residential Lea	rent; residential; eviction	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 4.902. Salary Step Placement on Assignment to a Different Posit	Los_Angeles-Sec._4.902._	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 4.1154.1. Health Insurance Premium Subsidy Program for Memb	Los_Angeles-Sec._4.1154	https://code	18-Apr-23	Residential Lea	tenant	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.310. Transient Occupancy Tax Revenues	Los_Angeles-Sec._5.310._	https://code	18-Apr-23	Residential Lea	occupancy	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.427. Creation and Administration of the Fund	Los_Angeles-Sec._5.427._	https://code	18-Apr-23	Residential Lease & Rental		residential
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.428. Creation And Administration Of The Fund	Los_Angeles-Sec._5.428._	https://code	18-Apr-23	Residential Lease & Rental		residential
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.431. Creation and Administration of Rental Rehabilitation Prog	Los_Angeles-Sec._5.431._	https://code	18-Apr-23	Residential Lea	rent	rent
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.500. Creation and Administration of the Funds	Los_Angeles-Sec._5.500._	https://code	18-Apr-23	Residential Lease & Rental		real property
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.519. Creation and Administration of the Channel Gateway/Ver	Los_Angeles-Sec._5.519._	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.522. Creation and Administration of the Affordable Housing Tr	Los_Angeles-Sec._5.522._	https://code	18-Apr-23	Residential Lea	housing	housing
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.528. Coastal Zone Affordable Housing Trust Fund	Los_Angeles-Sec._5.528._	https://code	18-Apr-23	Residential Lea	housing	housing
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.576. Creation and Administration of the Short-Term Rental En	Los_Angeles-Sec._5.576._	https://code	18-Apr-23	Residential Lea	rent	rent
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.582. Creation and Administration of the Housing Impact Trust	Los_Angeles-Sec._5.582._	https://code	18-Apr-23	Residential Lea	housing	housing
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 5.596. Creation and Administration of the Fund	Los_Angeles-Sec._5.596._	https://code	18-Apr-23	Residential Lease & Rental		rent
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 8.147. Sale of Real Property	Los_Angeles-Sec._8.147._	https://code	18-Apr-23	Residential Lea	real property	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 8.148.1. Establishment and Operation of ,ÄúOutside Expense/Ec	Los_Angeles-Sec._8.148.	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Charter and Administrative Code > ADMINISTRATIVE CODE >	Sec. 19.94.1. Duties of the Department of Building and Safety and th	Los_Angeles-Sec._19.94.	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 11.5.11. AFFORDABLE HOUSING	Los_Angeles-SEC._11.5.1	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.07.01. ,ÄúRE,Äù RESIDENTIAL ESTATE ZONE	Los_Angeles-SEC._12.07.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.08.1. RU RESIDENTIAL URBAN ZONE	Los_Angeles-SEC._12.08.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.08.3. RZ RESIDENTIAL ZERO SIDE YARD ZONE	Los_Angeles-SEC._12.08.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.08.5. ,ÄúRW1,Äù RESIDENTIAL WATERWAYS ZONE	Los_Angeles-SEC._12.08.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.09.5. ,ÄúRW2,Äù RESIDENTIAL WATERWAYS ZONE	Los_Angeles-SEC._12.09.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.10.5. RAS3 RESIDENTIAL/ACCESSORY SERVICES ZONE PURP(	Los_Angeles-SEC._12.10.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.11.5. RAS4 RESIDENTIAL/ACCESSORY SERVICES ZONE PURP(	Los_Angeles-SEC._12.11.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.38. DEDICATION OF STREETS BY LONG TERM LEASES	Los_Angeles-SEC._12.38.	https://code	18-Apr-23	Residential Lea	lease	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.39. LOW AND MODERATE HOUSING	Los_Angeles-SEC._12.39.	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.80. HOMELESS SHELTERS - EMERGENCIES - CITY OWNED AN	Los_Angeles-SEC._12.80.	https://code	18-Apr-23	Residential Lea	lease	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.95.2. CONVERSION PROJECTS: RESIDENTIAL; RESIDENTIAL T(	Los_Angeles-SEC._12.95.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 12.95.3. CONVERSION PROJECTS: COMMERCIAL/INDUSTRIAL; C	Los_Angeles-SEC._12.95.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 13.04. ,ÄúRPD,Äù RESIDENTIAL PLANNED DEVELOPMENT DISTR	Los_Angeles-SEC._13.04.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 13.13. ,ÄúRFA,Äù RESIDENTIAL FLOOR AREA DISTRICT	Los_Angeles-SEC._13.13.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 19.11. ANNUAL INSPECTION OF COMPLIANCE WITH FLOOR ARE	Los_Angeles-SEC._19.11.	https://code	18-Apr-23	Residential Lea	residential	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 19.14. FEES FOR ENFORCEMENT OF HOUSING COVENANTS	Los_Angeles-SEC._19.14.	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER I GENERAL PI	SEC. 19.18. AFFORDABLE HOUSING LINKAGE FEE	Los_Angeles-SEC._19.18.	https://code	18-Apr-23	Residential Lea	housing	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.98. OFFICE COMMERCIAL BUILDINGS, ETC., RENTALS	Los_Angeles-SEC._21.98.	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.99. RENTING ACCOMMODATIONS	Los_Angeles-SEC._21.99.	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.192. PERSONAL PROPERTY RENTAL	Los_Angeles-SEC._21.192	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.192.1. RENTAL - OUT OF STATE PROPERTY	Los_Angeles-SEC._21.192	https://code	18-Apr-23	Residential Lea	rent	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.193. SALE OF REAL PROPERTY	Los_Angeles-SEC._21.193	https://code	18-Apr-23	Residential Lea	real property	
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.3.1. CONSTITUTIONAL EXEMPTIONS	Los_Angeles-SEC._21.3.1	https://code	18-Apr-23	Residential Lease & Rental		tenant; occupancy
California	Los Angeles	Los Angeles Overview > Los Angeles Municipal Code > MUNICIPAL CODE > CHAPTER II* LICENSES,	SEC. 21.3.2. DEFINITIONS	Los_Angeles-SEC._21.3.2	https://code	18-Apr-23	Residential Lease & Rental		tenant; occupancy



Part 2: Text Classifier

Steps for Text Classification



Clarify the task

What are you trying to achieve? Clarify the aims and requirements.

Data quality checks

- Remove duplicates
- Remove null values
- Check languages

Exploratory Data Analysis

- Assess target class distribution - is there an imbalance?
- Class imbalance will cause issues with model performance where minority classes will get ignored, hence this might need to be addressed after baseline model is ran.
- What else does your data tell you?

Text preprocessing

- Remove punctuation
- Remove special characters
- Remove stop-words
- Lemmatise or stem words

Vectorisation

- Convert words into machine readable vectors
- Two methods for vectorisation:
  - Bag of words
  - Word embeddings
- With bag of words vectoriser, be sure to fit vectorizer to train data, then use this to transform test data

Model selection

- Try fitting the data to a few different classification models to see which performs best.
- For example, you could try a logisitic regression, linear SVM, random forest and naive bayes classifier.
- Proceed with the best performing model.

Baseline model performance

- Train and fit your data to the model BEFORE tuning or addressing imbalance. Use this score to compare with future iterations to understand model improvements
- If data are imbalanced, use alternative metrics to accuracy to measure model performance, such as recall, precision, or F1 score.

