Research Paper Content

# 1. Giới thiệu

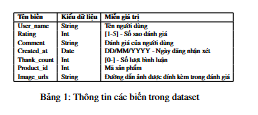
Với sự phát triển nhanh chóng của thương mại điện tử, các nền tảng bán hàng trực tuyến ngày càng trở thành nơi để người tiêu dùng đưa ra quyết định mua sắm. Một yếu tố quan trọng trong quyết định mua hàng chính là các đánh giá sản phẩm. Những nhận xét này không chỉ giúp người mua hiểu rõ hơn về chất lượng sản phẩm mà còn đóng vai trò quan trọng trong việc quyết định lựa chọn sản phẩm.  
  
Tuy nhiên, không phải tất cả các đánh giá đều hữu ích. Người tiêu dùng thường gặp khó khăn trong việc phân loại các đánh giá có thực sự hữu ích hay không. Điều này dẫn đến một vấn đề quan trọng: làm thế nào để đánh giá được mức độ hữu ích của các nhận xét tự động?  
  
Nghiên cứu này nhằm dự đoán mức độ hữu ích của các nhận xét sản phẩm trên sàn thương mại điện tử Tiki bằng cách sử dụng phương pháp học máy. Bằng cách phân tích các yếu tố ảnh hưởng đến mức độ hữu ích của nhận xét, nghiên cứu sẽ giúp người tiêu dùng tìm kiếm thông tin đáng tin cậy nhất khi mua sắm trực tuyến.

# 2. Nền tảng và các nghiên cứu liên quan

Việc đánh giá tính hữu ích của nhận xét sản phẩm trên các sàn thương mại điện tử đã thu hút sự quan tâm lớn. Các nhận xét đóng vai trò quan trọng trong quyết định mua hàng của người tiêu dùng, nhưng không phải tất cả nhận xét đều hữu ích. Vì vậy, việc phân loại nhận xét hữu ích và không hữu ích là rất quan trọng.  
  
Phương pháp học máy (Machine Learning)  
Nhiều nghiên cứu đã áp dụng các phương pháp học máy như Support Vector Machines (SVM) để phân loại nhận xét thành hữu ích hoặc không hữu ích. Ngoài SVM, các phương pháp như Random Forest và Logistic Regression cũng được sử dụng với độ chính xác khác nhau.  
  
Phân tích ngữ nghĩa và cảm xúc  
Các nghiên cứu đã chỉ ra rằng cảm xúc trong nhận xét ảnh hưởng lớn đến tính hữu ích. Nhận xét với cảm xúc tiêu cực thường ít hữu ích hơn. Ngoài ra, độ dài, cấu trúc ngữ pháp và mức độ chi tiết cũng là yếu tố quan trọng.  
  
Đặc trưng của người viết nhận xét  
Nhận xét từ những người có uy tín (ví dụ: người mua hàng nhiều lần) thường được coi là hữu ích hơn. Mức độ chủ quan của nhận xét cũng là yếu tố quan trọng, đặc biệt trong các sản phẩm có tính chất cảm nhận cao như mỹ phẩm và thời trang.

# 3. Dữ liệu

Dữ liệu sử dụng trong nghiên cứu này được thu thập từ các nhận xét sản phẩm trên sàn thương mại điện tử Tiki. Mỗi nhận xét bao gồm các thông tin sau:  
  
- Tên sản phẩm: Tên của sản phẩm được đánh giá.  
- Đánh giá: Mức đánh giá từ 1 đến 5 sao.  
- Nhận xét: Nội dung nhận xét của người dùng.  
- Ngày đăng: Thời gian người dùng đăng nhận xét.  
- Số lượt đánh giá: Số lượt bình chọn "Hữu ích" từ người dùng khác.  
  
Dữ liệu này đã được tiền xử lý bao gồm việc loại bỏ các từ dừng (stop words), xử lý từ vựng (stemming), và chuyển đổi về dạng vector cho các mô hình học máy. Tất cả dữ liệu đều được phân loại thành ba loại chính:  
- Có comment: Nhận xét có văn bản.  
- Có ảnh: Nhận xét có kèm hình ảnh.  
- Có cả comment và ảnh: Nhận xét có kèm văn bản và hình ảnh.

  
Số lượng sample: Dữ liệu được chia thành 3 loại: "Có comment", "Có ảnh", và "Có cả comment & ảnh". Đồ thị dưới đây thể hiện số lượng mẫu của từng loại.

Dưới đây là đồ thị thể hiện số lượng mẫu của từng loại nhận xét:

