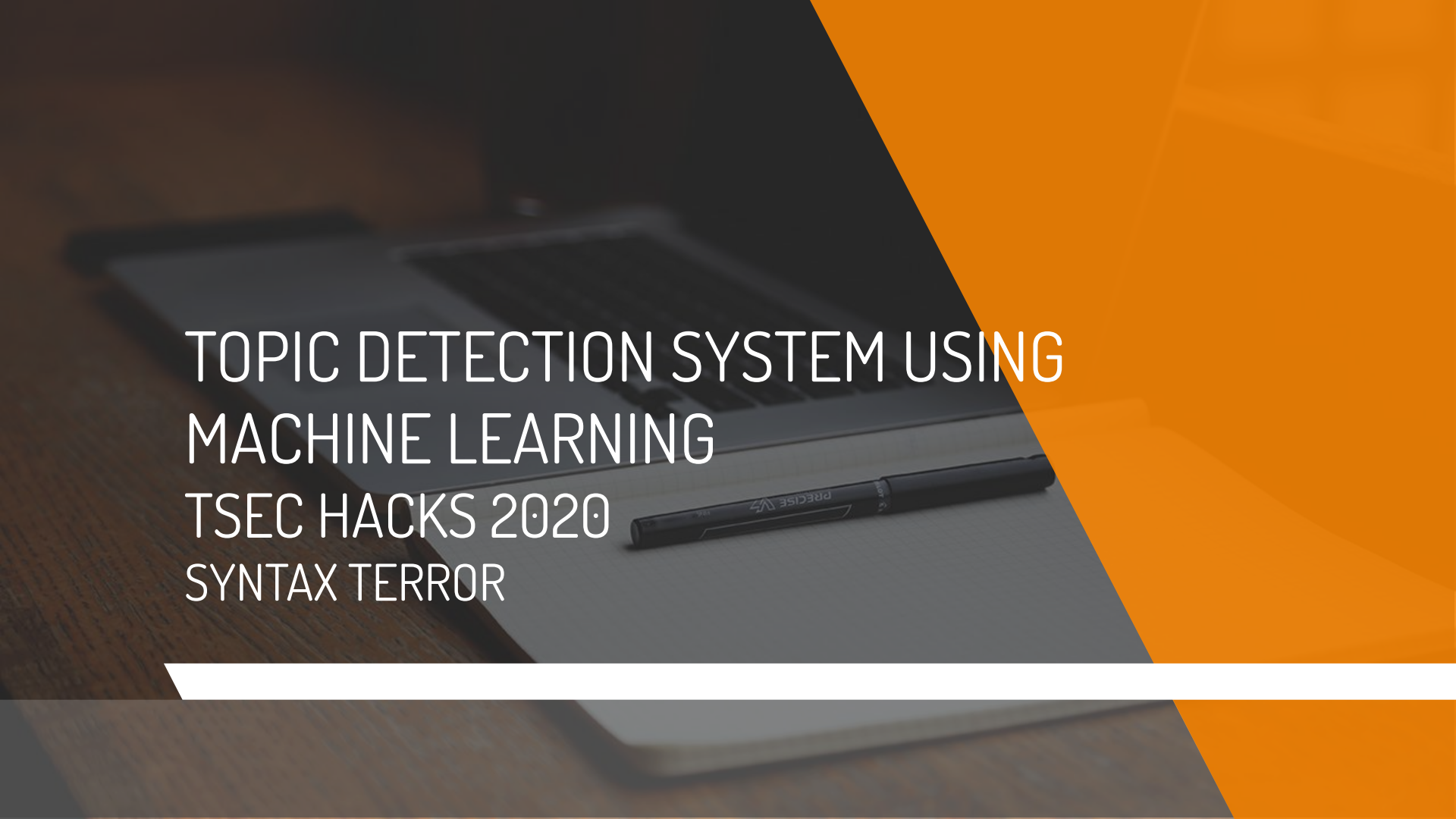
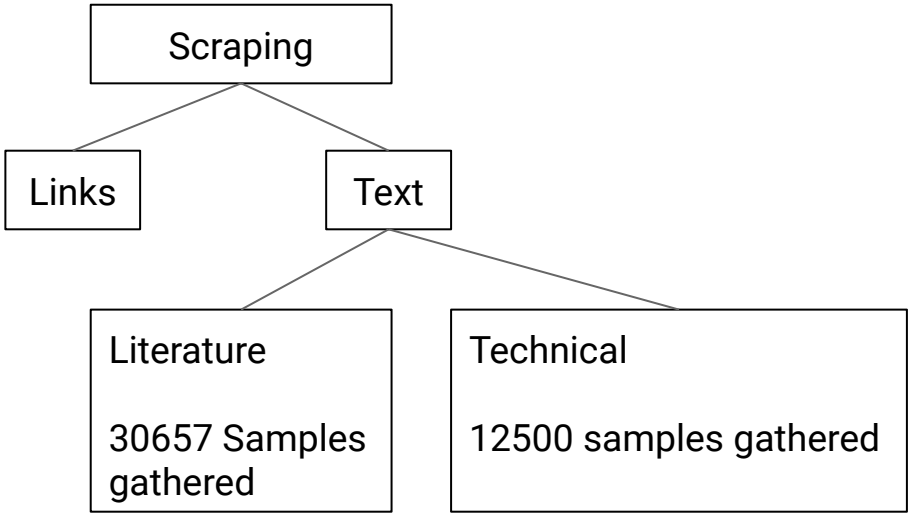