Imbalanced data processing

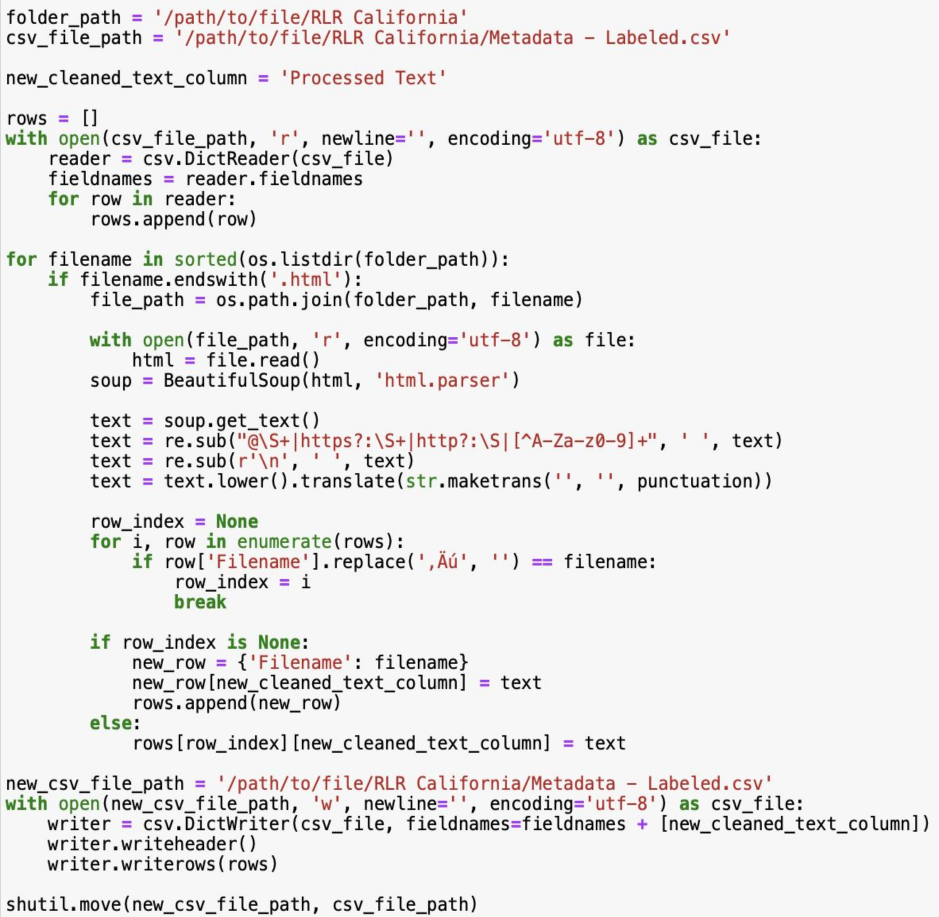
There are generally 5 methods for resolving class imbalance:

- Cost weight function to penalise misclassification of minority classes
- Oversample minority class
- Undersample majority class
- Synthesise instances of minority class (e.g. SMOTE)
- Text augmentation

Author: Lucy Dickinson

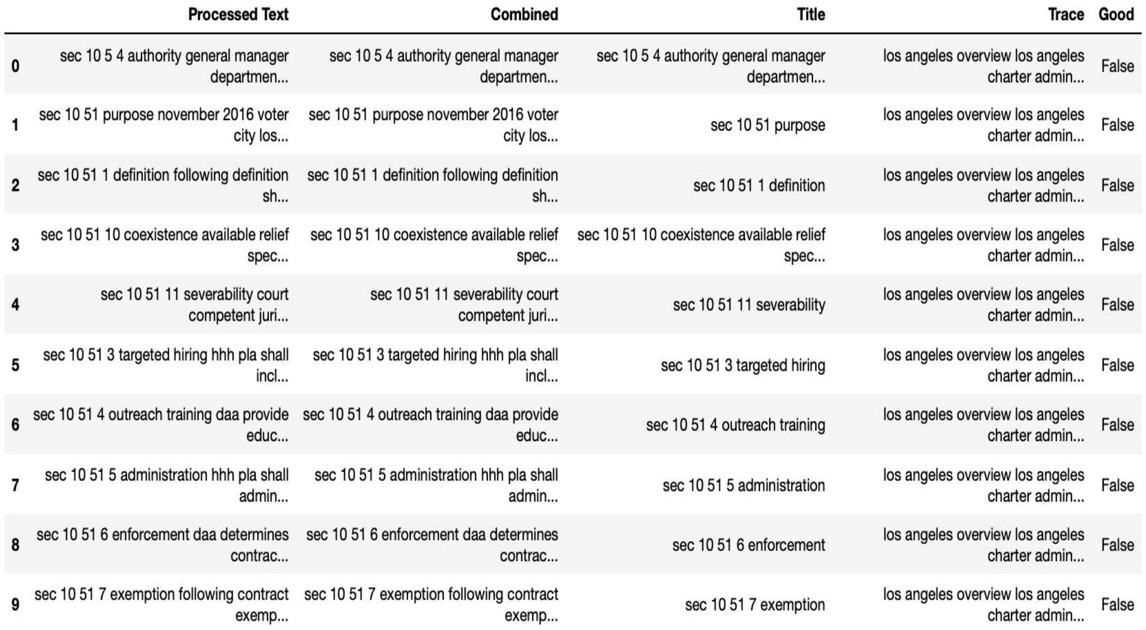
Data Preparation

- Cleaned HTML files by
  - Removing links,
  - Removing stop words,
  - Removing special characters, digits, and punctuation marks,
  - Removing leading and trailing whitespaces.
  - Converting text to lowercase
- Appended clean text to metadata csv by matching the filename in the specified folder with the ‘Filename’ column in the metadata csv.

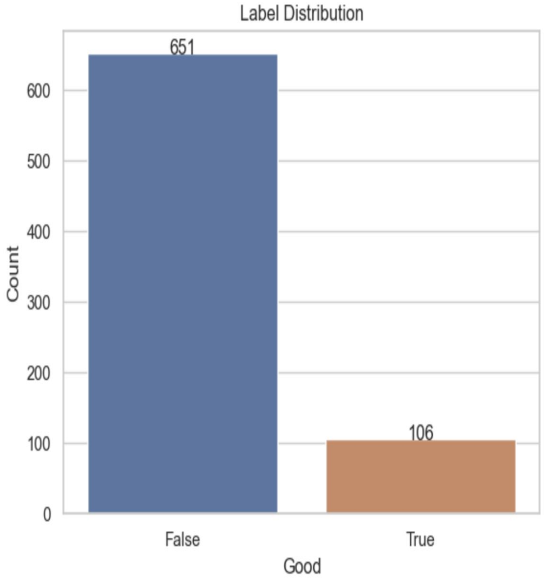


```python
folder_path = '/path/to/file/RLR California'
csv_file_path = '/path/to/file/RLR California/Metadata - Labeled.csv'

new_cleaned_text_column = 'Processed Text'

rows = []
with open(csv_file_path, 'r', newline='', encoding='utf-8') as csv_file:
    reader = csv.DictReader(csv_file)
    fieldnames = reader.fieldnames
    for row in reader:
        rows.append(row)

for filename in sorted(os.listdir(folder_path)):
    if filename.endswith('.html'):
        file_path = os.path.join(folder_path, filename)

        with open(file_path, 'r', encoding='utf-8') as file:
            html = file.read()
        soup = BeautifulSoup(html, 'html.parser')

        text = soup.get_text()
        text = re.sub("@\S+|https?:\S+|http?:\S|[^A-Za-z0-9]+", ' ', text)
        text = re.sub(r'\n', ' ', text)
        text = text.lower().translate(str.maketrans('', '', punctuation))

        row_index = None
        for i, row in enumerate(rows):
            if row['Filename'].replace(',Äú', '') == filename:
                row_index = i
                break

        if row_index is None:
            new_row = {'Filename': filename}
            new_row[new_cleaned_text_column] = text
            rows.append(new_row)
        else:
            rows[row_index][new_cleaned_text_column] = text

new_csv_file_path = '/path/to/file/RLR California/Metadata - Labeled.csv'
with open(new_csv_file_path, 'w', newline='', encoding='utf-8') as csv_file:
    writer = csv.DictWriter(csv_file, fieldnames=fieldnames + [new_cleaned_text_column])
    writer.writeheader()
    writer.writerows(rows)

shutil.move(new_csv_file_path, csv_file_path)
```

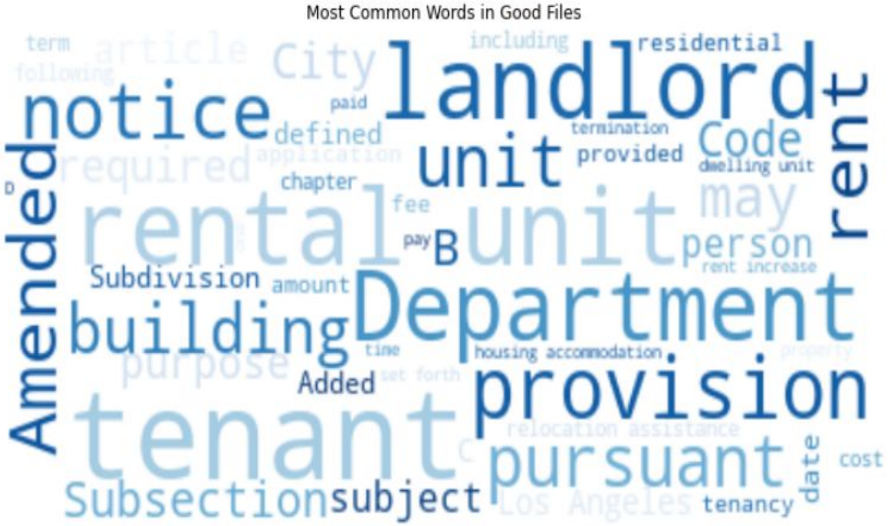
Dataset

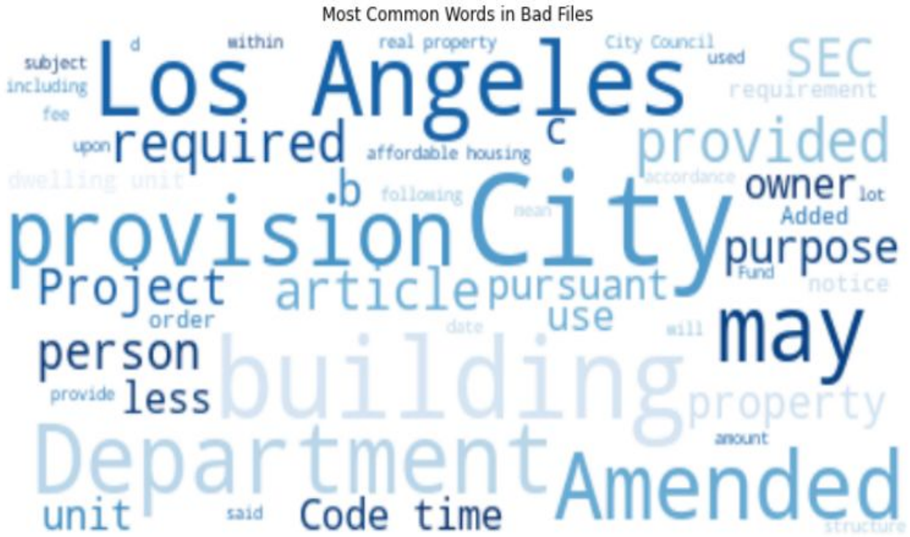


	Processed Text	Combined	Title	Trace	Good
0	sec 10 5 4 authority general manager departmen...	sec 10 5 4 authority general manager departmen...	sec 10 5 4 authority general manager departmen...	los angeles overview los angeles charter admin...	False
1	sec 10 51 purpose november 2016 voter city los...	sec 10 51 purpose november 2016 voter city los...	sec 10 51 purpose	los angeles overview los angeles charter admin...	False
2	sec 10 51 1 definition following definition sh...	sec 10 51 1 definition following definition sh...	sec 10 51 1 definition	los angeles overview los angeles charter admin...	False
3	sec 10 51 10 coexistence available relief spec...	sec 10 51 10 coexistence available relief spec...	sec 10 51 10 coexistence available relief spec...	los angeles overview los angeles charter admin...	False
4	sec 10 51 11 severability court competent juri...	sec 10 51 11 severability court competent juri...	sec 10 51 11 severability	los angeles overview los angeles charter admin...	False
5	sec 10 51 3 targeted hiring hhh pla shall incl...	sec 10 51 3 targeted hiring hhh pla shall incl...	sec 10 51 3 targeted hiring	los angeles overview los angeles charter admin...	False
6	sec 10 51 4 outreach training daa provide educ...	sec 10 51 4 outreach training daa provide educ...	sec 10 51 4 outreach training	los angeles overview los angeles charter admin...	False
7	sec 10 51 5 administration hhh pla shall admin...	sec 10 51 5 administration hhh pla shall admin...	sec 10 51 5 administration	los angeles overview los angeles charter admin...	False
8	sec 10 51 6 enforcement daa determines contrac...	sec 10 51 6 enforcement daa determines contrac...	sec 10 51 6 enforcement	los angeles overview los angeles charter admin...	False
9	sec 10 51 7 exemption following contract exemp...	sec 10 51 7 exemption following contract exemp...	sec 10 51 7 exemption	los angeles overview los angeles charter admin...	False

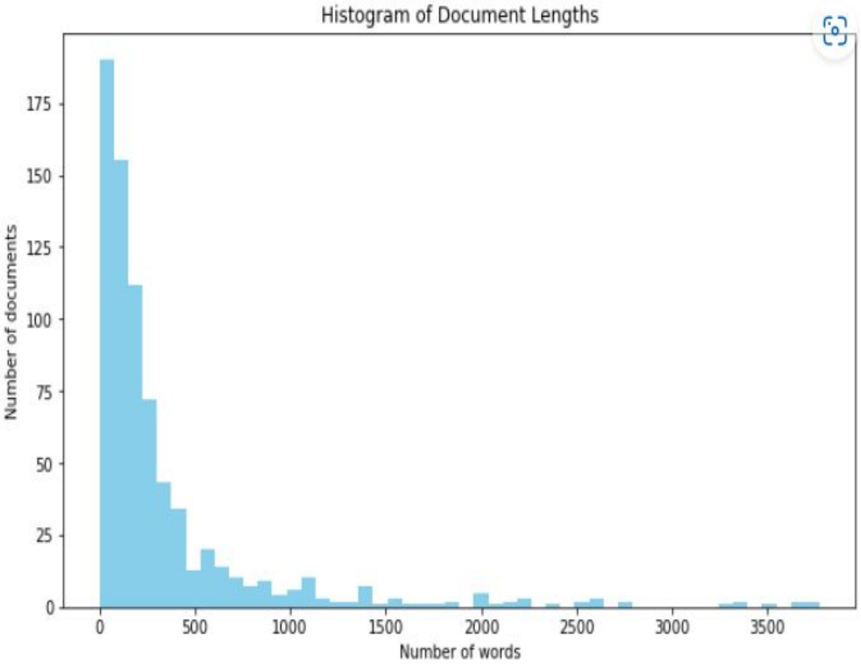


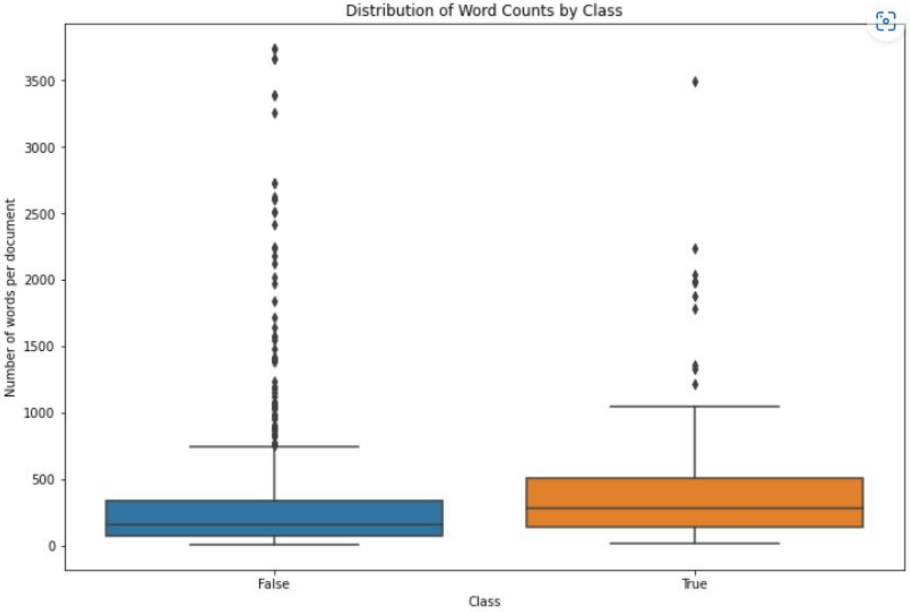
Most frequent words in Good files vs Bad files


