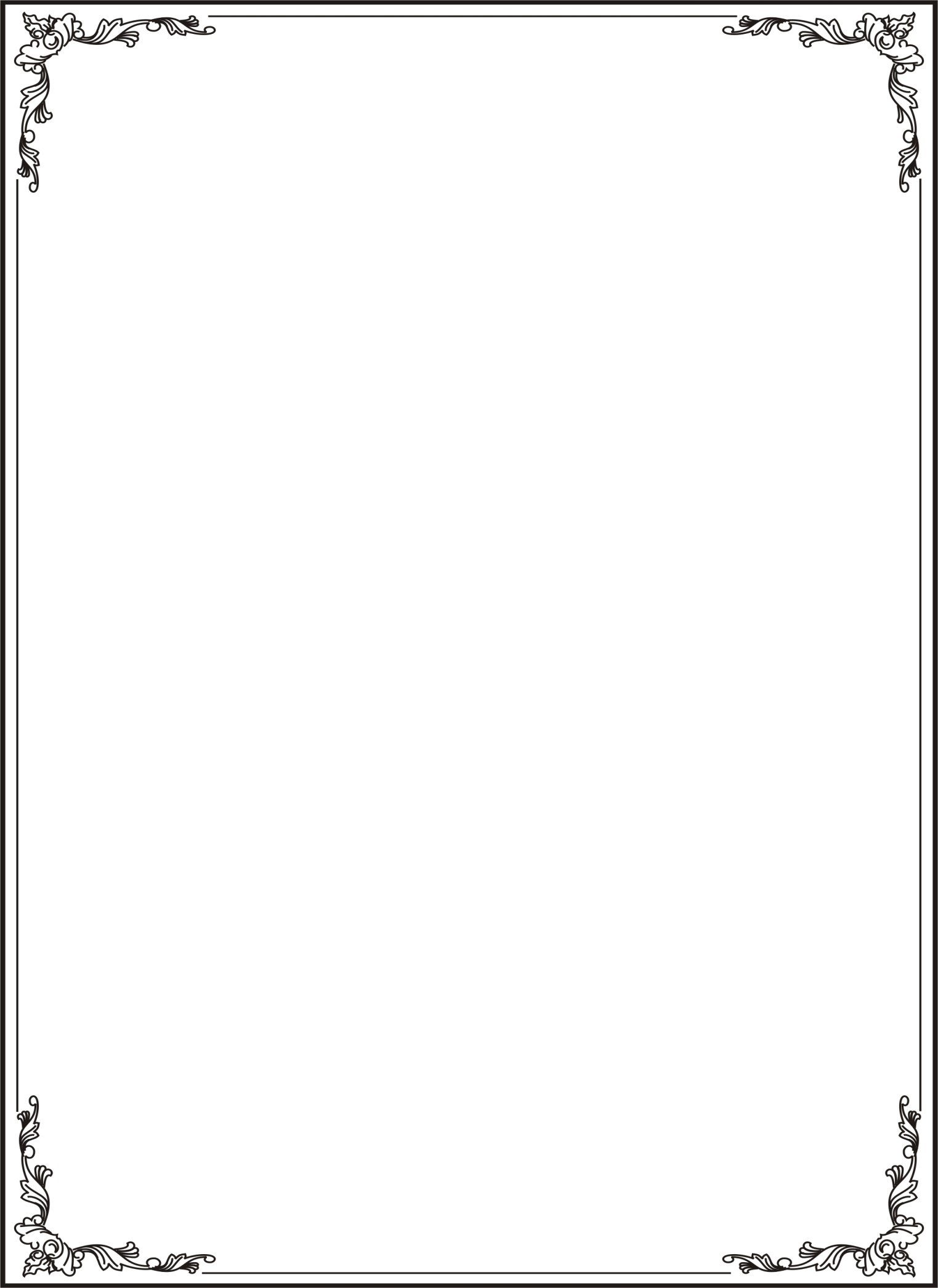
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC QUY NHƠN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO KHOA HỌC**

**Phân tích tình cảm và khai thác ý kiến bằng cách sử dụng**

**học sâu cho các bài đánh giá trên Google Play**

Sinh viên thực hiện: Ung Minh Hoài

MSSV: 4251050044

Lớp: CNTT K42A

Giảng viên hướng dẫn: Lê Quang Hùng

BÌNH ĐỊNH, 2022

**Mục lục**

[**1. Giới thiệu** 3](#_Toc114049431)

[**2. Công trình liên quan và Bối cảnh** 6](#_Toc114049432)

[**2.1 Công việc liên quan** 6](#_Toc114049433)

[**2.2 Kỹ thuật tính năng cho NLP** 7](#_Toc114049434)

[**2.3 Bộ phân loại học máy** 7](#_Toc114049435)

[**2.4 Mô hình học sâu tuần tự** 9](#_Toc114049436)

[**3 Bộ sưu tập Corpus** 9](#_Toc114049437)

[**4 Phân tích Corpus** 10](#_Toc114049438)

[**5 Xây dựng Mô hình và Kết quả** 13](#_Toc114049439)

[**5.1 Kết quả cho bộ phân loại học máy (Machine Learning Classifiers)** 13](#_Toc114049440)

[**5.2 Kết quả cho mô hình học sâu (Deep Learning Model)** 14](#_Toc114049441)

[**6. Kết luận** 16](#_Toc114049442)

[**Tài liệu tham khảo** 17](#_Toc114049443)

**Phân tích tình cảm và khai thác ý kiến bằng cách sử dụng**

**học sâu cho các bài đánh giá trên Google Play**

Sercan Sari and Murat Kalender

Khoa Kỹ thuật Máy tính, Đại học Yeditepe, Istanbul, Thổ Nhĩ Kỳ

{ssari, mkalender}@cse.yeditepe.edu.tr

**Abstract.** “Sentiment analysis and opinion mining” - Phân tích tình cảm và khai thác ý kiến ​​có vai trò quan trọng để theo dõi hành vi của người tiêu dùng. Với những tiến bộ gần đây trong kỹ thuật machine learning (học máy), vấn đề này đã được giải quyết và đi một chặng đường dài bằng tiếng Anh. Tuy nhiên, trong các ngôn ngữ phổ biến như tiếng Turkish (Thổ Nhĩ Kỳ), nó vẫn là một trong những chủ đề nóng. Trong nghiên cứu này, chúng tôi so sánh một số classification method (phương pháp phân loại) và một số deep learning method (phương pháp học sâu) để đưa ra phân tích tình cảm cho các review (bài đánh giá) của Thổ Nhĩ Kỳ mà chúng tôi đã thu thập được từ Google Play. Chúng tôi có accuracy (độ chính xác) dự đoán 87,30% cho Multinomial Naive Bayes và độ chính xác dự đoán 95,87% cho mô hình học sâu. Chúng tôi có kết quả đáng kể cả từ bộ phân loại học máy và mô hình học sâu. Mặc dù không có sự khác biệt giữa các bộ phân loại học máy khi chúng ta sử dụng các vector khác nhau, nhưng có sự khác biệt khi chúng ta xây dựng mô hình học sâu để dự đoán target value (giá trị mục tiêu). Mô hình này cũng có thể được áp dụng cho bất kỳ dữ liệu nào từ Twitter, Facebook hoặc bất kỳ tiểu blog nào khác.

**Keywords:** sentiment analysis · opinion mining · machine learning · text classification · deep learning

1. **Giới thiệu**

Hành vi của người dùng trực tuyến đã thay đổi khi tốc độ ngày càng tăng của dữ liệu trên internet. Theo thống kê [11], 70% người dùng trực tuyến thích mua sắm đồ điện tử cho biết rằng họ đã đọc các bài đánh giá trực tuyến trước khi mua sản phẩm. Tình hình này thúc đẩy các nghiên cứu về phân tích tình cảm và khai thác ý kiến ​​đối với các nguồn trực tuyến như Twitter và Facebook trên internet. Có một số nghiên cứu [18], [4], [14] phân tích các nền tảng blog như Twitter và Facebook để đưa ra ý kiến ​​của mọi người và phân loại chúng theo cảm tính. Nhiều loại thông tin khác nhau có thể được trích xuất từ ​​các nguồn này.

Ví dụ: các công ty sản xuất có thể thu thập thông tin về sản phẩm của họ và các đảng phái chính trị và tổ chức xã hội có thể xác định các sự kiện của họ theo kết quả của nghiên cứu này.

Với những tiến bộ gần đây trong kỹ thuật học máy, vấn đề này đã được giải quyết và đi một chặng đường dài bằng tiếng Anh. Tuy nhiên, trong các ngôn ngữ tổng hợp như tiếng Thổ Nhĩ Kỳ, nó vẫn là một trong những chủ đề nóng [13], [16], [25]. Mặc dù có một số khuyến cáo sử dụng blog, đặc biệt là Twitter, chẳng hạn như nhiều loại dữ liệu và tốc độ dữ liệu [18], một số tweet có thể có nội dung mỉa mai [5] và nó có thể có vấn đề để label (gắn nhãn) một dữ liệu khổng lồ như vậy bằng cách chỉ nhìn vào một số feature (tính năng) trong đó [7]. Những điều này có thể dẫn đến kết quả không hiệu quả và làm giảm hiệu suất của các mô hình học máy.

Để giải quyết những vấn đề này và để tạo ra các mô hình dự đoán chính xác hơn, chúng tôi đưa ra một cách tiếp cận khác. Trong tài liệu, đã có những nghiên cứu mà họ sử dụng các bài phê bình phim [19], [24], [10], [8], [9]. Cách tiếp cận mà chúng tôi đã theo dõi cũng tương tự như các nghiên cứu này. Chúng tôi sử dụng các bài đánh giá trên Google Play để loại bỏ ambiguity (sự mơ hồ/không rõ nghĩa) của văn bản. Bởi vì trong những phát biểu như vậy, mọi người trực tiếp đề cập đến ý kiến ​​của họ và đánh giá chúng trước khi gửi chúng. Kết quả của việc xếp hạng trước khi gửi, các vấn đề nói trên có thể được làm sáng tỏ một cách dễ dàng. Mặc dù, nghiên cứu của chúng tôi có những điểm tương đồng với tình trạng văn học hiện tại, nhưng điểm khác biệt chính của chúng tôi so với hiện trạng của các nghiên cứu là chúng tôi xây dựng một deep neural network (mạng lưới thần kinh sâu) để phân tích tình cảm và khai thác ý kiến ​​cho kho ngữ liệu văn bản Thổ Nhĩ Kỳ. Chúng tôi cũng đóng góp tài liệu bằng cách chia sẻ kho tài liệu mà chúng tôi đã chuẩn bị cho nghiên cứu này [1]. Trong nghiên cứu của mình, chúng tôi áp dụng các thuật toán học máy khác nhau và xây dựng một mạng lưới thần kinh sâu để phân tích cách các bài đánh giá trên Google Play ở Thổ Nhĩ Kỳ có thể được sử dụng cho mục đích phân tích tình cảm và khai thác ý kiến. Chúng tôi chọn các bài đánh giá vì những lý do sau:

* Phần đánh giá ứng dụng được sử dụng trực tiếp để bày tỏ ý kiến của người dùng.
* Đánh giá ứng dụng chứa các văn bản lớn và cập nhật.
* Có rất nhiều người khác nhau sử dụng một số lượng lớn các ứng dụng khác nhau có nghĩa là các bối cảnh khác nhau.

Chúng tôi đã thu thập tổng hợp 11000 bài đánh giá từ Google Play Thổ Nhĩ Kỳ bằng cách điều chỉnh xếp hạng của họ một cách đồng đều giữa hai bộ bài đánh giá:

1. reviews thuộc bốn và năm sao là cảm xúc tích cực (positive)
2. reviews thuộc về một và hai sao là cảm xúc tiêu cực. (negative)

|  |  |
| --- | --- |
| rating | review |
| 5 | Eskiden beri oynuyorum. Oyun muhte¸sem. Kesinlikle herkese tavsiye ediyorum. |
| 4 | Oyunu c¸ok be˘gendim. Yu¨kledi˘gim oyunların en gu¨zellerinden biri. |
| 2 | Telefonuma yaptı˘gım son gu¨ncelleme ile foto˘graftaki renkler bozulmaya ba¸sladı. |
| 1 | Program su¨rekli olarak kendini I˙ngilizce yapıyor. Sorunu ¸c¨ozu¨n lu¨tfen. |

**Table 1.** Examples of reviews on Google Play

Phần còn lại của bài báo sẽ được phân tích như sau:

* Phần 2, chúng tôi sẽ đề cập đến các công việc liên quan.
* Phần 3, chúng tôi sẽ làm rõ cách chúng tôi thu thập dữ liệu cho nghiên cứu của mình.
* Phần 4, sẽ đưa ra một phân tích về kho ngữ liệu mà chúng tôi sử dụng các mô hình đào tạo.
* Phần 5, chúng tôi trình bày việc xây dựng mô hình và kết quả nghiên cứu của chúng tôi.
* Phần 6, sẽ kết thúc bài báo.

**2. Nghiên cứu liên quan và Cơ sở lý thuyết**

**2.1 Nghiên cứu liên quan**

Đã có những nghiên cứu tập trung vào phân tích tình cảm và khai thác quan điểm trong tài liệu. Động lực chính đằng sau những nghiên cứu này đến từ khả năng theo dõi hành vi của người tiêu dùng và phạm vi rộng của dữ liệu.

Năm 2012, Kaya et al. [13] tích hợp các kỹ thuật phân loại tình cảm vào miền tin tức chính trị cho các trang tin tức Thổ Nhĩ Kỳ. Họ so sánh các thuật toán học tập có giám sát là Naive Bayes, Maximum Entropy, SVM và Mô hình ngôn ngữ N-Gram dựa trên ký tự để phân tích tình cảm đối với tin tức chính trị của người Thổ Nhĩ Kỳ. Kucuk et al. và cộng sự. [16] làm việc về Name entity recognition (nhận dạng thực thể được đặt tên) viết tắt là NER và báo cáo các thử nghiệm về NER trên các tweet của Thổ Nhĩ Kỳ.

