

**ĐẠI HỌC THÁI NGUYÊN**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

---

**LÊNG HOÀNG LÂM**

**PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT VÀ**  
**ỨNG DỤNG VÀO CÁC CƠ QUAN NHÀ NƯỚC TỈNH BẮC KẠN**

**Chuyên ngành: Khoa học máy tính**

**Mã số: 60 48 0101**

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**Người hướng dẫn khoa học: PGS.TS. ĐOÀN VĂN BAN**

**Thái Nguyên - 2017**  
**LỜI CAM ĐOAN**

Tôi xin cam đoan đây là sản phẩm nghiên cứu, tìm hiểu của cá nhân tôi. Các số liệu, kết quả trình bày trong luận văn là trung thực. Những nội dung trình bày trong luận văn hoặc là của bản thân, hoặc là được tổng hợp từ những nguồn tài liệu có nguồn gốc rõ ràng và được trích dẫn hợp pháp, đầy đủ.

Tôi xin hoàn toàn chịu trách nhiệm cho lời cam đoan của mình.

*Thái Nguyên, tháng 4 năm 2017*  
**HỌC VIÊN**

**Lèng Hoàng Lâm**

**LỜI CẢM ƠN**

Trân trọng cảm ơn các thầy giáo, cô giáo trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên; các giảng viên đến từ Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Trường Đại học Quốc gia Hà Nội... đã tạo điều kiện tốt nhất cho học viên trong quá trình học tập và làm luận văn. Đặc biệt, xin được bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới thầy giáo, PGS.TS. Đoàn Văn Ban, người đã định hướng và luôn tận tình chỉ bảo, hướng dẫn em trong việc nghiên cứu, thực hiện luận văn này.

Trong suốt quá trình học tập và thực hiện đề tài, học viên luôn nhận được sự ủng hộ, động viên của gia đình, đồng nghiệp, đặc biệt là sự quan tâm tạo điều kiện của Ban lãnh đạo Trung tâm Công nghệ thông tin và Truyền thông tỉnh Bắc Kạn - nơi học viên đang công tác. Xin trân trọng cảm ơn!

*Thái Nguyên, tháng 4 năm 2017*

**HỌC VIÊN**

**Lèng Hoàng Lâm**

**MỤC LỤC**

LỜI CAM ĐOAN .....	i
LỜI CẢM ƠN .....	ii
MỤC LỤC.....	iii
DANH MỤC CÁC TỪ VIẾT TẮT .....	v
DANH MỤC CÁC HÌNH.....	vi
DANH MỤC CÁC BẢNG.....	vii
MỞ ĐẦU.....	1
CHƯƠNG I. TỔNG QUAN VỀ PHÂN LOẠI VĂN BẢN TIẾNG VIỆT .....	3
1.1. Khai phá dữ liệu .....	4
1.2. Khai phá dữ liệu văn bản .....	7
1.3. Phân loại văn bản .....	11
1.3.1. Giới thiệu bài toán phân loại văn bản .....	11
1.3.2. Quy trình phân loại văn bản.....	12
1.3.3. Phân loại văn bản tiếng Việt.....	13
1.4. Đặc trưng của văn bản tiếng Việt .....	14
1.4.1. Các đơn vị của tiếng Việt .....	14
1.4.2. Ngữ pháp của tiếng Việt.....	17
1.4.3. Từ tiếng Việt.....	18
1.4.4. Câu tiếng Việt .....	20
1.4.5. Các đặc điểm chính tả và văn bản tiếng Việt .....	23
1.5. Công tác quản lý văn bản tại các cơ quan tỉnh Bắc Kạn .....	23
1.6. Kết luận chương 1 .....	25
CHƯƠNG II. CÁC KỸ THUẬT TRONG PHÂN LOẠI VĂN BẢN TIẾNG VIỆT.....	25
2.1. Tách từ trong văn bản .....	26
2.1.1. Phương pháp khớp tối đa.....	27
2.1.2. Mô hình tách từ bằng WFST và mạng Neural.....	28
2.1.3. Phương pháp học dựa vào sự biến đổi trạng thái .....	29
2.1.4. Loại bỏ từ dừng.....	31
2.2. Trọng số của từ trong văn bản .....	31
2.2.1. Phương pháp Boolean.....	32
2.2.2. Phương pháp dựa trên tần số .....	32

