Presentation Script: OCR-Driven Product Classification from Thai Receipts

Slide 1: Title Slide (30 seconds)

Good [morning/afternoon], everyone. I'm [Name] from the University of Phayao, Thailand. Today, I'll be presenting our research on "Leveraging OCR-Driven Information Extraction for Accurate Product Type Classification from Thai Receipt Data: An Ensemble Learning Approach."

Slide 2: Outline (30 seconds)

Here's an overview of what we'll cover today. We'll start with an introduction to the problem, then discuss our methodology, data collection and preprocessing, classification models, results, and finish with a discussion and conclusion.

Slide 3: Introduction (1 minute)

Receipt data plays a crucial role in family expense management, offering detailed insights into spending patterns. However, extracting and classifying product information from Thai receipts presents unique challenges due to the complexity of the Thai script and the absence of word boundaries. Our research aims to develop an accurate product type classification system using ensemble learning techniques on OCR-extracted data from Thai receipts.

Slide 4: Methodology (1 minute)

Our approach involves several key steps, as illustrated in this framework diagram. We start with OCR to extract text from receipt images, followed by preprocessing and tokenization of the Thai text. We then apply feature extraction techniques before feeding the data into various classification models, including both base classifiers and ensemble methods.

Slide 5: Data Collection and Preprocessing (2 minutes)

We collected 1,305 receipt images from 100 volunteers in Phayao Province, Thailand. From these, we extracted 5,087 product names across five categories. We used Tesseract OCR, specifically trained for Thai, to extract the text. The preprocessing stage involved text normalization and tokenization using PyThaiNLP, a library designed for Thai natural language processing. For feature extraction, we applied TF-IDF with considerations specific to Thai language characteristics, such as the use of character-level n-grams to capture sub-word information.

Slide 6: Classification Models (2 minutes)

We experimented with two groups of classification models. Our base classifiers included K-Nearest Neighbors, Logistic Regression, Naive Bayes, Decision Trees, and Support Vector Machines. Each of these offers unique strengths in handling text data. For ensemble methods, we implemented Random Forest, AdaBoost, Extra Trees, Bagging, and Majority Voting. These ensemble techniques were chosen to leverage the strengths of multiple models while mitigating individual weaknesses.

Slide 7: Results - Base Classifiers (2 minutes)

This slide shows the performance of our base classification algorithms. Among these, Support Vector Machines (SVM) demonstrated the most balanced and robust performance across all categories, achieving a weighted average F1-score of 92.26% and an accuracy of 92.51%. It performed particularly well in the Drinks category with 98.78% precision and 65.32% recall. Other classifiers showed varying strengths across different categories, highlighting the complexity of the classification task.

Slide 8: Results - Ensemble Methods (2 minutes)

Moving on to our ensemble methods, we see significant improvements. Majority Voting emerged as the top performer, achieving a weighted average F1-score of 91.74% and an accuracy of 91.92%. It showed balanced performance across all categories. The Extra Trees algorithm recorded the highest overall accuracy at 92.05%, excelling particularly in the Food and Health & Beauty categories. These results demonstrate the power of ensemble techniques in handling the complexities of Thai product name classification.

Slide 9: Discussion (2 minutes)

Our findings clearly show the superiority of ensemble methods over individual classifiers for this task. The success of Majority Voting and Extra Trees algorithms suggests their ability to handle diverse product names and potential OCR errors effectively. However, we also encountered challenges, including OCR errors in mixed Thai-English text, issues arising from faded or damaged receipts, and class imbalance in our dataset. These challenges highlight areas for potential improvement in future iterations of our system.

Slide 10: Conclusion and Future Work (1 minute, 30 seconds)

In conclusion, our research demonstrates the significant potential of leveraging OCR-driven information extraction and ensemble learning for accurate product type classification from Thai receipts. The success of our ensemble methods, particularly Majority Voting and Extra Trees, sets a strong foundation for future advancements in this field. Moving forward, we plan to explore advanced feature extraction techniques, methods to handle OCR errors more effectively, and the application of deep learning approaches. We also see potential in developing language-specific models that consider the unique characteristics of Thai in OCR and text processing.

Slide 11: Thank You (30 seconds)

Thank you for your attention. I'd be happy to answer any questions you may have about our research.