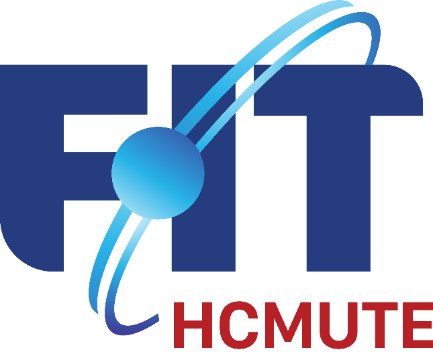


BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM

KHOA CÔNG NGHỆ THÔNG TIN



Đề Tài

Phân tích khám phá dữ liệu

và dự đoán sự thành công của ứng dụng

trên Google Play Store

Tp. Hồ Chí Minh, tháng 4 năm 2024

GVHD: Ths. Lê Minh Tân

Nhóm 1:

Nguyễn Trung Phiên 21110593

Sú Minh Luân 21110900

Hoàng Mai Hiếu 21110882

Lớp thứ 6 – Tiết 1-4

INDS331085\_23\_2\_01

ĐIỂM SỐ

|  |  |  |  |
| --- | --- | --- | --- |
| TIÊU CHÍ | NỘI DUNG | TRÌNH BÀY | TỔNG |
| ĐIỂM |  |  |  |

NHẬN XÉT

..................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

*Ký tên*

Ths. Lê Minh Tân

**LỜI CẢM ƠN**

Trong quá trình tìm hiểu nghiên cứu thực hiện đề tài tiểu luận môn Khoa Học Dữ Liệu, nhóm chúng em đã nhận được sự hỗ trợ nhiệt tình chỉ bảo tận tâm từ rất nhiều người, từ bạn bè, anh chị và thầy cô. Vì thế nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến với các cá nhân và tổ chức sau đây: Đầu tiên, chúng em xin cảm ơn thầy Lê Minh Tân – người hướng dẫn, giảng dạy của môn học này đã cung cấp cho chúng em kiến thức cơ bản lẫn cả nâng cao trong lĩnh vực Dữ Liệu và cũng đã hỗ trợ chúng em vô cùng nhiệt tình trong quá trình thực hiện đề tài. Chúng em cũng muốn bày tỏ lòng biết ơn của mình đến các nhà nghiên cứu, các chuyên gia trong lĩnh vực này, những người đã cống hiến cuộc đời mình, thanh xuân mình để tạo ra các công nghệ và thuật toán đột phá, thay đổi nhiều góc nhìn khác nhau trong Khoa Học Dữ Liệu. Ngoài ra, chúng em cũng không thể quên được sự hỗ trợ của những người bạn, anh chị xung quanh trong quá trình học tập và thực hiện đề tài này. Cuối cùng, chúng em xin chân thành cảm ơn Trường Đại Học Sư Phạm Kỹ Thuật Thành Phố Hồ Chí Minh và các nhà tài trợ, các công ty công nghệ đã cung cấp điều kiện và nguồn lực để chúng em có thể hoàn thành tiểu luận này một cách chỉnh chu nhất. Một lần nữa, chúng em xin gửi lời cảm ơn sâu sắc và mong rằng những đóng góp, nghiên cứu của chúng em sẽ mang lại giá trị tốt đẹp, hữu ích cho cộng đồng khoa học và công nghệ.

**DANH MỤC BẢNG**

[Bảng 1: Top 10 thể loại có số lượng ứng dụng vượt trội 13](#_Toc165327150)

[Bảng 2: Mối tương quan giữa số lượt cài đặt và các đặc trưng số khác 19](#_Toc165327151)

[Bảng 3: Những reviews phản ánh sự hài lòng trong quá khứ 37](#_Toc165327152)

[Bảng 4: Những giá trị sentiment không khớp với Review 39](#_Toc165327153)

[Bảng 5: Các reviews tích cực nhưng có rating thấp 41](#_Toc165327154)

[Bảng 6: Các reviews có nội dung chứa toàn chữ in hoa 42](#_Toc165327155)

[Bảng 7: Các reviews có chứa câu hỏi 42](#_Toc165327156)

[Bảng 8: Các reviews có chứa dấu chấm than “!” 42](#_Toc165327157)

[Bảng 9: Các giá trị đánh giá mô hình SVM trên tập train 49](#_Toc165327158)

[Bảng 10: Các giá trị đánh giá mô hình SVM trên tập test 50](#_Toc165327159)

[Bảng 11: Các giá trị đánh giá mô hình Gradient Boosting trên tập train 51](#_Toc165327160)

[Bảng 12: Các giá trị đánh giá mô hình Gradient Boosting trên tập test 51](#_Toc165327161)

[Bảng 13: Các giá trị đánh giá mô hình Random Forest trên tập train 52](#_Toc165327162)

[Bảng 14: Các giá trị đánh giá mô hình Random Forest trên tập test 53](#_Toc165327163)

**DANH MỤC HÌNH**

[Hình 1: Confusion matrix 8](#_Toc165327195)

[Hình 2: Biểu đồ số lượng ứng dụng theo thể loại trước khi gộp thể loại game 11](#_Toc165327196)

[Hình 3: Biểu đồ số lượng ứng dụng theo thể loại sau khi gộp thể loại Game 12](#_Toc165327197)

[Hình 4: Biểu đồ số lượt cài đặt của 10 thể loại phổ biến nhất về số lượng ứng dụng 14](#_Toc165327198)

[Hình 5: Biểu đồ số lượt cài đặt của tất cả thể loại 15](#_Toc165327199)

[Hình 6: Sự tăng trưởng lượt tải của 3 thể loại game phổ biến đầu Quý 1 2020 so với Quý 4 2019 16](#_Toc165327200)

[Hình 7: Chi tiêu tiêu dùng toàn cầu cho ứng dụng di động năm 2020 17](#_Toc165327201)

[Hình 8: Lượng cài đặt ứng dụng trên toàn cầu năm 2020 18](#_Toc165327202)

[Hình 9: Chi tiêu tiêu dùng toàn cầu cho ứng dụng Game năm 2020 18](#_Toc165327203)

[Hình 10: Lượng cài đặt ứng dụng Game trên toàn cầu năm 2020 18](#_Toc165327204)

[Hình 11: Biểu đồ boxplot về lượt cài đặt theo thể loại 19](#_Toc165327205)

[Hình 12: Bản đồ nhiệt của mối tương quan giữa các đặc trưng 20](#_Toc165327206)

[Hình 13: Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 2 từ 21](#_Toc165327207)

[Hình 14: Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 3 từ 21](#_Toc165327208)

[Hình 15: Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 4 từ 22](#_Toc165327209)

[Hình 16: Biểu đồ số lượng ứng dụng theo xếp hạng nội dung 23](#_Toc165327210)

[Hình 17: Biểu đồ số lượt cài đặt theo xếp hạng nội dung 23](#_Toc165327211)

[Hình 18: Biểu đồ loại ứng dụng theo số lượng ứng dụng 24](#_Toc165327212)

[Hình 19: Biểu đồ loại ứng dụng theo số lượt cài đặt 25](#_Toc165327213)

[Hình 20: Biểu đồ loại ứng dụng của 4 thể loại phổ biến nhất 25](#_Toc165327214)

[Hình 21: Biểu đồ tròn về tỉ lệ của 4 thể loại phổ biến nhất 26](#_Toc165327215)

[Hình 22: Biểu đồ top 10 nhà phát triển ứng dụng có tổng lượt cài đặt cao nhất 27](#_Toc165327216)

[Hình 23: Biểu đồ số lượng ứng dụng của top 10 nhà phát triển có lượt cài đặt cao nhất 28](#_Toc165327217)

[Hình 24: Biểu đồ top 10 nhà phát triển có số lượng ứng dụng nhiều nhất 28](#_Toc165327218)