2.3. Các mô hình biểu diễn văn bản.....	33
2.3.1. Mô hình Boolean .....	33
2.3.2. Mô hình xác suất.....	33
2.3.3. Mô hình không gian vector.....	34
2.4. Độ tương đồng văn bản.....	36
2.5. Thuật toán phân loại văn bản .....	39
2.5.1. Thuật toán Support Vector Machine (SVM) .....	39
2.5.2. Thuật toán K-Nearest Neighbor (kNN) .....	43
2.5.3. Thuật toán Naïve Bayers (NB) .....	44
2.6. Phân loại văn bản tiếng Việt.....	47
2.6.1. Trích chọn đặc trưng văn bản .....	47
2.6.2. Sử dụng thuật toán SVM để phân loại văn bản .....	50
2.7. Kết luận chương 2.....	53
<b>CHƯƠNG III. ÁP DỤNG THUẬT TOÁN SUPPORT VECTOR MACHINE</b> <b>PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT.....</b>	<b>54</b>
3.1. Ứng dụng SVM vào bài toán phân loại văn bản hành chính tiếng Việt tại các cơ quan nhà nước tỉnh Bắc Kạn.....	54
3.2. Áp dụng phân loại văn bản .....	56
3.3. Xây dựng chương trình thử nghiệm ứng dụng phân loại văn bản áp dụng vào máy tìm kiếm văn bản hành chính tiếng Việt .....	57
3.3.1. Mô tả bài toán .....	57
3.3.2. Quá trình tiền xử lý văn bản .....	59
3.3.3. Vector hóa và trích chọn đặc trưng văn bản .....	60
3.3.4. Đánh giá bộ phân lớp.....	60
3.3.5. Chương trình thực nghiệm.....	62
3.3.6. Kết quả thực nghiệm.....	62
3.4. Kết luận chương 3 .....	63
<b>KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....</b>	<b>64</b>
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>65</b>

## **DANH MỤC CÁC TỪ VIẾT TẮT**

<b>Từ viết tắt</b>	<b>Giải thích</b>
CSDL	Cơ sở dữ liệu
KDD	Knowledge Discovery from Data
IDF	Inverse Document Frequency
kNN	K-Nearest Neighbor
NB	Naïve Bayers
SVM	Support Vector Machine
S <sup>3</sup> VM	Semi-Supervised Support Vector Machine
TBL	Transformation - based Learning
TF	Term Frequency
WFST	Weighted Finite - State Transducer

## **DANH MỤC CÁC HÌNH**

Hình 1.1. Các bước trong quá trình phát hiện tri thức từ CSDL (KDD) .....	5
Hình 1.2. Quy trình phân loại văn bản .....	13
Hình 2.1. Biểu diễn văn bản theo mô hình xác suất .....	34
Hình 2.2. Minh họa hình học thuật toán SVM.....	40
Hình 2.3. Chi tiết giai đoạn huấn luyện .....	50
Hình 2.4. Mô hình SVM .....	51
Hình 3.1. Chi tiết giai đoạn huấn luyện .....	58
Hình 3.2. Chi tiết giai đoạn phân lớp .....	59

## **DANH MỤC CÁC BẢNG**

Bảng 3.1. Bộ dữ liệu thử nghiệm .....	62
Bảng 3.2. Kết quả phân lớp bộ dữ liệu kiểm tra .....	63
Bảng 3.3. Đánh giá hiệu suất phân lớp .....	63



## MỞ ĐẦU

### 1. Đặt vấn đề

Trong thời đại bùng nổ Công nghệ thông tin hiện nay, phương thức sử dụng văn bản giấy truyền thống đã dần được số hóa, chuyển sang dạng các văn bản điện tử lưu trữ trên máy tính và được chia sẻ, truyền tải trên mạng. Với rất nhiều tính năng ưu việt của tài liệu số như: Lưu trữ gọn nhẹ, linh hoạt; thời gian lưu trữ lâu dài; dễ hiệu chỉnh và đặc biệt tiện dụng trong trao đổi, chia sẻ nên ngày nay, số lượng văn bản điện tử được sử dụng trong các cơ quan nhà nước tăng lên rất nhanh chóng. Do đó, một vấn đề đặt ra là làm thế nào để có thể tìm kiếm và khai thác thông tin từ nguồn dữ liệu phong phú này. Các kỹ thuật để giải quyết vấn đề này được gọi là “Text Mining” hay Khai phá dữ liệu văn bản.