Most Common Words in Good Files


Most Common Words in Bad Files

Distribution of Word Count

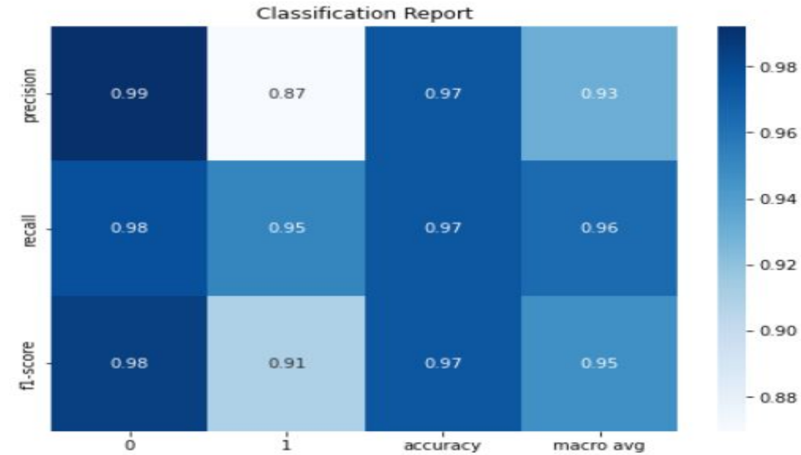




Machine Learning Algorithms

- Logistic Regression
- XGBoost
- KNN
- RNN
- LSTM

Logistic Regression Model Results

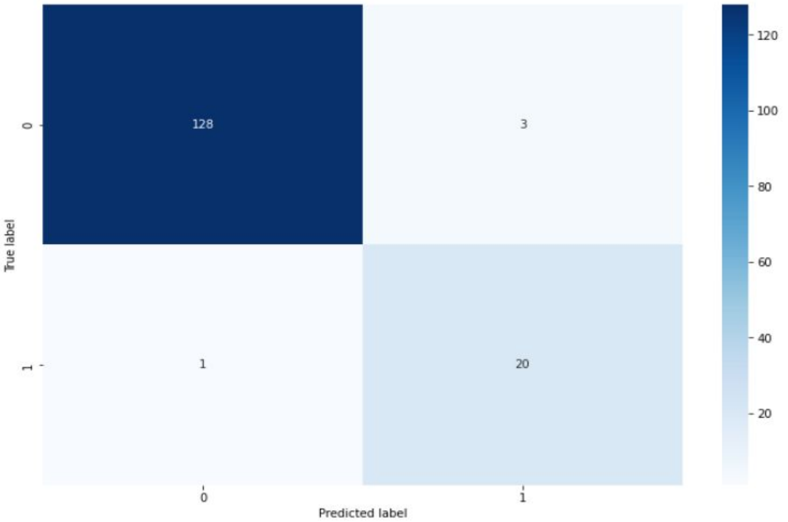


Classification Report

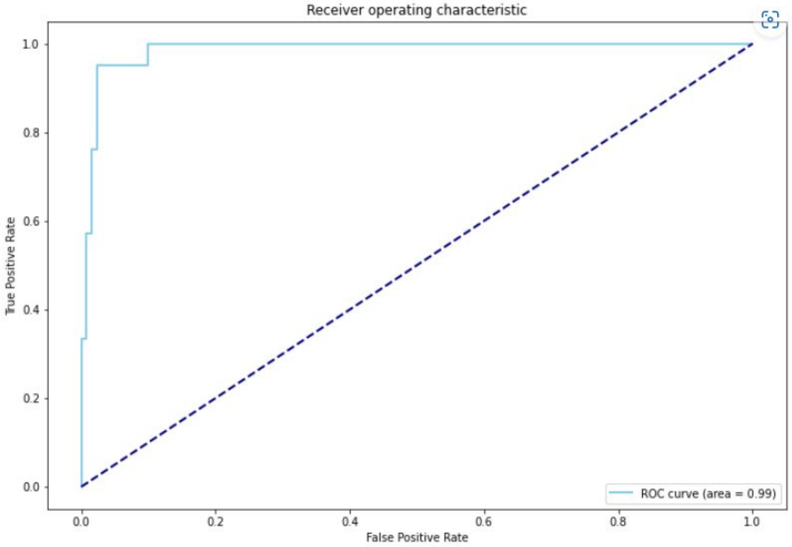
	0	1	accuracy	macro avg
precision	0.99	0.87	0.97	0.93
recall	0.98	0.95	0.97	0.96
f1-score	0.98	0.91	0.97	0.95

```
param_grid_lr = {'C': [0.01, 0.1, 1, 10, 100],
                 'penalty': ['l2'],
                 'max_iter': [100, 500, 10000]}
```

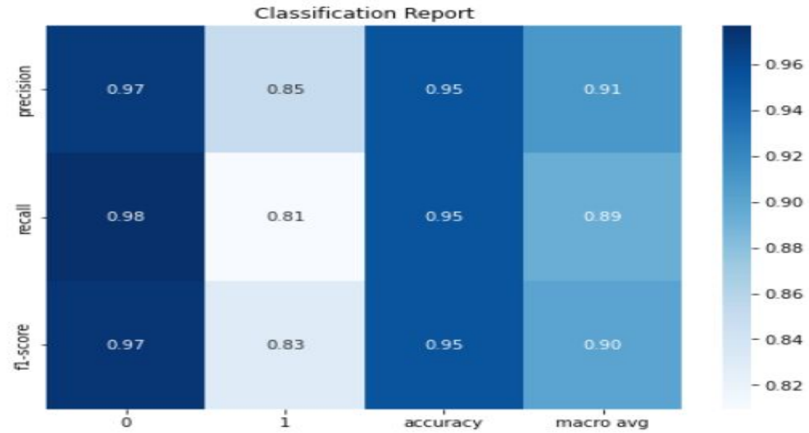
Best hyperparameters:
{'C': 100, 'max_iter': 100, 'penalty': 'l2'}



True label \ Predicted label	0	1
0	128	3
1	1	20

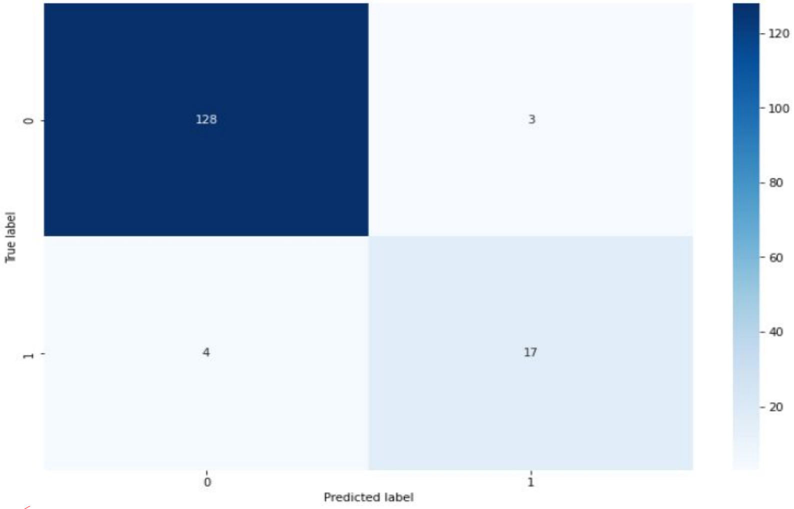


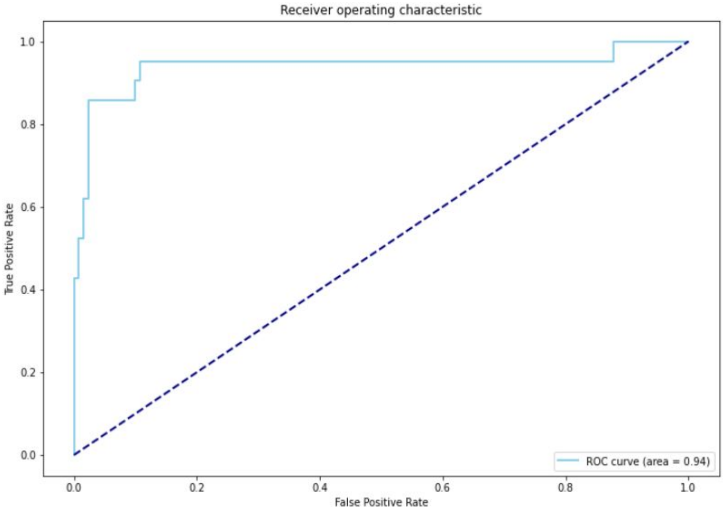
XGBoost Model Results



param_grid_xgb = {'learning_rate': [0.1, 0.01],
'max_depth': [3, 5, 7],
'n_estimators': [50, 100, 200]}

Best hyperparameters:
{'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100}





KNN (K-Nearest Neighbors)