[Hình 25: Biểu đồ boxplot của rating theo thể loại 29](#_Toc165327219)

[Hình 26: Biểu đồ word cloud về các từ xuất hiện nhiều nhất trong các reviews 33](#_Toc165327220)

[Hình 27: Thống kê của GOOGLE PLAY STORE về tỷ lệ số lượng ứng dụng thuộc thể loại game cập nhật đến ngày 24/4/2024 34](#_Toc165327221)

[Hình 28: Phân phối tần suất của các từ ứng với rating 1 34](#_Toc165327222)

[Hình 29: Phân phối tần suất của các từ ứng với rating 2 35](#_Toc165327223)

[Hình 30: Phân phối tần suất của các từ ứng với rating 3 35](#_Toc165327224)

[Hình 31: Phân phối tần suất của các từ ứng với rating 4 36](#_Toc165327225)

[Hình 32: Phân phối tần suất của các từ ứng với rating 5 36](#_Toc165327226)

[Hình 33: Hiện tượng bão 1\* trên trang đánh giá của game Gensin Impact 38](#_Toc165327227)

[Hình 34: Game Genshin Impact ở trang sản phẩm của GOOGLE PLAY STORE 38](#_Toc165327228)

[Hình 35: Biểu đồ box plot về phân phối của sentiment theo rating 40](#_Toc165327229)

[Hình 36: Rating trung bình của các review chứa toàn ký tự in hoa so với các review còn lại 45](#_Toc165327230)

[Hình 37: Rating trung bình của các review chứa dấu chấm than so với các review còn lại 45](#_Toc165327231)

[Hình 38: Rating trung bình của các review chứa câu hỏi so với các review còn lại 45](#_Toc165327232)

[Hình 39: Công thức đánh nhãn sự thành công (1) 47](#_Toc165327233)

[Hình 40: Các trường hợp ngoại lệ cho công thức (1) 47](#_Toc165327234)

[Hình 41: Công thức đánh nhãn sự thành công (2) 47](#_Toc165327235)

[Hình 42: Công thức cho đô 48](#_Toc165327236)

[Hình 43: Confusion matrix cho mô hình SVM trên tập train. 49](#_Toc165327237)

[Hình 44: Confusion matrix cho mô hình SVM trên tập test. 50](#_Toc165327238)

[Hình 45: Confusion matrix cho mô hình Gradient trên tập train. 50](#_Toc165327239)

[Hình 46: Confusion matrix cho mô hình Gradient Boosting trên tập test. 51](#_Toc165327240)

[Hình 47: Confusion matrix cho mô hình Random Forest trên tập train. 52](#_Toc165327241)

[Hình 48: Confusion matrix cho mô hình Random Forest trên tập test. 53](#_Toc165327242)

[Hình 49: Đánh giá các mô hình trên tập train 54](#_Toc165327243)

[Hình 50: Đánh giá các mô hình trên tập test 54](#_Toc165327244)

MỤC LỤC

[PHẦN 1: GIỚI THIỆU ĐỀ TÀI 1](#_Toc165327251)

[**1.1.** **Phạm vi dữ liệu** 1](#_Toc165327252)

[**1.2.** **Công cụ và giao thức thu thập dữ liệu** 2](#_Toc165327253)

[**1.3.** **Các yếu tố ảnh hưởng đến việc thu thập và sử dụng dữ liệu** 2](#_Toc165327254)

[**1.4.** **Bộ dữ liệu** 2](#_Toc165327255)

[**1.5.** **Phương pháp thực hiện** 2](#_Toc165327256)

[PHẦN 2: CƠ SỞ LÝ THUYẾT 3](#_Toc165327257)

[**2.1.** **Phân tích cảm xúc** 3](#_Toc165327258)

[**2.2.** **Kiểm định giả thuyết và ước lượng khoảng tin cậy cho trung bình** 3](#_Toc165327259)

[**2.2.1.** **Kiểm định giả thuyết** 3](#_Toc165327260)

[**2.2.2.** **Ước lượng khoảng tin cậy cho trung bình** 4](#_Toc165327261)

[**2.3.** **Các mô hình học máy** 5](#_Toc165327262)

[**2.3.1.** **SVM** 5](#_Toc165327263)

[**2.3.2.** **Gradient boosting** 6](#_Toc165327264)

[**2.3.3.** **Random forest** 6](#_Toc165327265)

[**2.4.** **Các phép đo đánh giá mô hình phân loại** 7](#_Toc165327266)

[**2.4.1.** **Confusion matrix** 7](#_Toc165327267)

[**2.4.1.** **Accuracy** 8](#_Toc165327268)

[**2.4.2.** **Precision** 8](#_Toc165327269)

[**2.4.3.** **Recall** 8](#_Toc165327270)

[**2.4.4.** **F1-Score** 9](#_Toc165327271)

[PHẦN 3: TIỀN XỬ LÝ DỮ LIỆU 10](#_Toc165327272)

[PHẦN 4: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU 11](#_Toc165327273)

[**4.1.** **Phân tích thông tin ứng dụng** 11](#_Toc165327274)

[**4.1.1.** **Thể loại và ứng dụng** 11](#_Toc165327275)

[**4.1.2.** **Thể loại và lượt cài đặt** 13](#_Toc165327276)

[**4.1.3.** **Tương quan số lượt cài đặt** 19](#_Toc165327277)

[**4.1.4.** **Tên ứng dụng** 20](#_Toc165327278)

[**4.1.5.** **Xếp hạng nội dung** 22](#_Toc165327279)

[**4.1.6.** **Loại ứng dụng** 24](#_Toc165327280)

[**4.1.7.** **Lượt cài đặt và nhà phát triển** 26](#_Toc165327281)

[**4.1.8.** **Ứng dụng và nhà phát triển** 27](#_Toc165327282)

[**4.1.9.** **Thể loại và rating** 29](#_Toc165327283)

[**4.1.10.** **Kiểm định giả thuyết** 30](#_Toc165327284)

[**4.2.** **Phân tích Review và rating** 32](#_Toc165327285)

[**4.2.1.** **Từ điển từ** 32](#_Toc165327286)

[**4.2.2.** **Phân tích từ điển từ** 33](#_Toc165327287)

[**4.2.3.** **Hiện tượng không khớp giữa các trường dữ liệu** 38](#_Toc165327288)

[**4.2.4.** **Kiểm định giả thiết** 41](#_Toc165327289)

[PHẦN 5: MÔ HÌNH DỰ ĐOÁN SỰ THÀNH CÔNG CỦA ỨNG DỤNG 47](#_Toc165327290)

[**5.1.** **Định nghĩa sự thành công của ứng dụng.** 47](#_Toc165327291)

[**5.2.** **Super vector machine (SVM).** 48](#_Toc165327292)

[**5.3.** **Gradient Boosting** 50](#_Toc165327293)

[**5.4.** **Random Forest.** 51](#_Toc165327294)

[**5.5.** **Đánh giá các mô hình và lựa chọn mô hình** 54](#_Toc165327295)

[PHẦN 6: KẾT LUẬN 56](#_Toc165327296)

[TÀI LIỆU THAM KHẢO 57](#_Toc165327297)

[ZEPPELIN REPORT 58](#_Toc165327298)

# **PHẦN 1: GIỚI THIỆU ĐỀ TÀI**

Nền tảng phân phối ứng dụng di động lớn như Google Play Store có đến hàng tỷ người dùng và hàng triệu ứng dụng trong khi có đến hàng ngàn ứng dụng mới được thêm vào mỗi ngày. Công việc phát triển ứng dụng di động đang chứng kiến sự cạnh tranh khốc liệt từ khắp nơi trên thế giới, việc những nhà phát triển ứng dụng di động có cái nhìn đúng đắn về hướng đi của mình để xây dựng được những ứng dụng đáp ứng được yêu cầu của người dùng trở nên vô cùng quan trọng. Vì hầu hết ứng dụng trên Play Store đều miễn phí, mô hình thu nhập thường không được làm rõ và thiếu thông tin chi tiết về cách mua hàng, quảng cáo trong ứng dụng,… Do đó, thành công của một ứng dụng thường được phản ánh chủ yếu dựa vào số lượng lượt tải xuống và các ratings/ reviews từ người dùng trong quãng thời gian nó tồn tại trên play store, chứ không chỉ dựa vào doanh thu mà ứng dụng đó tạo ra.

Có nhiều câu hỏi đặt ra như: Các đặc trưng khác nhau ảnh hưởng đến sự thành công của một ứng dụng như thế nào? Điều gì là yếu tố quyết định sự thành công của một ứng dụng, là ratings hay số lượt tải xuống? Thể loại ứng dụng nào có tiềm năng lớn đối với những nhà phát triển ứng dụng di động?. Để trả lời tất cả những câu hỏi trên, trong dự án này chúng tôi hướng tới việc phân tích khám phá dữ liệu về các ứng dụng thu thập từ Google Play Store để xác định các đặc trưng cần thiết cho sự thành công của một ứng dụng trong các danh mục khác nhau, kết hợp các đặc trưng này với cảm xúc của người dùng về ứng dụng để xây dựng các mô hình dự đoán sự thành công của một ứng dụng ngay sau khi nó được ra mắt trên Google Play Store.

* 1. **Phạm vi dữ liệu**
* Target population: Tất cả các ứng dụng có sẵn trên Google Play Store.
* Access frame: Dữ liệu lấy về bằng phương pháp web-scraping.
* Sample: Lọc ra các ứng dụng trong access frame mà có ít nhất n reviews/comments cho mỗi ứng dụng và tiến hành lấy ngẫu nhiên m ứng dụng từ bộ dữ liệu được lọc ra. (n, m là số lượng mà ta sẽ điều chỉnh tùy thuộc vào Access frame và bộ dữ liệu được lọc ra).
  1. **Công cụ và giao thức thu thập dữ liệu**
* Instruments: Sử dụng phương pháp web-scraping để thu thập dữ liệu về các ứng dụng trên google play store thông qua thư viện Google-Play-Scraper.
* Protocols: Thu thập url của các ứng dụng, sau đó Scraper sử dụng các url này và thực hiện lặp để trích xuất thông tin của mỗi ứng dụng và lưu tất cả thông tin vào một tệp CSV để sử dụng cho phân tích sau này.
  1. **Các yếu tố ảnh hưởng đến việc thu thập và sử dụng dữ liệu**
* Các ứng dụng mới liên tục được thêm vào, và các ứng dụng hiện tại được cập nhật, xếp hạng và đánh giá liên tục trên Google Play Store. Nên dữ liệu sử dụng cho dự án có thể nhanh chóng bị lỗi thời.
* Google Play có các giới hạn về tốc độ truy cập để ngăn chặn web scraping quá mức. Giới hạn này có thể gây chặn hoặc giảm hiệu suất của scraper khi thu thập dữ liệu.
* Việc thu thập thông tin có thể bị hạn chế bởi chính sách về quyền riêng tư của Google. Điều này đặc biệt quan trọng khi thu thập dữ liệu liên quan đến người dùng hoặc thông tin cá nhân.
* Không phải tất cả thông tin trên Google Play đều là độ tin cậy cao. Đặc biệt là với đánh giá và phản hồi người dùng, có thể có các đánh giá giả mạo hoặc không chính xác.
  1. **Bộ dữ liệu**

**New Google Play Store - Android Apps dataset**

* Tác giả: TUNG M PHUNG
* Link tải dữ liệu: [New Google Play Store - Android Apps dataset (kaggle.com)](https://www.kaggle.com/datasets/tungmphung/new-google-play-store-android-apps-dataset?select=app.csv)
  1. **Phương pháp thực hiện**

B1: Thu thập dữ liệu

B2: Tiền xử lý dữ liệu

B3: Phân tích khám phá dữ liệu để tìm ra insight bên trong dữ liệu về các đặc trưng ảnh hưởng đến sự thành công của 1 ứng dụng và tiến hành phân tích cảm xúc người dùng dựa trên các comments của mỗi ứng dụng.

B4: Kết hợp các đặc trưng quan trọng ảnh hưởng đến sự thành công với cảm xúc của người dùng về mỗi ứng dụng để xây dựng các model machine learning cho nhiệm vụ dự đoán sự thành công của ứng dụng.

B5: Đưa ra so sánh, kết luận và viết báo cáo.

# **PHẦN 2: CƠ SỞ LÝ THUYẾT**

1. **Phân tích cảm xúc**

Phân tích cảm xúc (Sentiment Analysis) là hoạt động diễn dịch và phân loại tự động các cảm xúc (tích cực, tiêu cực hoặc trung tính) từ dữ liệu văn bản như các bài đánh giá bằng chữ, các bài đăng trên mạng xã hội. Có thể xem phân tích cảm xúc là mô hình AI (AI model) được tạo ra từ kỹ thuật máy học. Nó có khả năng phân tích ngôn ngữ và nhận diện cảm xúc. AI model sẽ được “huấn luyện” dựa trên một bộ dữ liệu văn bản được dán nhãn. Cụ thể là gắn nhãn cụm từ / từ khoá với trạng thái cảm xúc tương ứng. Là tích cực, tiêu cực hay bình thường? Sau đó, AI model sẽ sử dụng những dữ liệu gắn nhãn này để học các cấu trúc văn bản kết hợp với mỗi trạng thái cảm xúc khác nhau.

AI model phân tích cảm xúc sẽ phân tích các đoạn văn bản mới và đưa ra dự đoán. Bằng cách nhận diện chính xác cụm từ cảm xúc trong văn bản, sau đó tính toán và đưa ra tổng điểm cảm xúc của đoạn văn đó. Trong báo cáo này chúng tôi sử dụng công cụ TextBlob API để dự đoán giá trị cảm xúc từ bình luận, sau đó sử dụng giá trị cảm xúc này cho việc phân tích sâu và khám phá những câu chuyện đằng sau review và rating của người dùng.

1. **Kiểm định giả thuyết và ước lượng khoảng tin cậy cho trung bình**
2. **Kiểm định giả thuyết**

Trong kiểm định giả thuyết thống kê, p-value là một công cụ quan trọng giúp chúng ta quyết định liệu có bác bỏ giả thuyết null (H0) hay không.

Giả thuyết Null (H0) và đối thuyết (H1): Trước khi tiến hành kiểm định, chúng ta cần đặt ra hai giả thuyết. Giả thuyết null thường đại diện cho tình huống "không có gì đặc biệt xảy ra" hoặc "không có hiệu ứng". Đối thuyết đại diện cho hiệu ứng mà chúng ta mong đợi hoặc muốn kiểm tra.

**Tính p-value**: p-value là xác suất của việc thu được kết quả kiểm định ít nhất bằng với giá trị thực tế được quan sát (hoặc cực đoan hơn) dưới giả định là giả thuyết null đúng. Nói cách khác, p-value cho chúng ta biết rằng nếu H0 đúng, có bao nhiêu khả năng thu được kết quả như chúng ta đã quan sát hoặc khả năng còn cao hơn.

**Mức Ý nghĩa (α):** Đây là ngưỡng mà chúng ta chọn để quyết định việc bác bỏ H0. Mức ý nghĩa thường được chọn là 0.05, nghĩa là có 5% khả năng bác bỏ giả thuyết null khi nó thực sự đúng (lỗi loại I).

**Quyết định:** Dựa vào p-value và mức ý nghĩa α:

* Nếu p-value ≤ α, chúng ta có cơ sở để bác bỏ giả thuyết null. Điều này cho thấy có bằng chứng đủ mạnh để ủng hộ đối thuyết.
* Nếu p-value > α, chúng ta không bác bỏ giả thuyết null. Điều này không đồng nghĩa với việc chứng minh được H0 đúng mà chỉ là không có đủ bằng chứng để bác bỏ nó.

Sử dụng kiểm định ANOVA test: hay còn được gọi là phân tích phương sai (Analysis of Variance), là một kỹ thuật thống kê tham số được sử dụng để phân tích sự khác nhau giữa giá trị trung bình của các biến phụ thuộc với nhau.

ANOVA one-way là một phương pháp thống kê được sử dụng để so sánh ba hoặc nhiều nhóm trên một biến phụ thuộc. Nó kiểm tra xem có sự khác biệt đáng kể nào giữa các nhóm hay không.

Giả thuyết null trong ANOVA one-way là **không có sự khác biệt đáng kể nào giữa các nhóm**. Cụ thể, giả thuyết null có thể được diễn đạt như sau:

: Không có sự khác biệt đáng kể giữa các nhóm trên biến phụ thuộc.

Giả thuyết này cho rằng trung bình của tất cả các nhóm là bằng nhau. Đối thuyết của giả thuyết null là giả thuyết thay thế, mà cho rằng ít nhất một cặp trung bình của các nhóm là khác nhau.

Khi thực hiện kiểm định ANOVA one-way, chúng ta tính toán một giá trị thống kê được gọi là F-statistic, được tính bằng tỉ lệ

trong đó: MSB: phương sai trung bình giữa các nhóm, MSW: phương sai trong nhóm.

Khi đó chúng ta có thể tính giá trị p-value, nếu giả thuyết H0 đúng thì xác suất F=F-statistic là bao nhiêu và so sánh với mức ý nghĩa α để bác bỏ hay chấp nhận giả thuyết .

1. **Ước lượng khoảng tin cậy cho trung bình**

Uớc lượng khoảng tin cậy cho trung bình là xác định một phạm vi mà chúng ta tin rằng giá trị trung bình thực sự của tổng thể có khả năng rơi vào.

**Khoảng Tin Cậy (CI)**: Khoảng tin cậy là khoảng giá trị được xác định dựa trên dữ liệu mẫu và một mức độ tin cậy cho trước.

**Mức Độ Tin Cậy ()**: thường là 95% hoặc 99%, là xác suất mà khoảng tin cậy chứa tham số tổng thể.

Khi mẫu lớn ( 30) và phương sai tổng thể chưa biết, phân phối z thường được sử dụng, công thức khoảng tin cậy của trung bình tổng thể sẽ là:

Trong đó:

là trung bình mẫu.

α : mức ý nghĩa ( α = 1 - )

là giá trị z tại (tra theo bảng phân phối z để tính)

𝑠 là độ lệch chuẩn mẫu.

là kích thước mẫu.

1. **Các mô hình học máy**
2. **SVM**

Mô hình Máy Vector Hỗ trợ (Support Vector Machine - SVM) là một phương pháp học máy giám sát được sử dụng chủ yếu trong bài toán phân loại và hồi quy.

**Phân loại biên cứng (Hard Margin Classification):** Trong bài toán phân loại nhị phân, SVM tìm một siêu mặt phẳng (hyperplane) sao cho khoảng cách (margin) đến điểm dữ liệu gần nhất từ mỗi lớp là lớn nhất. Siêu mặt phẳng này chính là biên quyết định để phân tách hai lớp.

**Vector Hỗ trợ (Support Vectors):** Là những điểm dữ liệu gần nhất với siêu mặt phẳng, đóng vai trò quan trọng trong việc xác định vị trí và độ nghiêng của siêu mặt phẳng.

**Phân loại biên mềm (Soft Margin Classification):** Để tăng cường khả năng tổng quát hóa và xử lý dữ liệu không tách biệt tuyến tính, SVM sử dụng phương pháp biên mềm cho phép một số điểm dữ liệu vi phạm margin. Điều này được thực hiện bằng cách đưa vào tham số hòa nhã (regularization parameter) C để kiểm soát độ lớn của vi phạm.

**Hạt nhân (Kernel Trick):** SVM có thể xử lý dữ liệu không tách biệt tuyến tính bằng cách áp dụng các hàm hạt nhân, chẳng hạn như hạt nhân đa thức (polynomial kernel), hạt nhân RBF (Radial Basis Function), để chuyển dữ liệu sang không gian đa chiều nơi có thể tìm được siêu mặt phẳng phân loại.

**Tối ưu hóa:** Việc tìm siêu mặt phẳng tối ưu là một bài toán tối ưu hóa lồi, nơi hàm mục tiêu là tối đa hóa margin và ràng buộc là các điểm dữ liệu phải nằm đúng về phía của lớp của chúng.

**Đa lớp:** Mặc dù SVM truyền thống được thiết kế cho bài toán phân loại nhị phân, nhưng nó có thể mở rộng để giải quyết các bài toán phân loại đa lớp thông qua kỹ thuật như one-vs-rest (OvR) hoặc one-vs-one (OvO).

Các mô hình SVM phức tạp và mạnh mẽ, được ưa chuộng trong các ứng dụng có dữ liệu có chiều cao, như nhận dạng hình ảnh và phân loại văn bản, vì chúng có khả năng phân loại dữ liệu phức tạp và không tách biệt tuyến tính mà không cần phải thêm vào nhiều đặc trưng thủ công.

1. **Gradient boosting**

Mô hình Gradient Boosting là một kỹ thuật mạnh mẽ của học máy giám sát được sử dụng trong cả bài toán phân loại và hồi quy. Nó hoạt động dựa trên nguyên lý tăng cường (boosting) một loạt các mô hình yếu (thường là cây quyết định) để tạo nên một mô hình dự đoán mạnh mẽ.

**Mô hình yếu (Weak Learners)**: Gradient Boosting kết hợp nhiều mô hình yếu, thường là cây quyết định, trong đó mỗi mô hình cố gắng khắc phục những lỗi của mô hình trước đó.

**Cải thiện lần lượt (Sequential Improvement):** Mỗi mô hình mới được xây dựng để cải thiện lỗi dự đoán của tổ hợp mô hình đã có. Điều này đạt được bằng cách áp dụng gradient descent vào không gian hàm số để tối thiểu hóa hàm mất mát.

**Hàm mất mát (Loss Function):** Một hàm mất mát được xác định để đo lường sai số giữa dự đoán và thực tế, ví dụ như sai số bình phương cho hồi quy hay độ mất mát logarit cho phân loại.

**Tối ưu hóa dựa trên Gradient:** Đối với mỗi bước, mô hình được tối ưu hóa thông qua việc tính gradient của hàm mất mát để xác định hướng và bước điều chỉnh cho mô hình kế tiếp.

**Tốc độ học (Learning Rate):** Một tham số tốc độ học nhỏ (𝜂η) được sử dụng để kiểm soát mức độ mỗi cây quyết định góp phần vào mô hình cuối cùng, giúp mô hình hội tụ và tránh overfitting.

**Số lượng cây (Number of Trees):** Số lượng các cây quyết định (được gọi là 'estimators' trong thuật ngữ của Gradient Boosting) cần được cân nhắc cẩn thận để đạt được hiệu suất tốt nhất mà không gây overfitting.

**Regularization:** Để ngăn chặn việc quá khớp dữ liệu, Gradient Boosting có thể áp dụng các phương pháp điều tiết như shrinkage (sử dụng một tốc độ học nhỏ) và subsampling (stochastic gradient boosting) để cải thiện khả năng tổng quát hóa.

