ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Đồ án
Xây dựng trò chơi Mummy Maze

Báo cáo kĩ thuật
CSC10013 - Cơ sở lập trình cho Trí tuệ nhân tạo

Sinh viên thực hiện:

Trần Đình Hưng

Vương Thành Phát

Lê Thanh Phi

Nguyễn Quốc Thái

Giảng viên hướng dẫn:

Lê Thanh Tùng

Trần Hoàng Quân

Thành phố Hồ Chí Minh, Ngày 11 tháng 1 năm 2026

MỤC LỤC

MỤC LỤC **1**

1 Introduction **2**

1.1 Current background 2

1.2 Position of the text classification problem 3

1.3 Scientific significance 4

REFERENCES **5**

CHƯƠNG 1

Introduction

1.1. Current background

1. The explosion of text data

   In recent years, the amount of digital documents and complex text has grown exponentially [1], [2]. IDC forecasts that global data volume will increase from 33 Zettabytes (in 2018) to 175 Zettabytes by 2025. This data source is extremely diverse, including emails, blogs, administrative documents, and personal communications, leading to information overload [3]. The current rate of text information generation has far exceeded the manual processing capabilities of humans, making the development of automatic classification systems urgent [1], [3].

2. The unstructured nature of data

   The input data for language processing tasks is raw and unstructured text [1]. Unlike other data types, text does not have an intrinsic arithmetic representation that computers can process immediately [1]. Most texts are written freely, lacking standardized formatting (except for some scientific articles), requiring reliance on keyword occurrences or semantic features for classification [3]. Notably, nearly 80% of business information exists in this form of unstructured text data [2].

3. Current challenges

   Modern text processing and classification face numerous significant challenges.

   - **Data representation:** It is necessary to convert unstructured text into a structured feature space [2]. However, this is difficult because the data contains a lot of "noise" (stops, spelling errors, slang) and a huge vocabulary (millions of words), causing problems with time and memory complexity [2].

- **Resource and technical requirements:** Traditional methods ("shallow learning") rely heavily on costly manual feature extraction and require expert knowledge [1]. In contrast, deep learning models, while powerful, require large amounts of data and high computational resources [2].
- **Transparency and reliability:** Deep learning models often suffer from the "black box" problem, meaning a lack of ability to explain how decisions are made [2]. Additionally, there are concerns that large language models act like "stochastic parrots"—memorizing training data without actually understanding the language—and are vulnerable to counterattacks [1].

1.2. Position of the text classification problem

1. In the context of Machine Learning

   Text classification is a typical **semi-supervised or supervised learning problem** in which documents are assigned to predefined labels based on content [3]. This field has undergone a dramatic paradigm shift: from "shallow learning" methods that rely on costly manual feature design to **deep learning** methods with the ability to automatically extract complex and nonlinear semantic features [1], [2].

2. The intersection between fields

   This problem is at the heart of the development of **Text Mining** and **Natural Language Processing (NLP)** [1]. It demonstrates a clear technological intersection when applying architectures from computer vision (such as CNNs to capture discriminant phrases) [1], [3] and graph theory (Graph Neural Networks) [1] to traditional probabilistic statistical models [2].

3. The foundational role in information systems

   Text classification is a core tool for addressing the problem of digital information overload [3]. Its role spans many important applications

   - **Information Retrieval and Filtering:** The foundation for search engines, spam filtering systems, and automated document organization [1], [2].
   - **Advanced Data Analysis:** Supports sentiment analysis, user opinion discovery, and

recommender systems [2], [3].

- **Knowledge Management:** Automates text summarization and management of unstructured data sources (accounting for up to 80% of business information) in fields such as healthcare, law, and business [2].

1.3. Scientific significance

1. Promoting research on Natural Language Processing (NLP)

   Text classification is a fundamental task and a major challenge, playing a key role in the development of the field of Text Mining and NLP [1], [2]. The success of algorithms in this field is based on the ability to model complex and non-linear relationships, requiring a deep understanding of modern machine learning methods [2].

2. Solving the data representation problem

   This is the core problem in transforming unstructured text data into a structured feature space that computers can process [1], [2]. Efforts to solve this problem have driven the shift from manual feature extraction methods (such as BoW, TF-IDF) to deep learning techniques capable of automatically learning complex semantic and syntactic features through word embedding and contextual representation models [1], [2].

3. Foundation for more complex problems

   Text classification techniques serve as a foundation for many advanced information processing applications such as Information Retrieval, Sentiment Analysis, Recommender Systems, and Text Summarization [1], [2]. In addition, it provides a methodological basis for solving in-depth problems in fields such as healthcare (medical record coding), law, and social sciences [2].

REFERENCES

[1] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, 2022, ISSN: 2078-2489. DOI: 10.3390/info13020083 [Online]. Available: https://www.mdpi.com/2078-2489/13/2/83

[2] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, 2019, ISSN: 2078-2489. DOI: 10.3390/info10040150 [Online]. Available: https://www.mdpi.com/2078-2489/10/4/150

[3] M. Zaveri, "Automatic text classification: A technical review," *International Journal of Computer Applications*, 2011. DOI: 10.5120/3358-4633 [Online]. Available: https://doi.org/10.5120/3358-4633