TOPIC DETECTION SYSTEM USING MACHINE LEARNING

TSEC HACKS 2020

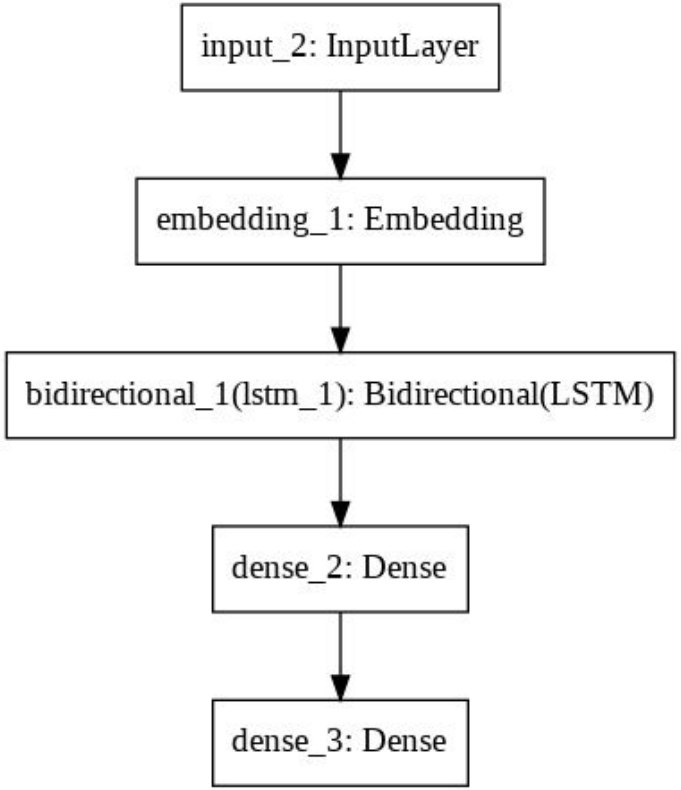
SYNTAX TERROR

Data Collection

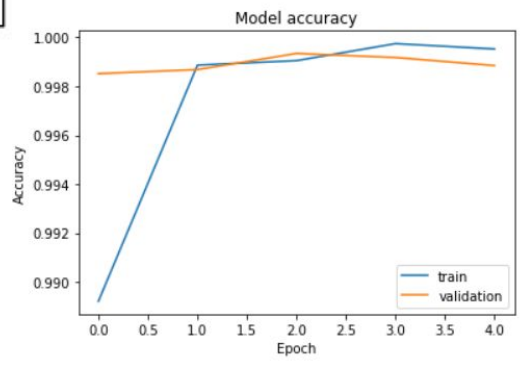


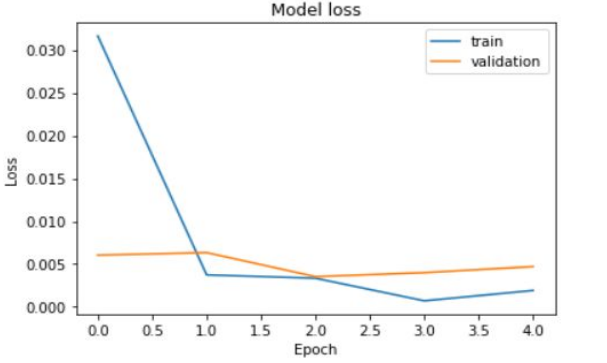
Requests: for getting the html content of the page
BeautifulSoup: for traversing the html
html5lib: for parsing the html content

