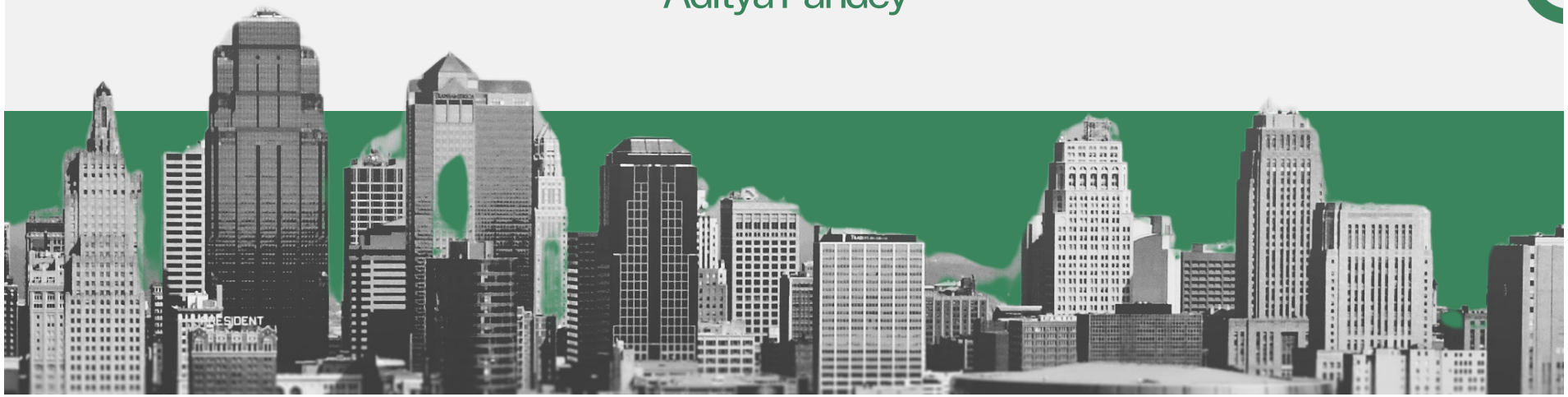
"Leveraging Machine Learning for Efficient Complaint Classification in the Banking Sector"

Aditya Pandey



Data Source and Description

- "Data from CFPB (JSON format)"
- "21,072 complaints"
- "Categories: Bank Accounts, Credit Cards, Theft, etc."
- "Unstructured text"

Data Cleaning Process

01.

"Removed punctuation (re.sub)"

02.

"Filtered numbers (re.sub)"

03.

"Lemmatization with spaCy"

04.

“Standardized text, reduce noise”

Model Selection and Methods

- Model Used: "DistilBERT model"
- "Why: 40% smaller, 97% BERT accuracy"
- "Classified into 5 categories"
- "Fine-tuned on complaint data"

Performance and Results

- "87% accuracy"
- "Metrics: Precision, Recall, F1"
- "Challenge: 'Others' vs. 'Theft' overlap"

Goals, Value, and Conclusion

"Goals: Automate classification"
"Value: Saves ~230 hours, ensures compliance"
"Conclusion: Effective, with room to refine"

Link: Hosted Model

Thank you very much!