Khai phá dữ liệu văn bản đề cập đến tiến trình trích lọc các mẫu hình thông tin hay tri thức đáng quan tâm hoặc có giá trị từ các tài liệu văn bản. Trong đó, phân loại văn bản là một bài toán cơ bản nhất của lĩnh vực khai phá dữ liệu văn bản. Phân loại văn bản là công việc phân tích nội dung của văn bản và sau đó ra quyết định (hay dự đoán) văn bản thuộc nhóm nào trong các nhóm văn bản đã cho trước. Văn bản được phân loại có thể thuộc một nhóm, nhiều nhóm, hoặc không thuộc nhóm văn bản mà ta đã định nghĩa trước. Phân loại văn bản có thể thực hiện bằng nhiều cách như sử dụng tiếp cận lý thuyết tập thô, cách tiếp cận theo luật kết hợp hoặc dựa trên cách tiếp cận máy học. Đây là một lĩnh vực mang tính khoa học cao, ứng dụng được rất nhiều trong các bài toán thực tế hiện nay như tìm kiếm thông tin, lọc văn bản, tổng hợp tin tức tự động, thư viện điện tử,... Do vậy, học viên quyết định chọn đề tài “*Phân loại văn bản hành chính tiếng Việt và ứng dụng vào các cơ quan nhà nước tỉnh Bắc Kạn*” để nghiên cứu, thực hiện luận văn tốt nghiệp của mình.

Mục tiêu của đề tài luận văn là khảo sát, tìm hiểu một số phương pháp

phân loại văn bản thường được sử dụng hiện nay, trên cơ sở đó đề xuất lựa chọn một phương án phân loại văn bản tiếng Việt tự động và ứng dụng thử nghiệm phân loại cho một đối tượng cụ thể là văn bản hành chính tiếng Việt.

## **2. Đối tượng và phạm vi nghiên cứu**

Đối tượng nghiên cứu bao gồm: Các thuật toán phân loại văn bản và các vấn đề liên quan đến bài toán phân loại văn bản tiếng Việt.

Phạm vi nghiên cứu của luận văn tập trung vào một số thuật toán phân loại văn bản thông dụng; các đặc trưng của văn bản tiếng Việt; các kỹ thuật liên quan trong xử lý phân loại văn bản và ứng dụng thuật toán học bán giám sát trong phân loại văn bản tiếng Việt.

## **3. Hướng nghiên cứu của đề tài**

Nghiên cứu lý thuyết cơ bản về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản với một số thuật toán phân loại văn bản thông dụng như Naïve Bayers, K-Nearest Neighbor, Support Vector Machine.

Nghiên cứu về các đặc trưng của văn bản tiếng Việt và các kỹ thuật liên quan trong xử lý phân loại văn bản tiếng Việt như tách từ, biểu diễn văn bản, đánh trọng số của từ, tính độ tương đồng văn bản.

Từ kết quả thu được tiến hành cài đặt ứng dụng trong bài toán phân loại văn bản hành chính tiếng Việt.

## **4. Những nội dung chính**

Nội dung chính của luận văn được trình bày trong 3 chương với tổ chức cấu trúc như sau:

### **Chương 1.** Tổng quan về phân loại văn bản tiếng Việt.

Chương này trình bày khái quát về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản tiếng Việt; đồng thời làm rõ các đặc trưng của văn bản tiếng Việt và giới thiệu sơ bộ về công tác quản lý văn bản tại các cơ quan thuộc tỉnh Bắc Kạn.