Model Architecture

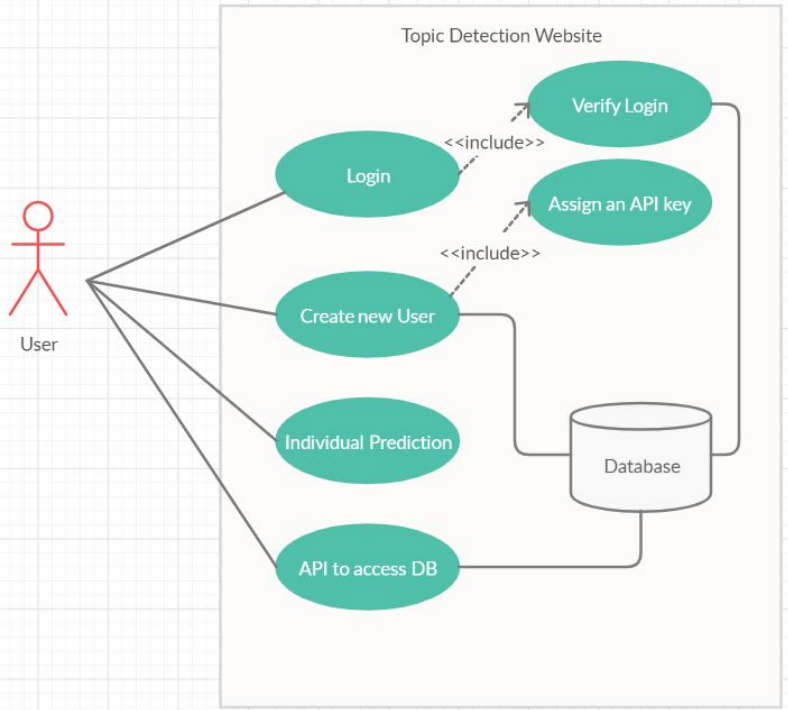


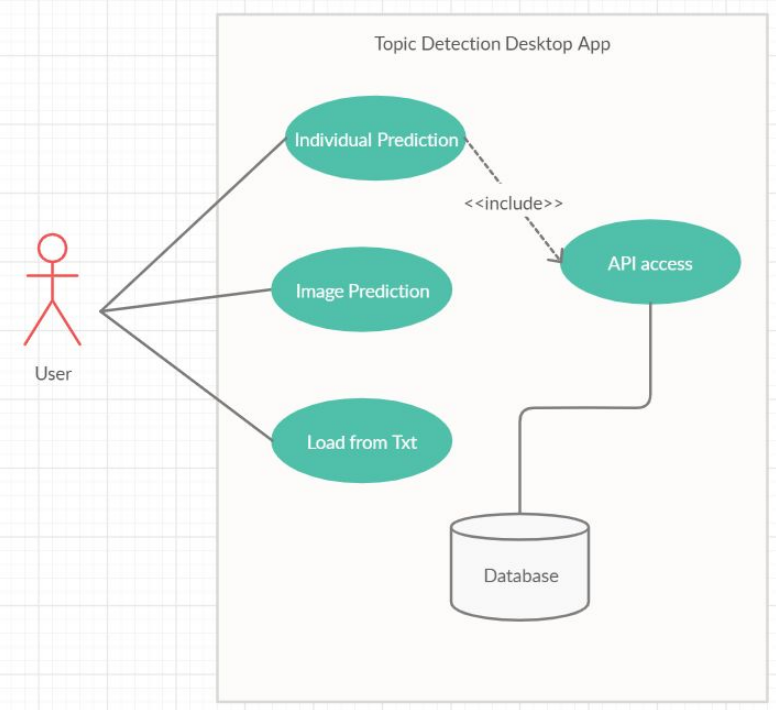
- Used SMOTE and pandas to oversample and merge the 2 datasets
- Due to the simplicity of the dataset the validation accuracy: 99.8%, precision: 99.7% and recall: 99.9%





User Interface





Future Developments

- Using Convolution across images to detect characters/words for handwritten images for on the go classification
- Search engine optimization
- Real time context searching like google lens

Technological Stack