**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI:**

# **Phân loại cảm xúc trong văn bản sử dụng Machine Learning - KHẢO SÁT ĐỀ TÀI -**

A blue button with text

AI-generated content may be incorrect.

**Giáo viên hướng dẫn: Đỗ Như Tài**

**Sinh viên thực hiện:**

**Trịnh Bảo Quân - 3119410335   
 Nguyễn Nhựt Hào - 3122410097**

**Ngô Khánh Tâm - 3122410370**

***Thành phố Hồ Chí Minh, tháng 05/2024***

**1. Giới thiệu**

**1.1 Bối cảnh và lý do chọn đề tài**

Mạng xã hội đã trở thành một phần không thể thiếu trong đời sống hiện đại. Hàng tỷ người dùng đăng tải trạng thái, bình luận, chia sẻ ý kiến mỗi ngày, tạo ra một lượng dữ liệu văn bản khổng lồ. Việc phân tích cảm xúc từ những dữ liệu này không chỉ giúp các doanh nghiệp hiểu khách hàng hơn mà còn hỗ trợ chính phủ và các tổ chức trong việc đo lường xu hướng dư luận.

Theo báo cáo của **Statista (2023)**, trung bình mỗi ngày có hơn **500 triệu tweet**, **293 tỷ email**, và **2,5 tỷ bài đăng trên Facebook** được tạo ra. Sự phát triển này đặt ra nhu cầu cấp thiết về các công cụ phân tích cảm xúc tự động để xử lý và hiểu nội dung khổng lồ này.

🔗 **Nguồn:** [Statista – Volume of data created each day on the Internet](https://www.statista.com/statistics/871513/worldwide-data-created/)

**1.2 Mục tiêu nghiên cứu**

Đề tài tập trung vào việc xây dựng hệ thống phân tích cảm xúc văn bản trên mạng xã hội bằng cách sử dụng các mô hình học máy và xử lý ngôn ngữ tự nhiên (NLP). Các mục tiêu cụ thể gồm:

* Khảo sát các phương pháp phân tích cảm xúc phổ biến hiện nay.
* Thu thập và xử lý tập dữ liệu từ mạng xã hội (Twitter, Facebook, Reddit).
* Ứng dụng các mô hình học máy như **Naïve Bayes, LSTM, và BERT** để phân loại cảm xúc.
* Đánh giá độ chính xác và tính khả thi của mô hình trong thực tế.

**2. Tổng quan nghiên cứu**

**2.1 Các phương pháp phân tích cảm xúc**

**1. Phương pháp dựa trên từ điển (Lexicon-based Approach):**

* Dựa vào danh sách từ vựng đã gán nhãn cảm xúc để xác định cảm xúc trong văn bản.
* Ví dụ: Sử dụng **SentiWordNet** hoặc **VADER** để gán điểm cảm xúc cho từng từ.
* **Hạn chế:** Không xử lý tốt ngữ cảnh hoặc cách diễn đạt phức tạp.

**2. Phương pháp dựa trên học máy (Machine Learning-based Approach):**

* Sử dụng các mô hình như **Naïve Bayes, SVM, hoặc Random Forest** để huấn luyện hệ thống phân loại cảm xúc.
* Ví dụ: Nghiên cứu của **Pang & Lee (2008)** đã chứng minh rằng mô hình SVM đạt độ chính xác lên đến **85%** khi phân loại cảm xúc phim.

🔗 **Nguồn:** [Pang & Lee – Opinion Mining and Sentiment Analysis](https://www.cs.cornell.edu/home/llee/omsa/omsa.html)

**3. Phương pháp dựa trên học sâu (Deep Learning-based Approach):**

* Sử dụng các mạng nơ-ron như **LSTM, GRU, BERT** để phân tích ngữ cảnh của văn bản.
* Ví dụ: **Devlin et al. (2019)** đã giới thiệu mô hình **BERT**, giúp cải thiện độ chính xác trong phân tích cảm xúc lên đến **90%**.

🔗 **Nguồn:** [Devlin et al. – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805)

**2.2 Ứng dụng thực tế**

Phân tích cảm xúc trên mạng xã hội được ứng dụng trong nhiều lĩnh vực:

* **Marketing & Quảng cáo:** Các công ty có thể phân tích phản hồi của khách hàng để cải thiện chiến lược tiếp thị.
* **Chính trị & Xã hội:** Theo dõi xu hướng dư luận về các vấn đề chính trị, xã hội.
* **Tài chính:** Dự đoán biến động thị trường dựa trên cảm xúc nhà đầu tư.

Ví dụ, nghiên cứu của **Bollen et al. (2011)** cho thấy rằng phân tích cảm xúc từ Twitter có thể dự đoán biến động của thị trường chứng khoán với độ chính xác lên đến **87%**.

🔗 **Nguồn:** [Bollen et al. – Twitter mood predicts the stock market](https://arxiv.org/abs/1010.3003)

**3. Phương pháp thực hiện**

**3.1 Thu thập dữ liệu**

* Sử dụng **API của Twitter, Facebook, Reddit** để thu thập dữ liệu.
* Dữ liệu được làm sạch bằng cách loại bỏ stop words, biểu tượng cảm xúc, đường dẫn, v.v.
* Ví dụ, sử dụng công cụ **Tweepy** để thu thập tweet có chứa từ khóa "happy" và "sad".

**3.2 Huấn luyện mô hình**

* Dữ liệu được chia thành tập **huấn luyện (80%)** và **kiểm tra (20%)**.
* Áp dụng các mô hình học máy và học sâu để huấn luyện.
* Đánh giá bằng các chỉ số như **độ chính xác (Accuracy), độ nhạy (Recall), và F1-score**.

**4. Kết quả mong đợi**

Dự kiến mô hình sẽ đạt độ chính xác trên **85%** khi phân loại cảm xúc từ văn bản. Ngoài ra, hệ thống sẽ giúp tự động hóa việc phân tích dư luận, hỗ trợ các doanh nghiệp và tổ chức đưa ra quyết định nhanh chóng và chính xác hơn.

**5. Kết luận**

Đề tài **"Phân tích cảm xúc văn bản trên mạng xã hội"** là một hướng nghiên cứu quan trọng và có tính ứng dụng cao. Việc áp dụng AI và NLP vào phân tích cảm xúc giúp khai thác dữ liệu hiệu quả hơn, từ đó đưa ra những quyết định chính xác hơn trong nhiều lĩnh vực.

🔗 **Nguồn:** [Liu, B. – Sentiment Analysis and Opinion Mining](https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html)