Năm 2015, Yildirim et al.[25] báo cáo tác động của các lớp tiền xử lý đối với việc phân loại tình cảm của các văn bản trên mạng xã hội Thổ Nhĩ Kỳ. Mặc dù có một số lợi ích khi sử dụng tiểu blog, đặc biệt là Twitter và Facebook, chẳng hạn như nhiều loại dữ liệu và tốc độ của dữ liệu [18], một số văn bản tiểu blog có thể có nội dung mỉa mai [5] và nó có thể tạo ra sự mơ hồ để làm việc với một dữ liệu khổng lồ như vậy chỉ bằng cách nhìn vào một số tính năng trong đó [7]. Những loại tính năng này của dữ liệu có thể gây ra kết quả không hiệu quả. Chúng tôi sử dụng các bài đánh giá trên Google Play để loại bỏ sự mơ hồ của các văn bản. Bởi vì trong những tuyên bố như vậy, mọi người trực tiếp đề cập đến ý kiến ​​của họ và đánh giá chúng trước khi gửi chúng. Kết quả là, các vấn đề nói trên có thể được làm sáng tỏ một cách dễ dàng. Trong tài liệu, đã có những nghiên cứu khác mà họ sử dụng các bài phê bình phim [19], [24], [10], [8], [9]. Cách tiếp cận mà chúng tôi đã theo dõi cũng tương tự như các nghiên cứu này.

Nền tảng của nghiên cứu của chúng tôi bao gồm ba phần chính là feature engineering (kỹ thuật tính năng), machine learning classifiers (bộ phân loại học máy) và sequential deep learning model (mô hình học sâu tuần tự). Trong phần kỹ thuật tính năng, chúng tôi sẽ giải thích lý do tại sao chúng tôi cần các biểu diễn khác nhau của dữ liệu văn bản để trích xuất thông tin từ chúng. Trong phần thứ hai, chúng tôi sẽ đề cập đến các bộ phân loại học máy mà chúng tôi đã sử dụng trong nghiên cứu này. Cuối cùng, chúng tôi sẽ cung cấp chi tiết về mạng nơ-ron sâu mà chúng tôi đã sử dụng trong nghiên cứu của mình.

**2.2 Kỹ thuật tính năng cho NLP**

Trong quy trình ngôn ngữ tự nhiên (Natural language process - NLP), chúng ta không thể trực tiếp sử dụng văn bản để trích xuất thông tin và xây dựng mô hình học máy. Để sử dụng thông tin văn bản, chúng ta cần chuyển đổi nó thành các giá trị số. Một mô hình đơn giản và đầy đủ để giúp chúng ta có thể sử dụng các tài liệu văn bản trong học máy được gọi là Mô hình Bag-of-Words, hay BoW [12]. Mô hình đơn giản này thực sự tập trung vào sự xuất hiện của mỗi từ trong một văn bản. Bằng cách sử dụng phương pháp này, chúng ta có thể dễ dàng mã hóa mọi văn bản dưới dạng vectơ có độ dài cố định được mã hóa với độ dài của từ vựng mà chúng ta đã biết. Trong phạm vi của nghiên cứu này, chúng tôi sẽ giải thích vectơ hóa cho các tính năng dựa trên văn bản và phân tích hai trong ba cách khác nhau để sử dụng mô hình này trong thư viện scikit-learning [20] là CountVectorizer, TfidfVectorizer và HashVectorizer. Vì chúng tôi sẽ sử dụng CountVectorizer và TfidfVectorizer, các phần phụ sau đây sẽ giải thích chi tiết về chúng.

**CountVectorizer** cung cấp và triển khai cả đếm lần xuất hiện và mã hóa (tokenization). Nó xây dựng một kho từ vựng của những từ đã biết. Nó sử dụng từ vựng này để mã hóa dữ liệu văn bản mới. Mặc dù nó là một giải pháp tốt, có một số hạn chế khi sử dụng CountVectorizer chẳng hạn như các từ không liên quan xảy ra quá nhiều lần.

**TfidfVectorizer** là một phương pháp thay thế để tính tần số từ. Nó được gọi là Tần suất thuật ngữ - Tần suất tài liệu nghịch đảo (Term Frequency – Inverse Document frequency) viết tắt là TFIDF, là một phần của điểm số thuộc về mỗi từ.

* **Tần suất thuật ngữ**: Tần suất một từ nhất định xuất hiện trong văn bản
* **Tần suất tài liệu nghịch đảo**: Điều này thực sự điều chỉnh các từ xuất hiện rất nhiều lần trong văn bản.

**2.3 Bộ phân loại học máy**

Trong tiểu mục này, chúng tôi sẽ cung cấp một số chi tiết về các bộ phân loại học máy mà chúng tôi đã sử dụng trong nghiên cứu của mình.

**Multionmial Naive Bayes** bộ phân loại đa thức Naive Bayes dựa trên định lý Naive Bayes [17]. Đây là một cách tiếp cận cơ bản và đơn giản để xây dựng bộ phân loại vì nó nhanh chóng và dễ thực hiện. Trong [21], họ thảo luận rằng mặc dù nó có hiệu quả thực sự tốt, nhưng nó lại ảnh hưởng đến chất lượng kết quả vì những giả định của nó. Để loại bỏ những nhược điểm đó, họ giới thiệu phương pháp. Multionnal Naive Bayes (MNB). Mô hình MNB phân phối các từ trong ngữ liệu dưới dạng đa thức. Họ giả sử văn bản như một chuỗi các từ và cho rằng vị trí của các từ được tạo ra độc lập với nhau.