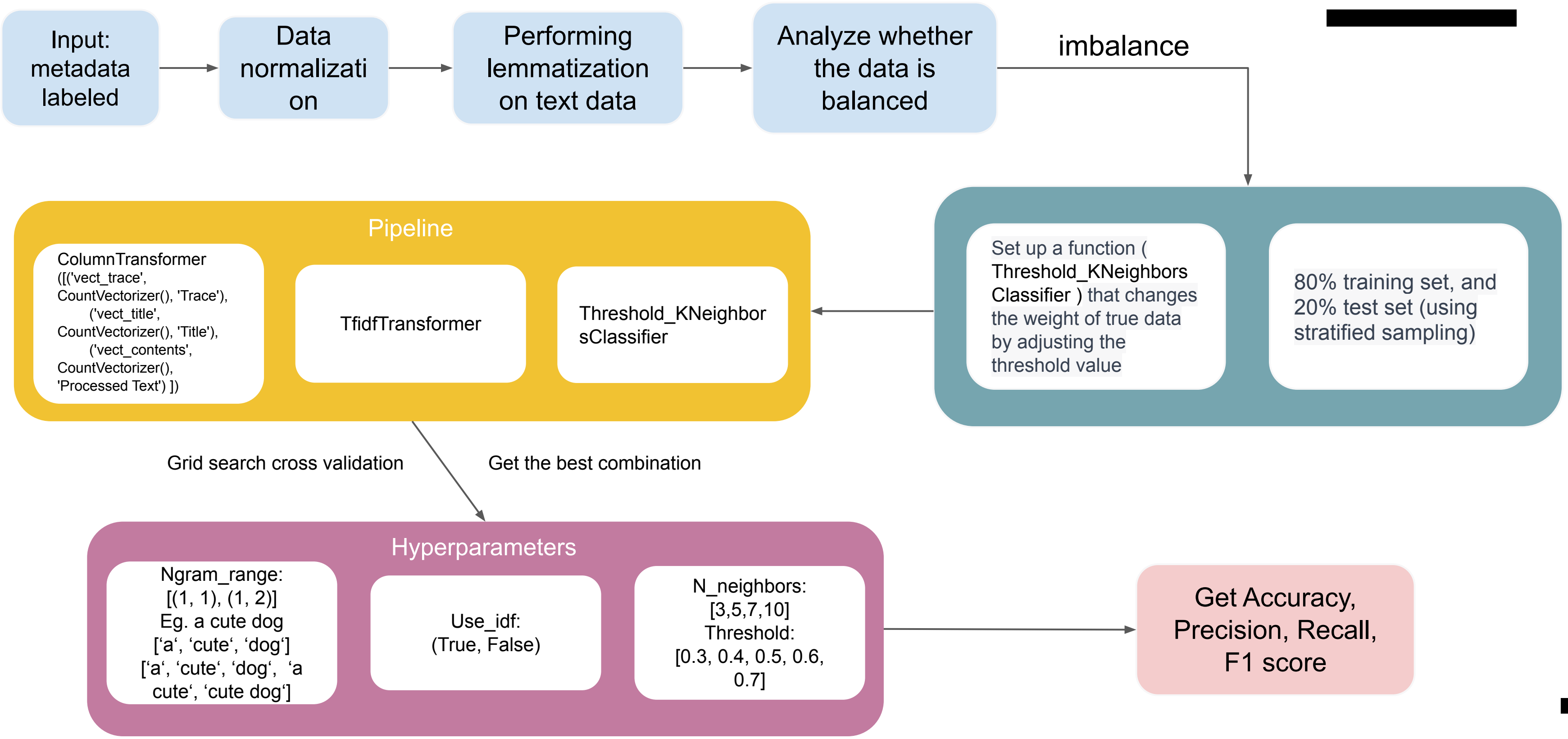
Working principle: Calculate the distance between the data point to be classified or predicted and each point in the training dataset. Select the K nearest training data points to the data point to be predicted. The data point to be predicted is assigned to the class with the most neighboring points.

Advantage：The KNN algorithm is very intuitive, easy to understand, and implement. It has strong adaptability and can adjust the complexity of the model by choosing an appropriate K value. It is a non-parametric method, meaning it does not make any assumptions about the data. This allows KNN to adapt to various types of data distributions, especially when the data distribution is unknown or non-linear.

Disadvantage: KNN requires calculating the distance between the test sample and all training samples during computation, which may lead to high computational complexity. It is sensitive to noise and outliers.

Research objectives and expected outcomes: The objective is to find a good model that can predict text relevance, with the expected outcome of the model's accuracy being above 85%.

Simple Flow Chart



Input: metadata labeled

Data normalization

Performing lemmatization on text data

Analyze whether the data is balanced

imbalance

Set up a function (Threshold_KNeighbors Classifier) that changes the weight of true data by adjusting the threshold value

80% training set, and 20% test set (using stratified sampling)

Pipeline

ColumnTransformer ([('vect_trace', CountVectorizer(), 'Trace'), ('vect_title', CountVectorizer(), 'Title'), ('vect_contents', CountVectorizer(), 'Processed Text')])

TfidfTransformer

Threshold_KNeighborsClassifier

Grid search cross validation

Get the best combination

Hyperparameters

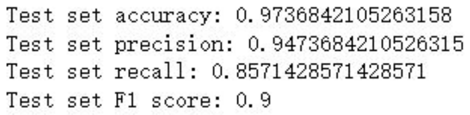
Ngram_range: [(1, 1), (1, 2)] Eg. a cute dog ['a', 'cute', 'dog'] ['a', 'cute', 'dog', 'a cute', 'cute dog']

Use_idf: (True, False)

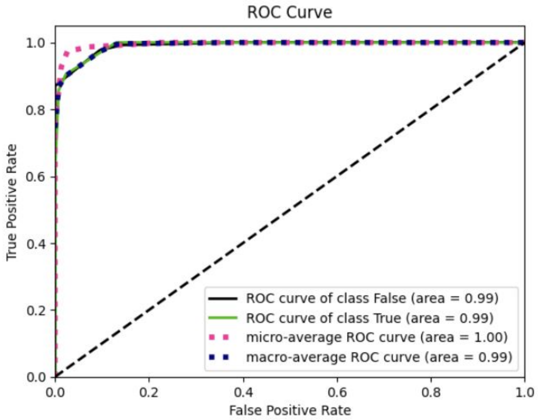
N_neighbors: [3,5,7,10] Threshold: [0.3, 0.4, 0.5, 0.6, 0.7]

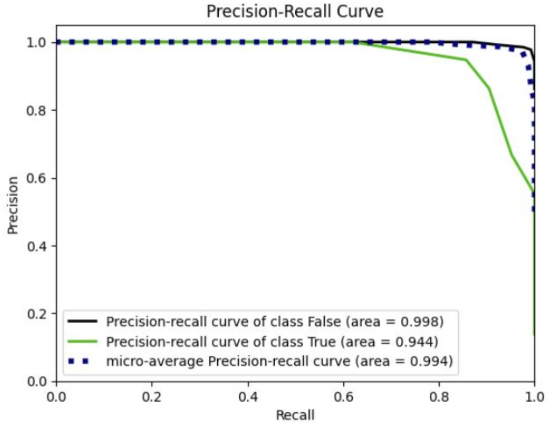
Get Accuracy, Precision, Recall, F1 score

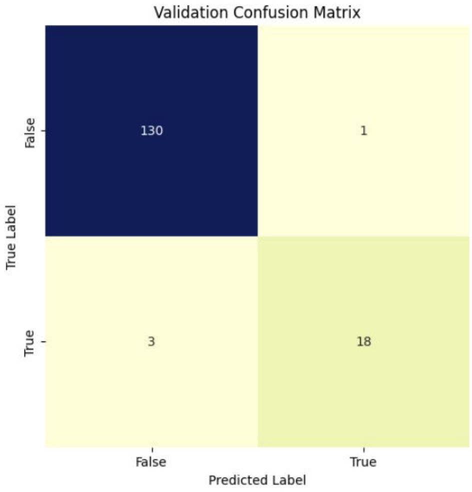
KNN Result

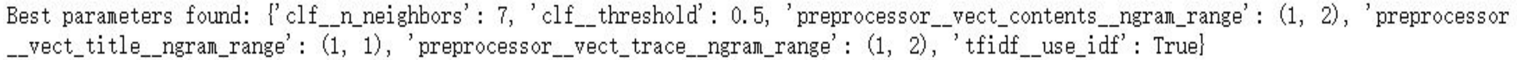


```
Test set accuracy: 0.9736842105263158
Test set precision: 0.9473684210526315
Test set recall: 0.8571428571428571
Test set F1 score: 0.9
```



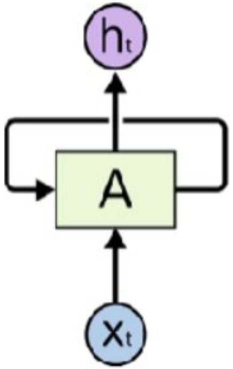






```
Best parameters found: {'clf__n_neighbors': 7, 'clf__threshold': 0.5, 'preprocessor__vect_contents__ngram_range': (1, 2), 'preprocessor
__vect_title__ngram_range': (1, 1), 'preprocessor__vect_trace__ngram_range': (1, 2), 'tfidf__use_idf': True}
```

RNN (Recurrent Neural Network)-GRU



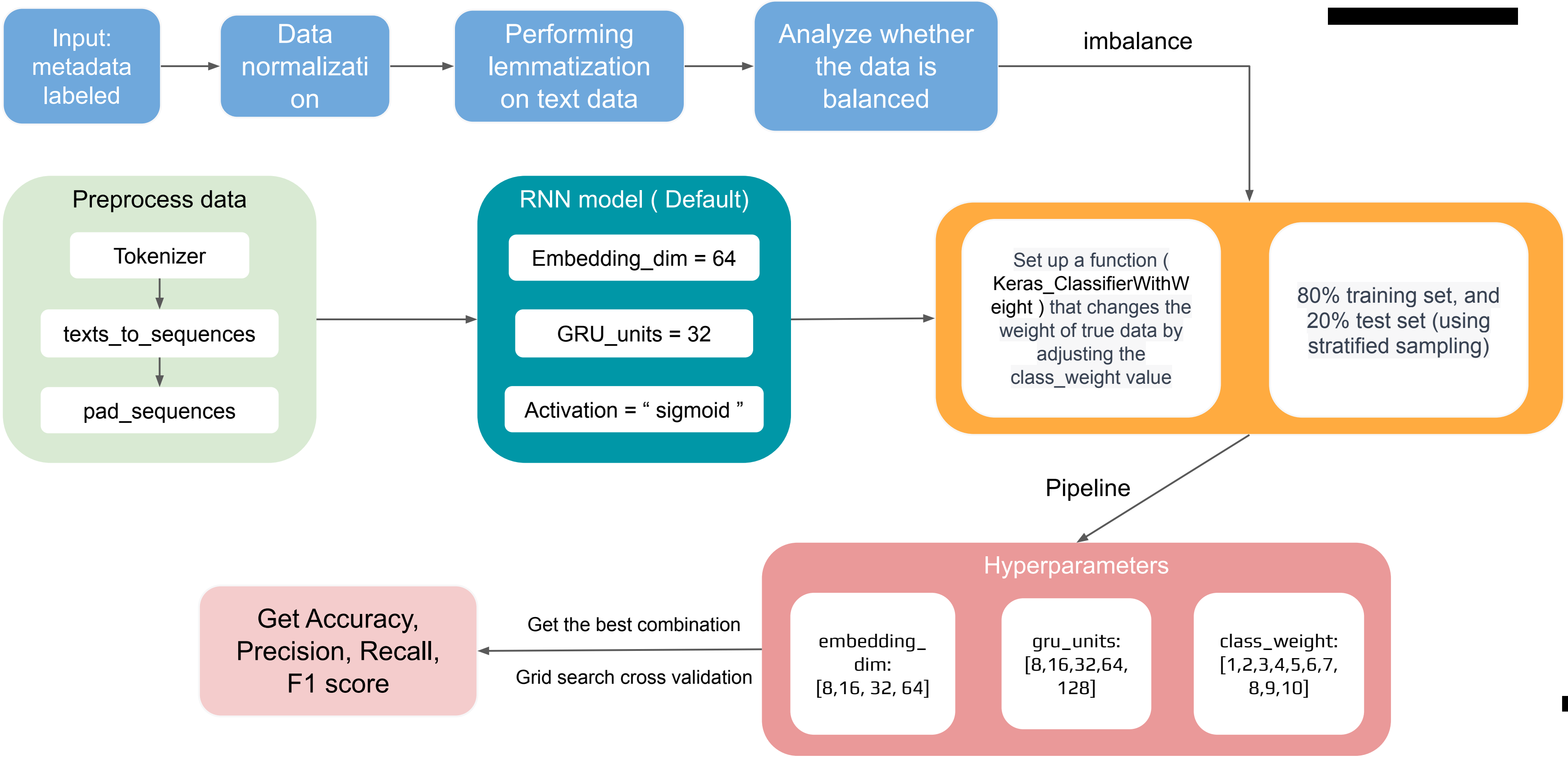
Working principle: In RNNs, the connections between neurons form directed cycles, allowing the network to have persistent internal states, thus capturing temporal dependencies in sequences. Gated Recurrent Units (GRUs) are an improved RNN unit that introduces gating mechanisms to address the vanishing/exploding gradient problem faced by traditional RNNs. GRUs introduce two gates, the update gate, and the reset gate. Through the control of these gates, GRUs can selectively retain or ignore historical information, thereby learning more effective temporal dependencies in long sequences.

Advantage：GRU can solve the vanishing gradient problem. Computational efficiency high. Good performance on NLP.

Disadvantage: For more complex tasks, GRUs may not have the same expressive power as LSTMs, as LSTMs have more gates and parameters. Parallelism limitations

Research objectives and expected outcomes: The objective is to find a good model that can predict text relevance, with the expected outcome of the model's accuracy being above 85%.

Simple Flow Chart



Input: metadata labeled

Data normalization

Performing lemmatization on text data

Analyze whether the data is balanced

imbalance

Set up a function (Keras_ClassifierWithWeight) that changes the weight of true data by adjusting the class_weight value

80% training set, and 20% test set (using stratified sampling)

Preprocess data

- Tokenizer
- texts_to_sequences
- pad_sequences

RNN model (Default)

- Embedding_dim = 64
- GRU_units = 32
- Activation = " sigmoid "

Pipeline

Hyperparameters

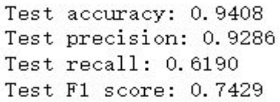
- embedding_dim: [8,16, 32, 64]
- gru_units: [8,16,32,64, 128]
- class_weight: [1,2,3,4,5,6,7, 8,9,10]

Get the best combination

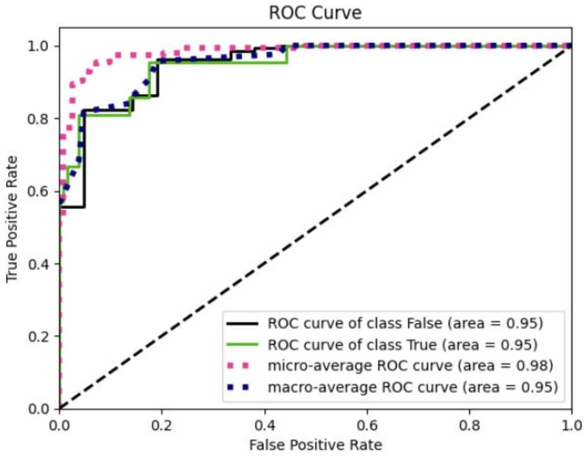
Grid search cross validation

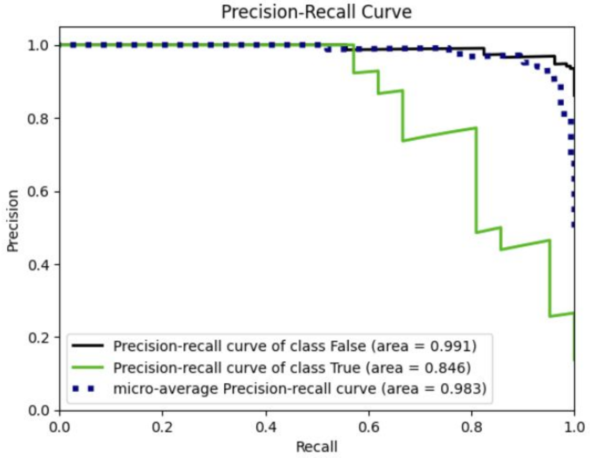
Get Accuracy, Precision, Recall, F1 score

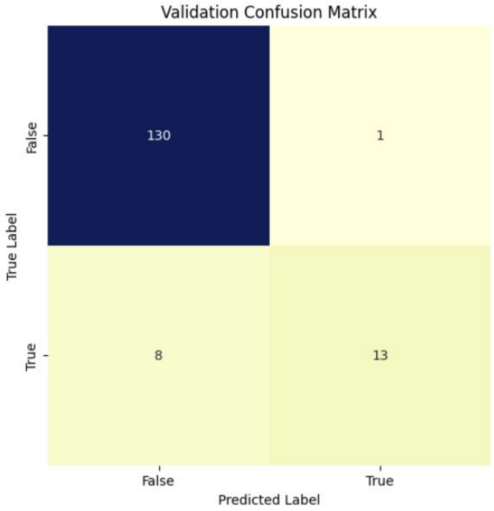
RNN-GRU Result

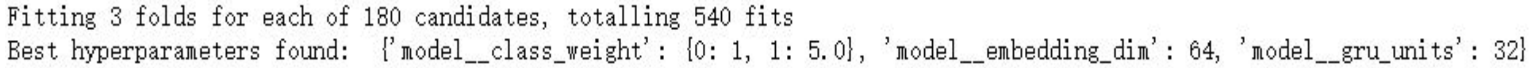


```
Test accuracy: 0.9408
Test precision: 0.9286
Test recall: 0.6190
Test F1 score: 0.7429
```



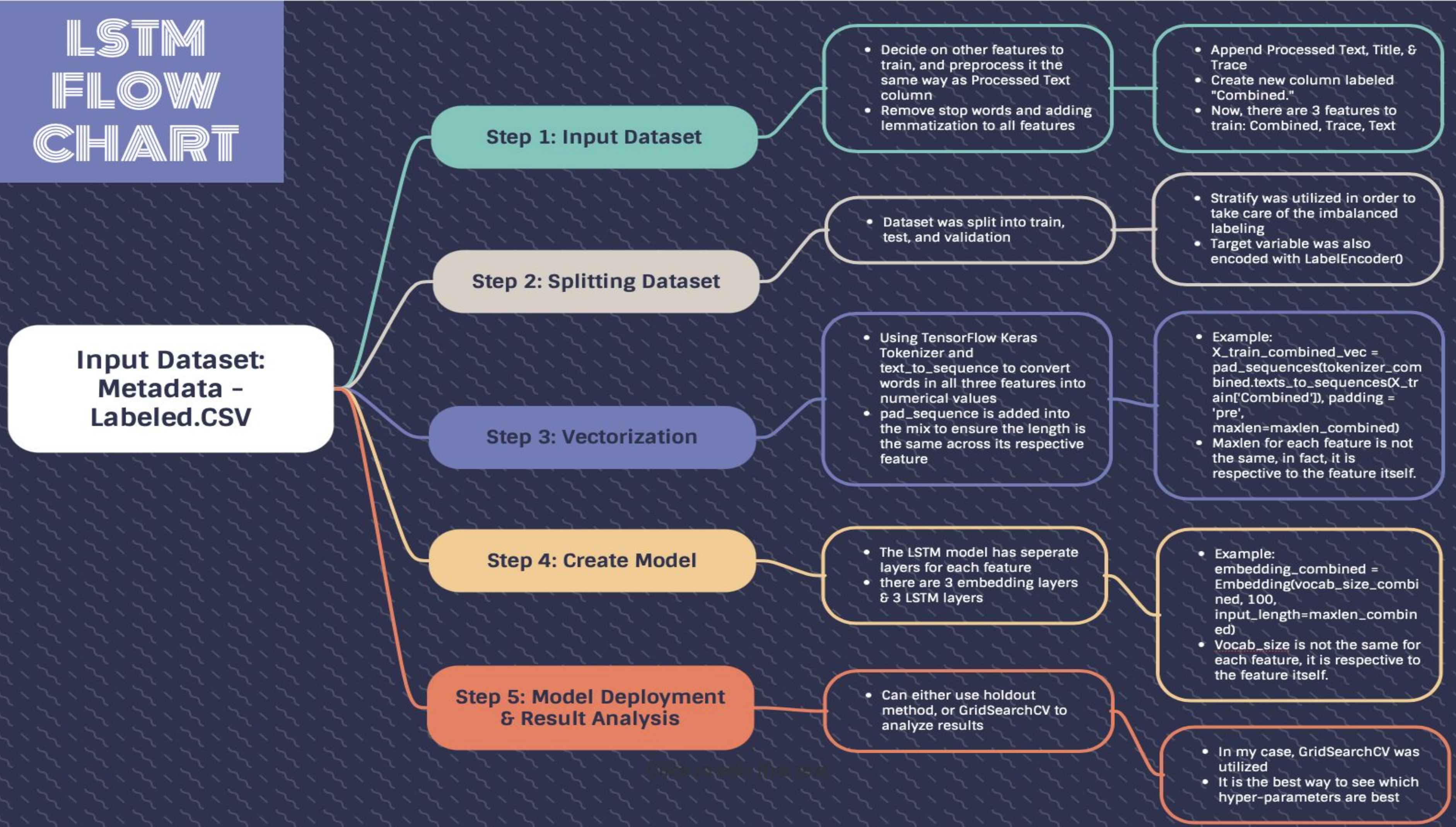






```
Fitting 3 folds for each of 180 candidates, totalling 540 fits
Best hyperparameters found:  {'model__class_weight': {0: 1, 1: 5.0}, 'model__embedding_dim': 64, 'model__gru_units': 32}
```

LSTM FLOW CHART



Input Dataset: Metadata - Labeled.CSV

Step 1: Input Dataset

- Decide on other features to train, and preprocess it the same way as Processed Text column
- Remove stop words and adding lemmatization to all features

- Append Processed Text, Title, & Trace
- Create new column labeled "Combined."
- Now, there are 3 features to train: Combined, Trace, Text

Step 2: Splitting Dataset

- Dataset was split into train, test, and validation

- Stratify was utilized in order to take care of the imbalanced labeling
- Target variable was also encoded with LabelEncoder()

Step 3: Vectorization

- Using TensorFlow Keras Tokenizer and text_to_sequence to convert words in all three features into numerical values
- pad_sequence is added into the mix to ensure the length is the same across its respective feature

- Example: X_train_combined_vec = pad_sequences(tokenizer_combined.texts_to_sequences(X_train['Combined']), padding = 'pre', maxlen=maxlen_combined)
- Maxlen for each feature is not the same, in fact, it is respective to the feature itself.

Step 4: Create Model

- The LSTM model has seperate layers for each feature
- there are 3 embedding layers & 3 LSTM layers

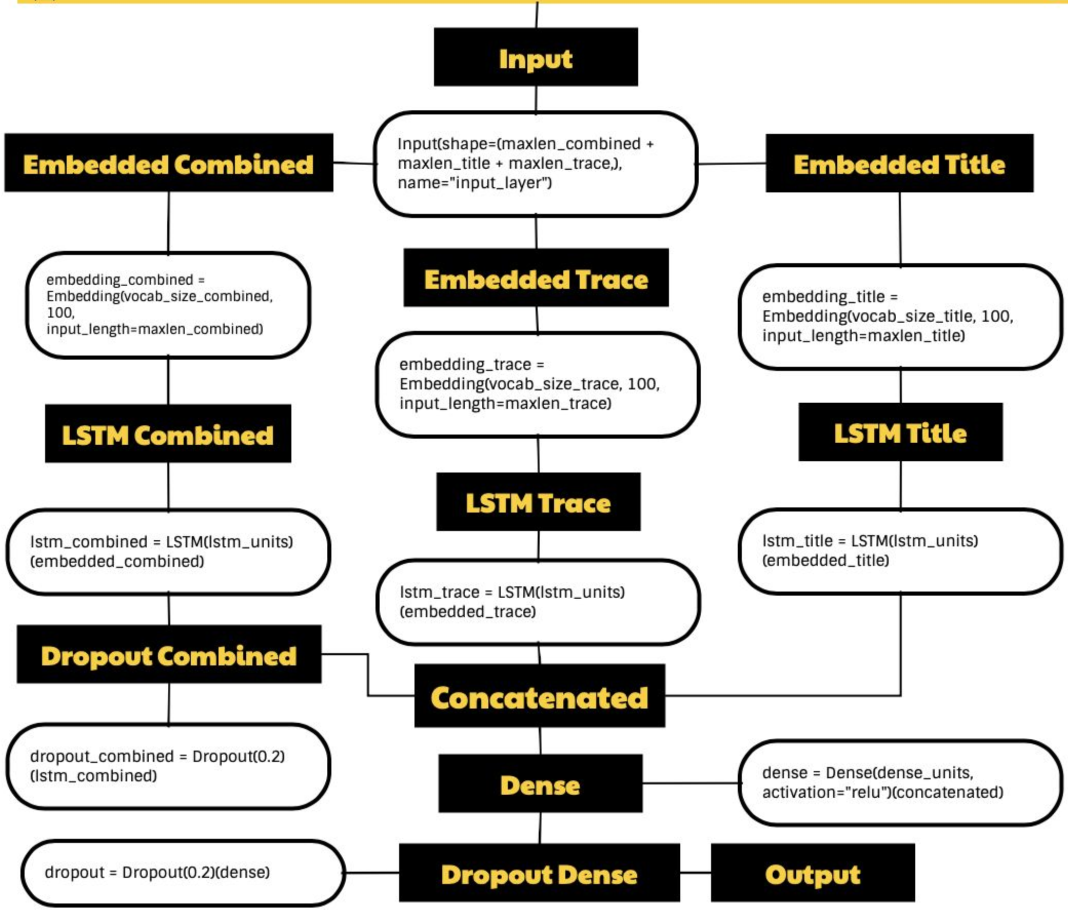
- Example: embedding_combined = Embedding(vocab_size_combined, 100, input_length=maxlen_combined)
- Vocab_size is not the same for each feature, it is respective to the feature itself.

Step 5: Model Deployment & Result Analysis

- Can either use holdout method, or GridSearchCV to analyze results

- In my case, GridSearchCV was utilized
- It is the best way to see which hyper-parameters are best

LSTM MODEL ARCHITECTURE



Input

```
Input(shape=(maxlen_combined + maxlen_title + maxlen_trace,), name="input_layer")
```

Embedded Combined

```
embedding_combined = Embedding(vocab_size_combined, 100, input_length=maxlen_combined)
```

LSTM Combined

```
lstm_combined = LSTM(lstm_units)(embedded_combined)
```

Dropout Combined

```
dropout_combined = Dropout(0.2)(lstm_combined)
```

Embedded Trace

```
embedding_trace = Embedding(vocab_size_trace, 100, input_length=maxlen_trace)
```

LSTM Trace

```
lstm_trace = LSTM(lstm_units)(embedded_trace)
```

Embedded Title

```
embedding_title = Embedding(vocab_size_title, 100, input_length=maxlen_title)
```

LSTM Title

```
lstm_title = LSTM(lstm_units)(embedded_title)
```

Concatenated

Dense

```
dense = Dense(dense_units, activation="relu")(concatenated)
```

Dropout Dense

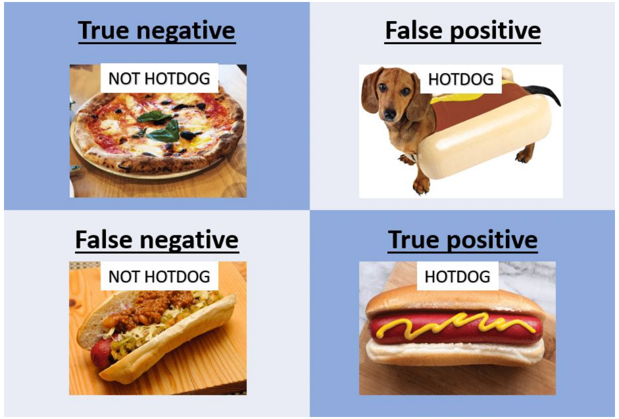
```
dropout = Dropout(0.2)(dense)
```

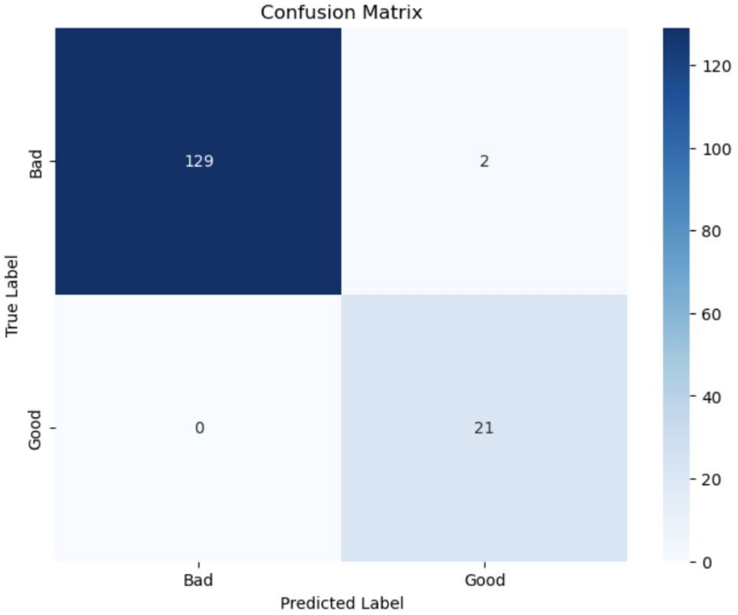
Output

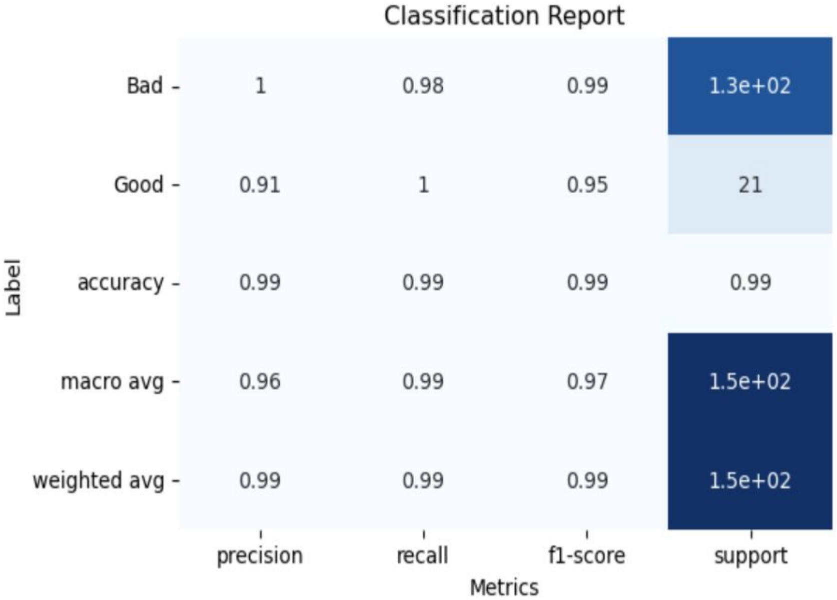
FYI: maxlen and vocab_size is respective to feature.

```
(lstm_units=128,
dense_units=64)
```

LSTM Model Results (Test Set)



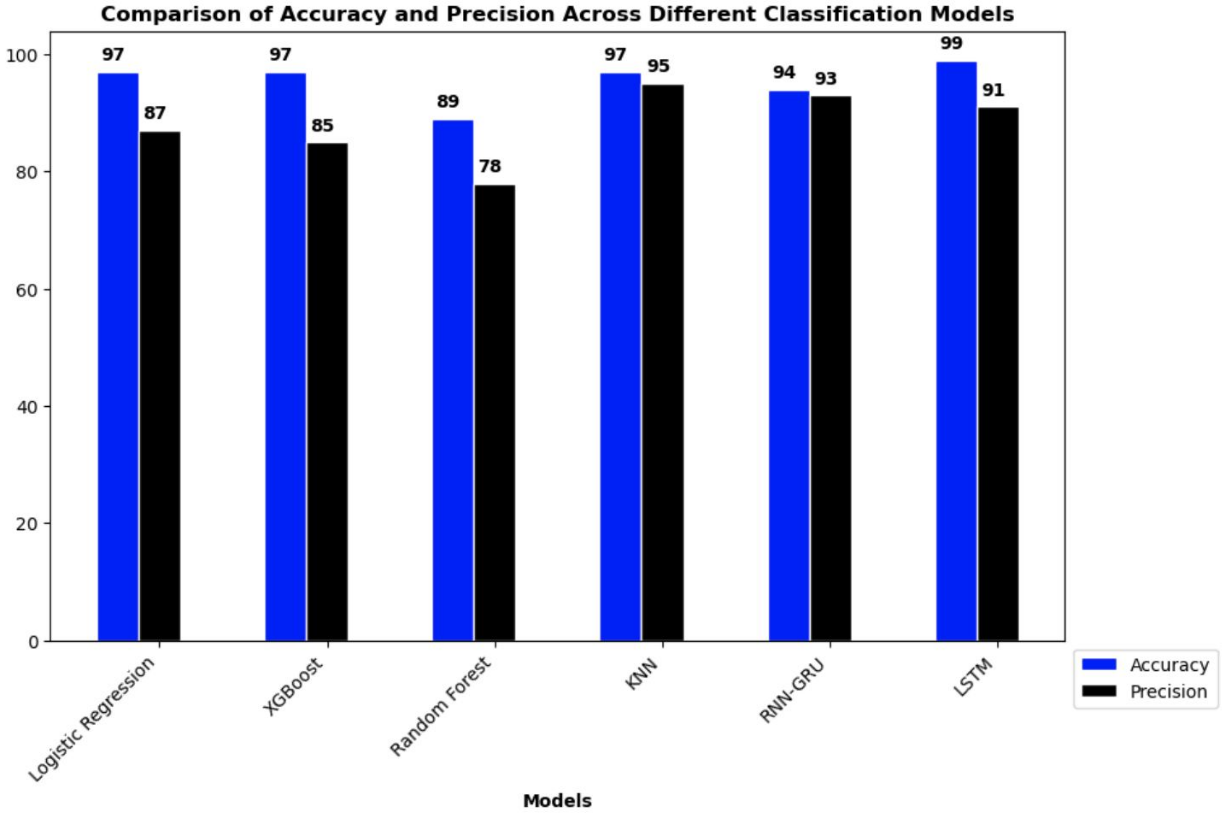



Classification Report

Label	precision	recall	f1-score	support
Bad	1	0.98	0.99	1.3e+02
Good	0.91	1	0.95	21
accuracy	0.99	0.99	0.99	0.99
macro avg	0.96	0.99	0.97	1.5e+02
weighted avg	0.99	0.99	0.99	1.5e+02

Metrics

Recap & Conclusion: Model Results



THANK
YOU!

References:

https://builtin.com/machine-learning/lemmatization

https://towardsdatascience.com/step-by-step-basics-text-classifier-e666c6bac52b

https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

https://www.kaggle.com/code/abhishek/approaching-almost-any-nlp-problem-on-kaggle