UNIVERSITI TEKNOLOGI MARA

INTEGRATING AN ENHANCED HIDDEN MARKOV MODEL WITH FEATURE SUBSTITUTION FOR SHORT-TEXT E-COMMERCE PRODUCT CLASSIFICATION

NORSYELA BINTI MUHAMMAD NOOR MATHIVANAN

PhD

November 2025

ABSTRACT

Automatic product classification based on short-text data is essential for managing the vast information generated on e-commerce platforms. As a subset of text classification, product classification assigns items to predefined categories. Within this domain, product title classification is challenging due to text brevity, inconsistent terminology, and noisy information. Accurate classification is crucial for enhancing the online shopping experience by improving organization, retrieval, and recommendation. Despite the rapid growth of data on e-commerce platforms, existing classification models continue to face challenges with accuracy and efficiency. These difficulties arise from short and inconsistent product descriptions, varying terminology across sellers and noisy or redundant information that complicates classification. This research addresses these challenges by leveraging Hidden Markov Models (HMMs), which capture sequential data through probabilistic modeling of hidden states and transitions. The study improves HMM performance in short-text product title classification through two key innovations which are feature substitution using Latent Dirichlet Allocation (LDA) and weighted parameter estimation in HMM. Traditional HMMs often degrade with complex data, while rigid emission parameters limit adaptability. Feature substitution reduces sparsity and redundancy in text, whereas weighted parameter estimation increases flexibility in parameter learning. To overcome these limitations, weighted parameter estimation is integrated into the HMM framework. This enhancement addresses the rigidity of emission parameters and improves adaptability to complex and diverse product data, which increases the model's flexibility and overall performance. This study proposes three methods to enhance HMM performance in product title classification. The propose method I (FS-LDA) substitutes semantically related features within the same product category to reduce sparsity and strengthen representation. The proposed method II focuses on adjusting emission parameters based on information from the training data, which allows the model to adapt more effectively to complex and imbalanced distributions. The proposed method III integrates FS-LDA with weighted parameter estimation in a unified framework, combining the advantages of both techniques to achieve improved classification outcomes. Experiments across five e-commerce datasets show significant improvements over traditional HMMs. Method III achieved over 95% accuracy in binary classification and Fl-Scores above 90%. In multi class scenarios, Fl-Scores exceeded 70%, demonstrating robustness across categories. The proposed methods also outperformed Naive Bayes and Support Vector Machines, particularly in short-text multiclass tasks. Beyond e-commerce, validation in spam filtering and occupational data mining confirmed substantial gains in accuracy and Fl-Scores. In conclusion, the proposed method III is the most effective approach to enhance HMM-based product title classification. The scope of this research is explicitly focused on short-text product title classification in e-commerce, contributing to the body of knowledge on text classification using HMMs. This study also provides a foundation for developing user-friendly tools, libraries, and documentation to facilitate the integration of enhanced HMM-based classifiers into existing e-commerce systems.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful

Alhamdulillah, all praises to Allah Subhanahu Wa Ta'ala for giving me the strength and courage to complete this thesis.

Dear self, thank you for being brave enough to complete your study. Thank you for being the voice that kept telling me not to give up.

Special appreciation goes to my supervisor, Prof. Dr. Nor Azura Md. Ghani. She is the reason I was able to successfully complete my study. Thank you for the eye-opening experiences throughout this long and challenging journey. Not to be forgotten, my co-supervisor, Datuk Prof. Dr. Roziah Mohd Janor, for her support and guidance.

From the beginning, my family has played a vital role in my success. Thank you to Mak, Abah, Abang Leh, Kak Sha, Adah, Aiman, Adik, Yusuf, and Hana for your tremendous support and prayers. My sincere appreciation to my husband, Nur Hakim Lim for lending me a shoulder to cry on. Though your weirdness, stubbornness, and occasional ignorance often test my patience, it is your love that melts my heart the most and gives me the strength to keep moving forward.

I dedicate this PhD to my late beloved mother-in-law, whose love and kindness profoundly touched my life during the three wonderful years I was fortunate to know her.

Finally, I would like to thank all those involved in the making of Dr. Norsyela Muhammad Noor Mathivanan. May Allah Subhanahu Wa Ta'ala return your kindness.

And to my little one who came into my life at the end of this journey, you are the sweetest surprise and my greatest blessing.

Thank you from the bottom of my heart.

TABLE OF CONTENTS

		Page
CONFIRMATION BY PANEL OF EXAMINERS		**ii**
AUTHOR'S DECLARATION		**iii**
ABSTRACT		**iv**
ACKNOWLEDGEMENT		**v**
TABLE OF CONTENTS		**vi**
LIST OF TABLES		**xi**
LIST OF FIGURES		**xiv**
LIST OF ABBREVIATIONS		**xvii**
CHAPTER 1 INTRODUCTION		**1**
1.1	Introduction	1
1.2	Research Motivation	1
1.3	Problem Statement	3
1.4	Research Questions	4
1.5	Research Objectives	4
1.6	Research Significance	5
1.7	Research Scope and Limitation	6
1.8	Thesis Outline	7
CHAPTER 2 LITERATURE REVIEW		**8**
2.1	Introduction	8
2.2	Text Classification Overview	8
	2.2.1 Process Involvement	10
	2.2.1.1 Data Extraction	10
	2.2.1.2 Data Pre-processing	11
	2.2.1.3 Feature Extraction	15
	2.2.1.4 Feature Selection	16
	2.2.1.5 Data Partition	18
	2.2.1.6 Data Classification	19

CHAPTER 1
INTRODUCTION

1.1 Introduction

Chapter One describes the basis of conducting this research. This chapter aims to emphasize the importance of this research through research motivation, research gaps, problem statement, research questions, and research objectives. Besides that, this chapter also includes the significance and limitations of the study. The main topic focuses on classifying e-commerce product data, specifically utilizing the Hidden Markov Model and the enhancement for more accurate classification. This chapter is essential in highlighting the reasons for conducting this research, outlining its expected contributions to both theory and practice. In particular, the research contributes by proposing improved methodologies for short-text product title classification.

1.2 Research Motivation

Automatic text classification (TC) is crucial for organizing the massive amount of textual data that has accumulated over time. TC refers to the task of automatically assigning pre-defined categories to text by learning from labeled data. It is widely applied in various domains such as classifying in-patient discharge summaries (Fernandes et al., 2021; Roussinov et al., 2022), flora data classification (Molina-Venegas et al., 2021), legal document classification (Rusiya & Jamatia, 2023; Stow et al., 2023), and patent document classification (Haghighian Roudsari et al., 2022; Miric et al., 2023). However, TC can be challenging due to the ambiguous meanings of terms, the high dimensionality of the data, and the need for models that are both accurate and efficient for large datasets.

Product classification is a specific application of TC that assigns pre-defined product categories to product-related text data, such as product metadata (titles and descriptions). This process helps organize the large-scale data generated on e-commerce platforms. Automated e-commerce product classification has been recognised as an important task for managing online product data (Moiseev, 2016) and is typically