**K-Nearest Neighbors** KNN (K-Nearest Neighbors) là một trong những thuật toán phân loại đơn giản nhất và được sử dụng rộng rãi. KNN là một thuật toán không tham số và lười học và nó được sử dụng để phân loại và hồi quy [3]. Ý tưởng chính đằng sau phương pháp lân cận gần nhất là tìm nhãn cho điểm mới theo chỉ số khoảng cách gần nhất với nó và dự đoán nhãn từ chúng. Số lượng hàng xóm có thể được xác định bởi người dùng.

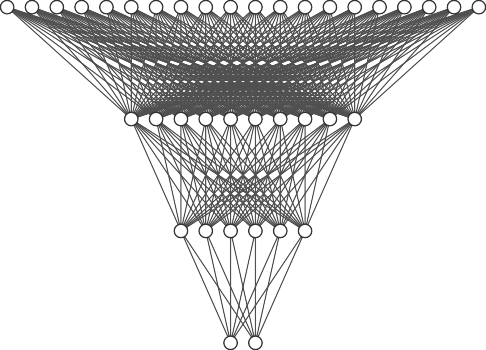
**Decision Tree learning** học cây quyết định là một trong những phương pháp thường được sử dụng trong phân tích dự báo [22]. Mục đích là xây dựng mô hình dự đoán đúng nhãn của các biến mục tiêu từ các biến đầu vào. Một cây được tạo ra bằng cách tách các biến đầu vào. Các đặc điểm phân loại có một vai trò quan trọng trong khi tách cây [23]. Có một số lợi thế khi sử dụng cây quyết định. Nó đơn giản để giải thích cây quyết định và nó sử dụng mô hình white-box (hộp trắng).

Trong nghiên cứu của chúng tôi, chúng tôi chọn các bộ phân loại ở trên và phân tích ảnh hưởng của các bộ vectơ khác nhau trên các mô hình khác nhau. Chúng tôi cũng đã xây dựng mô hình học sâu tuần tự để dự đoán đúng giá trị mục tiêu. Mạng học sâu mà chúng tôi đã sử dụng trong nghiên cứu của mình có thể được nhìn thấy trong Hình 1. Chúng tôi sẽ giải thích chi tiết về nó trong phần Xây dựng Mô hình và Kết quả.

**2.4 Mô hình học sâu tuần tự**

Deep learning là một loạt các phương pháp học máy. Nó dựa trên mạng nơ-ron nhân tạo (artifical neural networks). Việc sử dụng nhiều lớp (layer) làm cho quá trình học tập trở nên sâu sắc và đó là nơi bắt nguồn của tính từ "sâu". Mô hình tuần tự là một trong những cách đơn giản nhất để xây dựng một mô hình trong Keras, một mô hình học sâu [6]. Nó được xây dựng trên TensorFlow 2.0 [2]. Nó cung cấp cho chúng ta cách xây dựng mô hình theo từng lớp.

Trong nghiên cứu của mình, chúng tôi cũng đã xây dựng mô hình học sâu tuần tự để dự đoán giá trị mục tiêu theo thông tin dạng văn bản.

 Input Layer

Hidden Layer

Hidden Layer

Output Layer

Hình 1. Deep neural network với 2 hidden layers

**3 Bộ sưu tập Corpus**

Có một số phương pháp thu thập dữ liệu chẳng hạn như API, Chúng tôi đã sử dụng phương pháp quét web để trích xuất dữ liệu từ Google Play Thổ Nhĩ Kỳ. Chúng tôi chọn 112 ứng dụng phổ biến nhất và chúng tôi cũng trích xuất 100 bài đánh giá hữu ích nhất từ mỗi ứng dụng. Bằng cách đó, chúng tôi đạt được để thu thập các đánh giá phù hợp nhất. Chúng tôi chỉ đơn giản là thu thập phần văn bản của các bài đánh giá và xếp hạng của họ.

Chúng tôi loại bỏ các xếp hạng có 3 sao. Chúng tôi sử dụng quy trình dưới đây để gắn nhãn các bài đánh giá theo xếp hạng của họ:

* Positive label cho các bài đánh giá có 4 sao hoặc 5 sao
* Negative label cho các bài đánh giá có 1 sao hoặc 2 sao

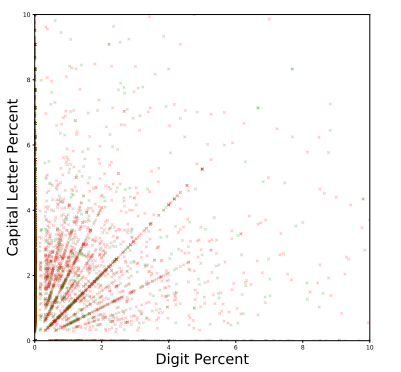
Hai loại dữ liệu được gắn nhãn này được sử dụng để đào tạo một bộ phân loại để dự đoán xem bài đánh giá đã cho có cảm tính tích cực (positive) hay tiêu cực (negative). Trong nghiên cứu của chúng tôi, chúng tôi đặc biệt sử dụng ngôn ngữ Thổ Nhĩ Kỳ. Danh mục của 112 ứng dụng mà chúng tôi đã chọn để trích xuất đánh giá có thể được xem trong Bảng 2 bên dưới.

|  |  |
| --- | --- |
| **Categories** | |
| Art & Design | Dating |
| Augmented Reality | Daydream |
| Auto & Vehicles | Education |
| Beauty | Entertainment |
| Books & Reference | Events |
| Business | Finance |
| Comics | Food & Drink |
| Communication | Health & Fitness |

**Bảng 2**. Danh mục ứng dụng trong Google Play

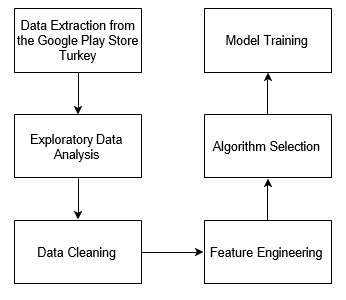
**4 Phân tích Corpus**

Để xây dựng mô hình học máy, chúng tôi kiểm tra dữ liệu của mình và khám phá các tính năng mới như độ dài của bài đánh giá, tỷ lệ phần trăm viết hoa trong bài đánh giá mà chúng tôi có thể sử dụng để phân loại. Trước khi lập mô hình, chúng tôi kiểm tra mối tương quan giữa các tính năng bổ sung và giá trị mục tiêu. Hình 2 dưới đây mô tả mối tương quan giữa phần trăm chữ số và phần trăm chữ in hoa trong các bài đánh giá.



Hình 2. Phân phối giữa phần trăm chữ số và phần trăm chữ in hoa trong các bài đánh giá

Các tính năng khác mà chúng tôi trích xuất từ kho tài liệu của mình là tỷ lệ phần trăm chữ số, cách sử dụng dấu chấm than, độ dài của bài đánh giá và tỷ lệ phần trăm viết hoa. Chúng tôi thấy rằng không có tính năng nào trong số này liên quan đến giá trị mục tiêu. Các giá trị tương quan như trong Hình 2 cho chúng ta thấy không có mối quan hệ nào giữa các tính năng này và giá trị target được sử dụng để phân tích tình cảm. Thiết kế tổng thể của chúng tôi có thể được nhìn thấy trong Hình 3. Như chúng tôi đã đề cập ở trên, chúng tôi đã thu thập hơn 10.000 đánh giá từ Google Play Thổ Nhĩ Kỳ.



Hình 3. Thiết kế tổng thể

Sau khi chúng tôi thu thập dữ liệu, chúng tôi đã khám phá dữ liệu và chúng tôi đã chia sẻ một số phát hiện của chúng tôi như trong Hình 3. Sau các bước này, chúng tôi bắt đầu làm sạch dữ liệu của mình bằng cách áp dụng tất cả các quy trình như sau:

* Thay thế các biểu tượng cảm xúc tương tự bằng một từ khóa xác định
* Loại bỏ dấu chấm câu
* Thay thế một số chữ cái Thổ Nhĩ Kỳ bằng các chữ cái tiếng Anh tương ứng của chúng
* Viết thường văn bản
* Loại bỏ các chữ số
* Loại bỏ các khoảng trắng thừa và dấu chấm câu

Sau các thủ tục này, cũng có các yêu cầu đặc biệt để trích xuất các tính năng từ dữ liệu văn bản do cấu trúc của nó. Văn bản phải được phân tích cú pháp để loại bỏ các token (mã thông báo) và sau đó, các từ phải được encoded (mã hóa) thành các giá trị số như số nguyên hoặc float để được sử dụng trong các mô hình học máy. Để đạt được điều đó, chúng tôi sử dụng thư viện scikit-learning, một thư viện máy học phần mềm miễn phí [20]. Bằng cách sử dụng scikit-learning, chúng tôi đều thực hiện mã hóa và trích xuất tính năng của kho ngữ liệu của chúng tôi.

Chúng tôi đã sử dụng hai loại phương pháp trích xuất tính năng là CountVectorizer và TfidfVectorizer và so sánh kết quả của chúng về hiệu quả với độ chính xác của dự đoán.

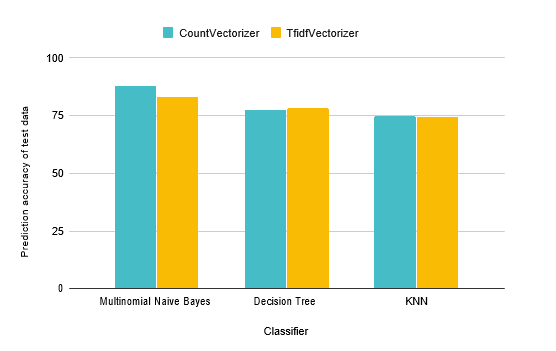
Trong nghiên cứu của mình, chúng tôi đều sử dụng bộ phân loại học máy và mô hình học sâu để lựa chọn thuật toán và đào tạo mô hình. Chúng ta sẽ trình bày chi tiết về các phần lựa chọn thuật toán và đào tạo mô hình trong các phần sau.

**5 Xây dựng Mô hình và Kết quả**

Chúng tôi xây dựng một bộ phân loại bằng cách sử dụng Multinomial Naive Bayes dựa trên định lý Bayes [15]. Chúng tôi cũng xây dựng các mô hình với Decision Tree và K-Nearest neighbors (KNN). Như chúng tôi đã đề cập trước đó, cả hai chúng tôi đều sử dụng CountVectorizer và TfidfVectorizer để thiết kế tính năng và phân tích kết quả.

**5.1 Kết quả cho bộ phân loại học máy (Machine Learning Classifiers)**

Biểu đồ thanh trong Hình 4 cho thấy độ chính xác dự đoán của các bộ phân loại học máy mà chúng tôi đã sử dụng trong nghiên cứu của mình. Có thể thấy trong hình, không có sự khác biệt đáng kể về độ chính xác của dự đoán khi chúng ta sử dụng CountVectorizer hoặc TfidfVectorizer.



Hình 4. Dự đoán độ chính xác của bộ phân loại

Kết quả dự đoán chính xác có thể được nhìn thấy trong Bảng 3. Có thể thấy, trong khi có sự khác biệt khi chúng tôi sử dụng bộ phân loại Multinomial Naive Bayes, không có sự khác biệt khi chúng tôi sử dụng bộ vectơ khác nhau cho decision tree và KNN classifier.

|  |  |  |
| --- | --- | --- |
|  | **CountVectorizer** | **TfidfVectorizer** |
| **MNB** | 87.30% | 83.13% |
| **DT** | 77.93% | 78.09% |
| **KNN** | 74.48% | 74.15% |

Bảng 3. Bảng dự đoán độ chính xác

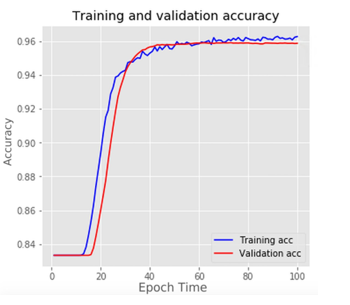
**5.2 Kết quả cho mô hình học sâu (Deep Learning Model)**

Như chúng tôi đã đề cập, chúng tôi cũng đã xây dựng một mô hình học sâu tuần tự. Trong mô hình của chúng tôi, chúng tôi có 2 lớp ẩn và chúng tôi áp dụng tính năng bỏ qua cho 20% số nút để tránh trang bị quá mức trong mô hình của chúng tôi. Hình 1 mô tả mạng học sâu mà chúng tôi đã sử dụng trong nghiên cứu của mình.

|  |  |  |
| --- | --- | --- |
|  | **Training Data** | **Test Data** |
| **CountVectorizer** | 99.73% | 95.71% |
| **TfidfVectorizer** | 95.97% | 95.87% |

Bảng 4. Kết quả dự đoán chính xác cho mô hình học sâu

Như có thể thấy trong Bảng 4, kết quả của chúng tôi chỉ ra rằng chúng tôi có độ chính xác dự đoán 95,87% trên dữ liệu thử nghiệm cho TfidfVectorizer và độ chính xác dự đoán 95,71% trên dữ liệu thử nghiệm cho CountVectorizer. Tuy nhiên, như có thể thấy trong Hình 5 và Hình 6, mặc dù chúng ta chính xác sử dụng cùng một mô hình cho cả hai, có thể overfitting khi chúng ta sử dụng CountVectorizer. Ngoài ra, khi chúng ta nhìn vào kết quả trong Bảng 4, có thể thấy rằng có overfitting trong mô hình. Trong khi độ chính xác của dự đoán là 99,73% đối với dữ liệu đào tạo, độ chính xác của dữ liệu thử nghiệm thấp hơn dữ liệu đào tạo gần 5%. Điều này có thể cung cấp cho chúng tôi một manh mối rằng có thể có overfitting khi chúng tôi sử dụng cùng một mô hình với CountVectorizer. Để loại bỏ overfitting, chúng tôi có thể cần phải tối ưu hóa các thông số của mô hình.



Hình 5. Đào tạo và xác nhận độ chính xác cho TfidfVectorizer



Hình 6. Đào tạo và xác nhận độ chính xác cho CountVectorizer

**6. Kết luận**

Dễ thấy rằng phân tích cảm tính và khai thác ý kiến ​​đóng một vai trò quan trọng để theo dõi hành vi của người tiêu dùng. Với những tiến bộ gần đây trong kỹ thuật học máy, vấn đề này đã được giải quyết và đi một chặng đường dài bằng tiếng Anh. Tuy nhiên, trong các ngôn ngữ phổ biến như tiếng Thổ Nhĩ Kỳ, nó vẫn là một trong những chủ đề nóng. Trong nghiên cứu này, chúng tôi so sánh một số phương pháp phân loại và phương pháp học sâu để đưa ra phân tích tình cảm đối với các bài đánh giá của Thổ Nhĩ Kỳ mà chúng tôi đã thu thập được từ Google Play. Chúng tôi đã giải thích các bộ phân loại học máy mà chúng tôi đã sử dụng và chúng tôi đã mô tả thiết kế tổng thể của mình. Chúng tôi có kết quả đáng kể cả từ bộ phân loại học máy và mô hình học sâu. Kết quả dự đoán chính xác của chúng tôi là khả quan. Chúng tôi có độ chính xác dự đoán 87,30% cho Multinomial Naive Bayes và độ chính xác dự đoán 95,87% cho mô hình học sâu. Các thí nghiệm của chúng tôi cho thấy rằng chúng tôi có thể bị thừa khi sử dụng các kỹ thuật vectơ khác nhau. Mặc dù không có sự khác biệt giữa các bộ phân loại học tập khi chúng ta sử dụng các công cụ vector hóa khác nhau, nhưng có sự khác biệt khi chúng ta xây dựng mô hình học sâu để dự đoán giá trị mục tiêu. Mặc dù chúng tôi xây dựng mô hình của mình từ các bài đánh giá mà chúng tôi đã trích xuất từ ​​Google Play, nhưng mô hình này cũng có thể được áp dụng cho bất kỳ dữ liệu nào từ Twitter, Facebook hoặc bất kỳ blog nhỏ nào khác.

**Tài liệu tham khảo**

1. https://github.com/ssari-memory/corpus-turkish-reviews

2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Man´e, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Vi´egas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org

3. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric re- gression. The American Statistician 46(3), 175–185 (1992)

4. Anjaria, M., Guddeti, R.M.R.: Influence factor based opinion mining of twitter data using supervised learning. In: 2014 Sixth International Conference on Com- munication Systems and Networks (COMSNETS). pp. 1–8. IEEE (2014)

5. Bouazizi, M., Ohtsuki, T.: Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. pp. 1594– 1597 (2015)

6. Chollet, F., et al.: Keras. https://keras.io (2015)

7. C¸ oban, O¨ ., O¨ zyer, B., O¨ zyer, G.: Tu¨rk¸ce twitter mesajlarının duygu analizi. In: Signal Processing and Communications Applications Conference (SIU) (2015)

8. Demirtas, E., Pechenizkiy, M.: Cross-lingual polarity detection with machine trans- lation. In: Proceedings of the Second International Workshop on Issues of Senti- ment Discovery and Opinion Mining. pp. 1–8 (2013)

9. Gezici, G., Yanıkoglu, B.: Sentiment analysis in turkish. In: Turkish Natural Lan- guage Processing, pp. 255–271. Springer (2018)

10. Ghorbel, H., Jacot, D.: Sentiment analysis of french movie reviews. In: Advances in Distributed Agent-Based Retrieval Tools, pp. 97–108. Springer (2011)

11.Gordon,K.: Topic: Online reviews, https://www.statista.com/topics/4381/online- reviews/

12. Harris, Z.S.: Distributional structure. Word 10(2-3), 146–162 (1954)

13. Kaya, M., Fidan, G., Toroslu, I.H.: Sentiment analysis of turkish political news. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and In- telligent Agent Technology. vol. 1, pp. 174–180. IEEE (2012)

14. Khan, F.H., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision support systems 57, 245–257 (2014)

15. Klon, A.E., Glick, M., Davies, J.W.: Combination of a naive bayes classifier with consensus scoring improves enrichment of high-throughput docking results. Journal of medicinal chemistry 47(18), 4356–4359 (2004)

16. Ku¨¸cu¨k, D., Steinberger, R.: Experiments to improve named entity recognition on turkish tweets. arXiv preprint arXiv:1410.8668 (2014)

17. Maron, M.E.: Automatic indexing: an experimental inquiry. Journal of the ACM (JACM) 8(3), 404–417 (1961)

18. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10, pp. 1320–1326 (2010)

19. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empir- ical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)

20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

21. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 616–623 (2003)

22. Rokach, L., Maimon, O.Z.: Data mining with decision trees: theory and applica- tions, vol. 69. World scientific (2008)

23. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)

24. Vural, A.G., Cambazoglu, B.B., Senkul, P., Tokgoz, Z.O.: A framework for sen- timent analysis in turkish: Application to polarity detection of movie reviews in turkish. In: Computer and Information Sciences III, pp. 437–445. Springer (2013)

25. Yıldırım, E., C¸ etin, F.S., Eryi˘git, G., Temel, T.: The impact of nlp on turkish sen-timent analysis. Tu¨rkiye Bili¸sim Vakfı Bilgisayar Bilimleri ve Mu¨hendisli˘gi Dergisi 7(1), 43–51 (2015)