Gradient Boosting là một kỹ thuật mạnh mẽ và linh hoạt, nhưng nó cũng đòi hỏi sự chú ý đến việc tinh chỉnh tham số để tránh overfitting và đảm bảo mô hình hội tụ tốt. Nó đã được chứng minh là rất hiệu quả trong nhiều tình huống thực tế từ dự đoán tài chính đến phân loại hình ảnh.

1. **Random forest**

Mô hình Random Forest là một thuật toán học máy giám sát, được sử dụng rộng rãi trong cả bài toán phân loại và hồi quy. Nó là một ví dụ của "Ensemble Learning", một kỹ thuật mà kết hợp nhiều mô hình dự đoán (thường là cây quyết định) để tạo ra một mô hình tổng hợp mạnh mẽ hơn.

**Ensemble of Decision Trees:** Random Forest xây dựng nhiều cây quyết định độc lập với nhau. Mỗi cây được huấn luyện trên một mẫu ngẫu nhiên của dữ liệu (bootstrap sample), và trong quá trình phân chia một nút, một tập con ngẫu nhiên của các đặc trưng được sử dụng (feature bagging).

**Bagging (Bootstrap Aggregating):** Mỗi cây trong rừng được xây dựng từ một mẫu có hoàn lại của dữ liệu huấn luyện (bootstrap sample), giúp tăng độ đa dạng của các mô hình và giảm overfitting.

**Feature Randomness:** Khi tạo các điểm phân chia trong cây, Random Forest chọn một tập con ngẫu nhiên của các đặc trưng thay vì sử dụng toàn bộ đặc trưng. Điều này giúp giảm sự phụ thuộc giữa các cây, từ đó tăng cường hiệu suất của mô hình tổng thể.

**Prediction Aggregation:** Để đưa ra dự đoán, Random Forest tổng hợp kết quả từ tất cả các cây. Đối với bài toán phân loại, kỹ thuật "voting" (bình chọn) được sử dụng, trong đó dự đoán cuối cùng là kết quả từ đa số các cây. Đối với hồi quy, trung bình các dự đoán từ tất cả các cây sẽ được lấy làm kết quả cuối cùng.

**Generalization Performance:** Thông qua việc kết hợp kết quả từ nhiều cây, Random Forest thường có khả năng tổng quát hóa tốt, giảm rủi ro của việc overfitting so với việc sử dụng một cây quyết định duy nhất.

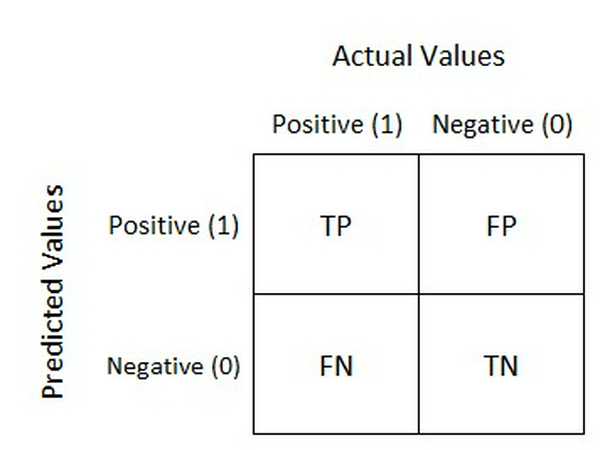
**Handling Missing Values:** Random Forest có thể xử lý dữ liệu thiếu một cách hiệu quả bằng cách sử dụng thông tin từ các nút trong cây để thay thế giá trị thiếu hoặc qua quá trình imputation.

**Feature Importance:** Một lợi ích của Random Forest là khả năng đánh giá mức độ quan trọng của từng đặc trưng đối với dự đoán mô hình. Điều này được xác định dựa trên mức độ mà mỗi đặc trưng cải thiện hiệu suất của cây khi được sử dụng trong các điểm phân chia.

Random Forest thích hợp với các bộ dữ liệu lớn và phức tạp, có khả năng cung cấp kết quả đáng tin cậy mà không cần phải thực hiện quá nhiều tiền xử lý dữ liệu, làm cho nó trở thành một công cụ rất hữu ích và mạnh mẽ trong lĩnh vực học máy.

1. **Các phép đo đánh giá mô hình phân loại**
2. **Confusion matrix**

Đây là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:



1. Confusion matrix

Đối với bài toán đánh giá sự thành công của ứng dụng. Thì mô hình sẽ đánh giá mô hình có là thành công hay không. Success là Positive, còn Unsuccess là Negative

- TP (Dương tính đúng): dự đoán chính xác những ứng dụng thành công

- TN (Âm tính đúng: dự đoán chính xác những ứng dụng không thành công

- FP (Dương tính sai): dự đoán những ứng dụng là thành công nhưng thực sự ứng dụng đó là không thành công

- FN (Âm tính sai): dự đoán những ứng dụng là không thành công nhưng thực sự ứng dụng đó là thành công

1. **Accuracy**

Accuracy là tỷ lệ giữa số lượng dự đoán đúng với tổng số lượng dự đoán, đây là độ đo đơn giản nhất. Công thức tính độ chính xác là:

1. **Precision**

Precision là tỷ lệ giữa số lượng dự đoán dương tính đúng (true positives) so với tổng số dự đoán dương tính (cả đúng và sai). Độ đo này đánh giá sự chính xác của mô hình đối với việc dự đoán trên nhãn dương tính. Công thức tính Precision:

1. **Recall**

Recall là tỷ lệ giữa số lượng dự đoán dương tính đúng so với tổng số mẫu có nhãn là dương tính. Nó đánh giá xem mô hình có tìm ra hết được các mẫu dương tính không. Công thức tính Recall:

1. **F1-Score**

Đây là độ đo đánh giá sự cân bằng của 2 độ đo Precision và Recall, tùy vào bài toán nó sẽ có cách đánh giá khác nhau. Có công thức tính sau:

# **PHẦN 3: TIỀN XỬ LÝ DỮ LIỆU**

Bộ dữ liệu ban đầu gồm 2 file là app.csv (53.732 dòng, 18 cột) và comment.csv (1.468.173 dòng, 4 cột) trong đó app.csv chứa các thông tin mô tả về ứng dụng và comment.csv chứa thông tin chi tiết về bình luận, đánh giá của người dùng về ứng dụng.

Đối với file app.csv, đầu tiên loại bỏ hết các dòng chứa giá trị null, sau đó thực hiện chuẩn hóa dữ liệu từ kiểu chữ sang kiểu số với các trường dữ liệu sau:

* reviews: loại bỏ các ký tự không phải số, sau đó đưa sang kiểu int
* cost\_label: Có các trường hợp đặc biệt như: ‘Install’, ‘<Giá tiền>’ + ‘now reduce to Install, ‘<Giá tiền>’ + ‘now reduce to ‘<Giá tiền>’, ‘<Giá tiền>’. Sẽ được chuyển về thành ‘0’, ‘0’, ‘<Giá tiền>’, <’Giá tiền>’
* size: Từ dữ liệu với 2 đơn vị ‘M’ và ‘k’, ta đưa về cùng một đơn vị là ‘k’.
* installs: Chuẩn hóa dữ liệu sang số (Ngầm hiểu là ≥)
* content\_rating: chuẩn hóa dữ liệu sang số (Ngầm hiểu ý nghĩa là ≥)

Đối với file comment.csv, thực hiện loại bỏ các dòng chứa giá trị null. Để hiểu được cảm xúc của người dùng, chúng tôi sử dụng TextBlob API để tạo thêm 2 cột mới là Polarity (sentiment) và subjectivity tương ứng với mỗi review, polarity có giá trị trong [-1, 1], càng gần 1 thì review càng tích cực và ngược lại thì càng tiêu cực, subjectivity có giá trị thuộc [0, 1], càng gần 1 thì review càng chủ quan và ngược lại thì càng khách quan. Sử dụng hàm gom nhóm để tạo ra bảng mới chứa giá trị raing trung bình, polarity trung bình, polarity trung vị, subjectivity trung bình, subjectivity trung vị cho mỗi app và thực hiện kết bảng mới này với bảng trong file app.csv để có được bảng dữ liệu mới đặt trong file train.csv sử dụng cho huấn luyện mô hình dự đoán sự thành công của ứng dụng sau này.

# **PHẦN 4: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU**

1. **Phân tích thông tin ứng dụng**

Bộ dữ liệu sử dụng cho phân tích trong phần này nằm trong file App.csv. Nó bao gồm 53731 ứng dụng (sau khi đã tiền xử lý).

1. **Thể loại và ứng dụng**

Cột “thể loại” trong tập dữ liệu của chúng tôi có 47 thể loại khác nhau. Hình 1 thể hiện biểu đồ cột về số lượng ứng dụng theo từng thể loại. Từ hình vẽ, có thể thấy rõ thể loại Education và Tools chiếm ưu thế nhất trên GOOGLE PLAY STORE.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

Mô tả được tạo tự động

1. Biểu đồ số lượng ứng dụng theo thể loại trước khi gộp thể loại game

Với thể loại Education, các ứng dụng của nó có thể kể đến như: Duolingo, Learn Languages Free, Learn Math with el Chavo, Beema App - Indore Insurance Institute, v.v… Với thể loại Tools, các ứng dụng của nó có thể kể đến như Google Lens, Calculator, Google Translate, Samsung Calculator, Gallery, v.v…

Ngoài ra, trong bộ dữ liệu, các thể loại trò chơi không phân loại theo 1 thể loại chung, mà chúng được tách ra các thể loại cụ thể. Và dưới đây là những thể loại trò chơi khác nhau mà chúng tôi đã tìm ra được:

* + Action
  + Casual
  + Arcade
  + Puzzle
  + Racing
  + Role Playing
  + Board
  + Card
  + Casino
  + Word
  + Trivia
  + Strategy

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

1. Biểu đồ số lượng ứng dụng theo thể loại sau khi gộp thể loại Game

Hình 3 thể hiện số lượng ứng dụng theo thể loại sau khi đã gộp các thể loại trên vào Game. Sau khi gộp, tổng số thể loại giảm xuống còn 36 và ‘Game’ là thể loại chiếm ưu thế nhất. Để rõ hơn, khi quan sát bảng 1, rõ ràng là thể loại Game thống trị Google Play Store với số lượng vượt trội so với ba thể loại xếp ngay sau đó là Education, Tools và Books & Reference.

|  |  |  |
| --- | --- | --- |
| Place | Genre | Frequency |
| 1 | Game | 10955 |
| 2 | Education | 3292 |
| 3 | Tools | 3237 |
| 4 | Books & Reference | 2128 |
| 5 | Productivity | 1934 |
| 6 | Health & Fitness | 1588 |
| 7 | Music & Audio | 1547 |
| 8 | Lifestyle | 1521 |
| 9 | Sports | 1458 |
| 10 | Personalization | 1375 |

1. Top 10 thể loại có số lượng ứng dụng vượt trội

Điều này thật dễ hiểu với thời đại số phát triển, nhu cầu về sử dụng Internet để giải trí, xả stress tăng cao. Nhà phát triển sẽ phải tập trung vào các ứng dụng phục vụ nhu cầu đó của người dùng. Và đó là lí do tại sao mà thể loại Game có số lượng ứng dụng cao vượt trội như vậy.

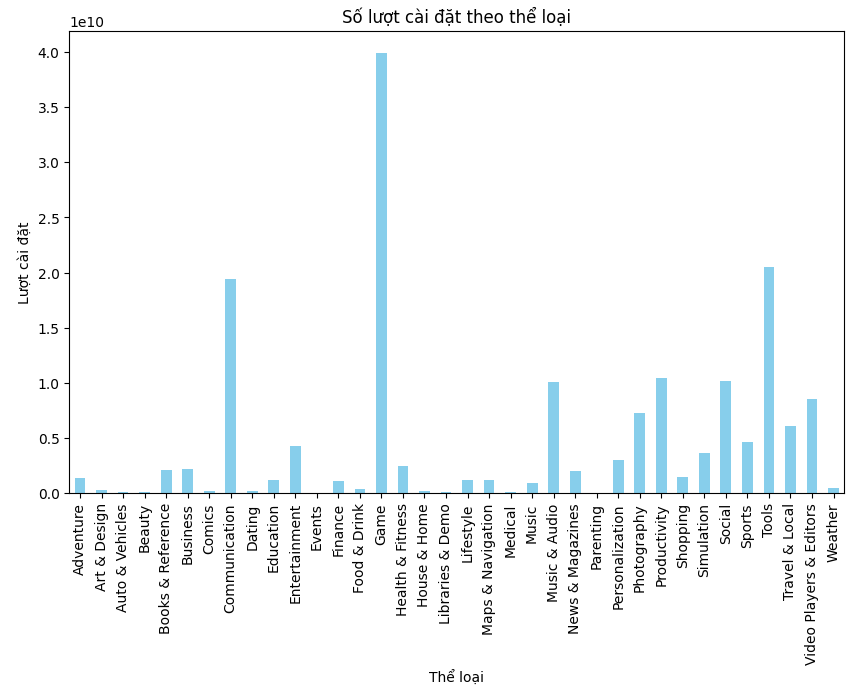
1. **Thể loại và lượt cài đặt**

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

1. Biểu đồ số lượt cài đặt của 10 thể loại phổ biến nhất về số lượng ứng dụng

Ngược lại với kết quả so sánh ở phần 1.1, việc so sánh 10 thể loại phổ biến về số lượng ứng dụng theo lượt cài đặt lại cho thấy một câu chuyện khác. Dựa trên hình 3, ta thấy thể loại Education từ vị trí số 2 về số lượng ứng dụng nay về thành vị trí gần cuối cùng về số lượt cài đặt, Books & Reference cũng không khá hơn khi từ vị trí số 4 thành vị trí số 8. Điều này có thể cho thấy rằng, mặc dù có rất nhiều ứng dụng trong thể hai thể loại này, nhưng chúng lại không nhận được nhiều sự quan tâm, lựa chọn từ phía người dùng. Do đó, nhà phát triển muốn thu hút lượng lớn người dùng nên chọn thể loại dựa trên số lượt cài đặt lớn thay vì dựa trên số lượng lớn ứng dụng trong thể loại đó.



1. Biểu đồ số lượt cài đặt của tất cả thể loại

Để có cái nhìn tổng quan hơn, quan sát Hình 5 để so sánh tất cả các danh mục với số lượt cài đặt. Ta thấy rằng thể loại Game vẫn giữ vị trí số 1 về lượt tải trên GOOGLE PLAY STORE. Vì bộ dữ liệu này được thu thập năm 2020 và năm 2020 cho thấy sự bùng nổ của thể loại game cùng với lượt cài đặt của chúng so với năm 2019 (Hình 6). Nguyên nhân là vì ảnh hưởng của đại dịch Covid-19, hàng tỷ người phải ở nhà cách ly, nhu cầu sử dụng điện thoại để xả stress của người dùng tăng cao, cũng từ đó mà các ứng dụng liên quan về giải trí được quan tâm mạnh mẽ.

Trên Google Play, các ứng dụng thuộc thể loại Tools có số lượt cài đặt nhiều nhất sau Game, đạt 7 tỷ lượt tải xuống lần đầu tiên. Con số này thể hiện mức tăng trưởng 18.6% so với năm trước từ gần 6 tỷ USD.

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Sơ đồ

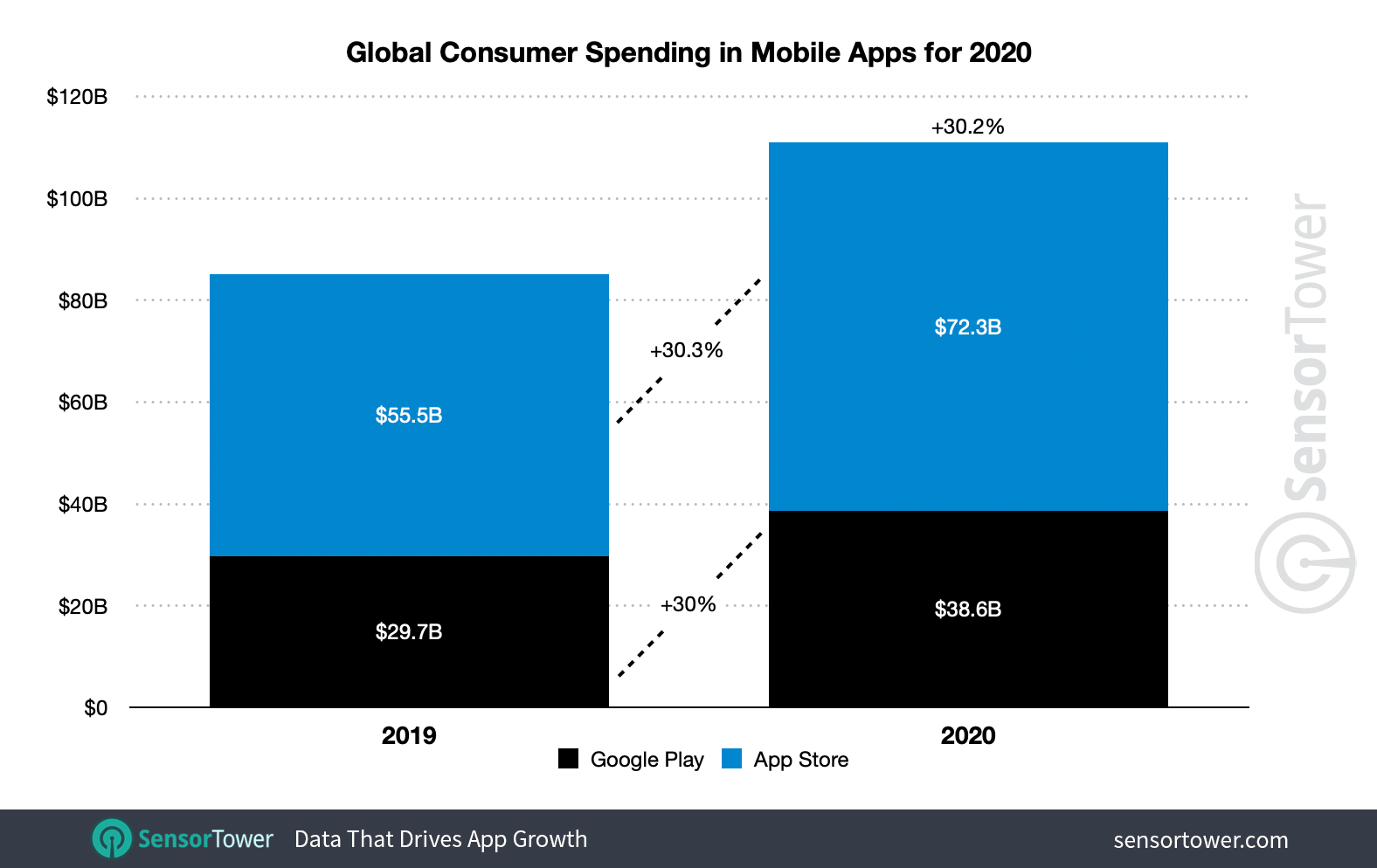
Mô tả được tạo tự động

1. Sự tăng trưởng lượt tải của 3 thể loại game phổ biến đầu Quý 1 2020 so với Quý 4 2019

Tiếp sau game là các thể loại cũng để giải trí và giao tiếp như: Communication, Music & Audio. Đáng chú ý là Tools vẫn chiếm được số lượt cài đặt cao thứ 2, chứng tỏ rằng những ứng dụng của thể loại này rất hữu ích cho người dùng.

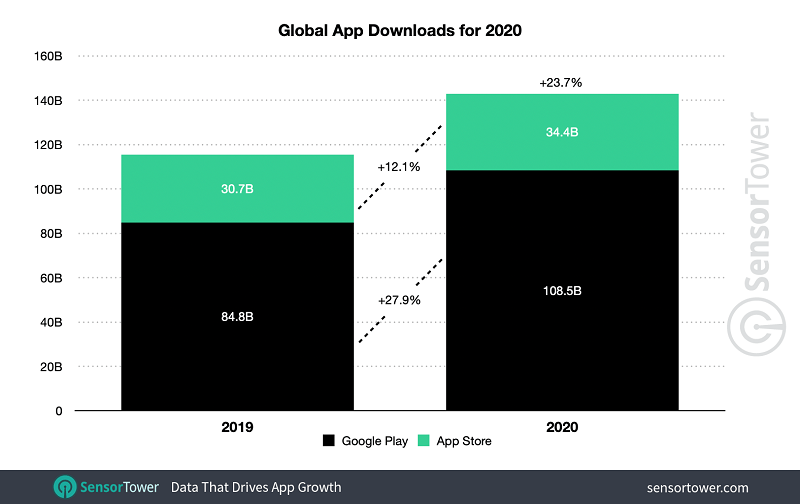
Theo nguồn Sensor Tower, số lượt cài đặt lần đầu tiên lập kỷ lục mới vào năm 2020 khi tổng cộng App Store và Google Play đạt gần 143 tỷ, tăng 23.7% so với mức xấp xỉ 115.5 tỷ vào năm 2019. Con số này cao hơn khoảng 14 phần trăm so với mức tăng trưởng so với cùng kỳ của cả hai cửa hàng trong năm trước.

Trên toàn cầu, chi tiêu của người tiêu dùng trên Google Play, tăng 30% so với cùng kỳ từ 29.7 tỷ USD trong 2019 lên 38.6 tỷ USD.



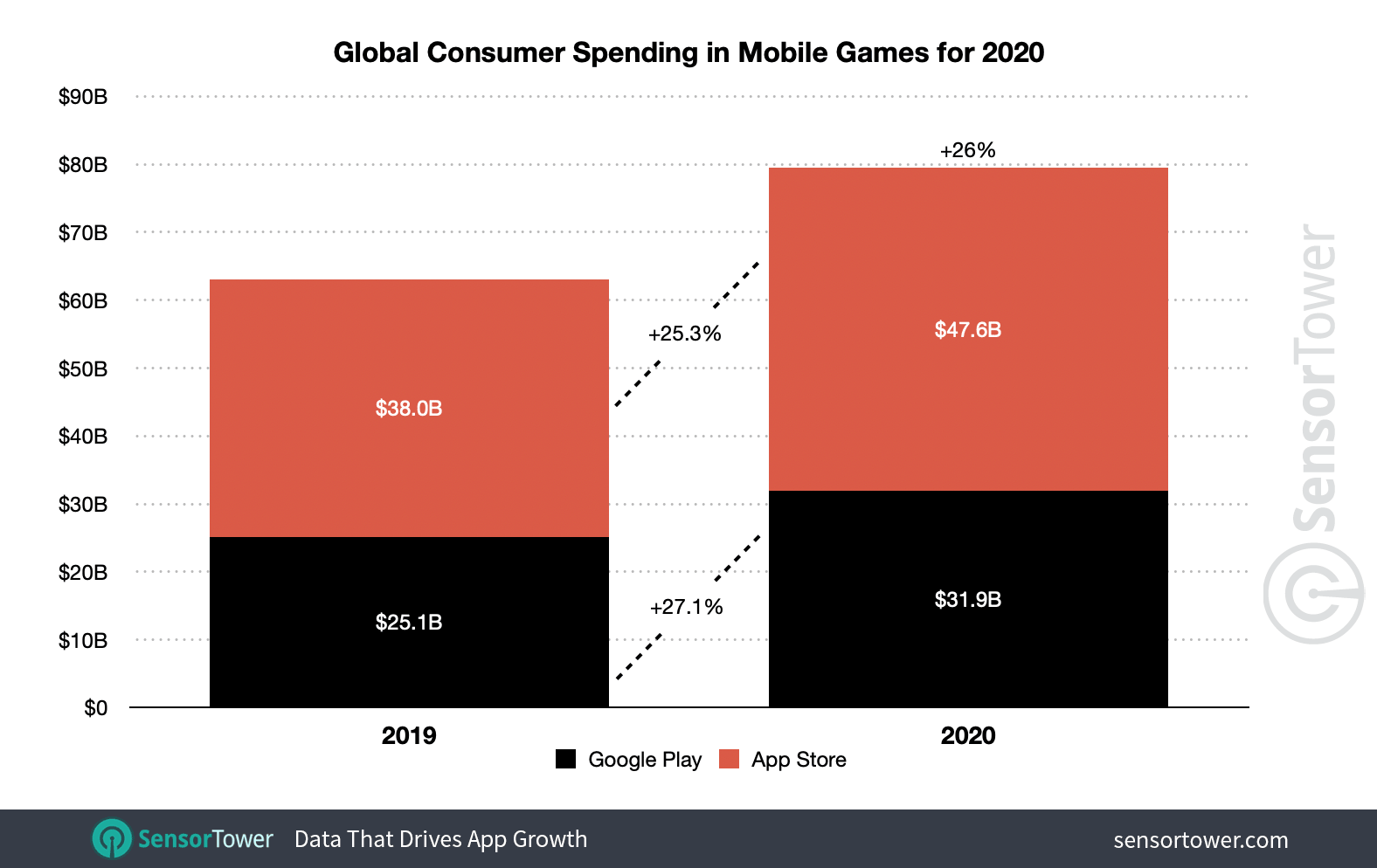
1. *Chi tiêu tiêu dùng toàn cầu cho ứng dụng di động năm 2020*

Số lượt cài đặt lần đầu tiên lập kỷ lục mới vào năm 2020 khi tổng cộng App Store và Google Play đạt gần 143 tỷ, tăng 23.7% so với mức xấp xỉ 115.5 tỷ vào năm 2019. Con số này cao hơn khoảng 14 điểm phần trăm so với mức tăng trưởng so với cùng kỳ của cả hai cửa hàng trong năm trước.



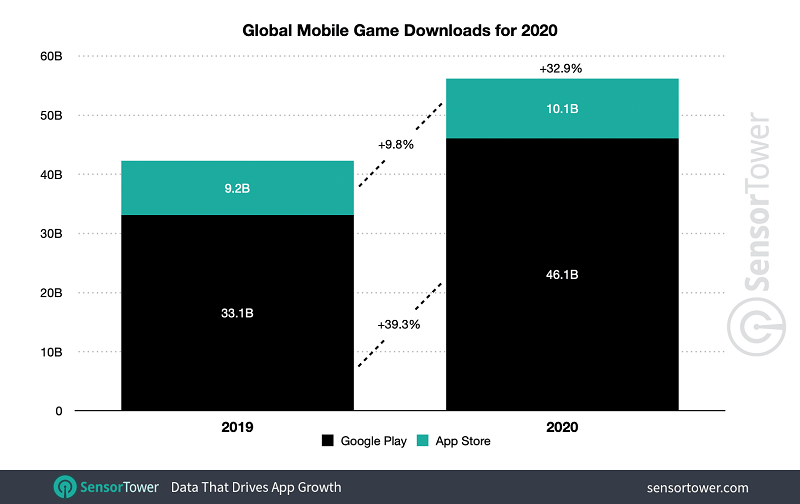
1. Lượng cài đặt ứng dụng trên toàn cầu năm 2020

Trò chơi trên Google Play đạt gần 32 tỷ đô la trong chi tiêu của người tiêu dùng trên toàn thế giới vào năm 2020, tăng 27% so với mức 25 tỷ đô la một chút so với năm trước. Tiêu đề có doanh thu cao nhất trên nền tảng của Google là Moon Active’s Coin Master, với doanh thu tăng 122% so với cùng kỳ lên 799 triệu đô la từ tổng số 360 triệu đô la của năm 2019.



1. Chi tiêu tiêu dùng toàn cầu cho ứng dụng Game năm 2020

Khoảng 82% tổng số lượt cài đặt trò chơi hoặc hơn 46 tỷ lượt là trên Google Play, tăng 39.3% so với cùng kỳ năm ngoái.



1. Lượng cài đặt ứng dụng Game trên toàn cầu năm 2020

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Mô tả được tạo tự động

1. Biểu đồ boxplot về lượt cài đặt theo thể loại

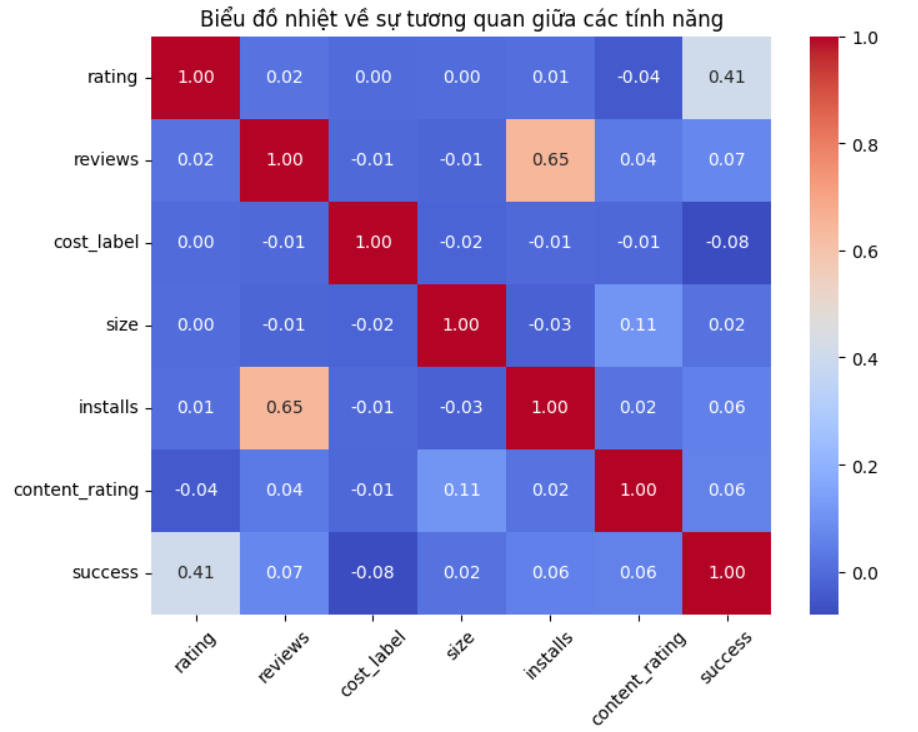
Hình 11 cho thấy các giá trị ngoại lai cao hơn các giá trị bình thường, điều này cho thấy sự phân bố không đồng đều của lượt cài đặt trong thị trường ứng dụng. Các ứng dụng ngoại lai có thể là những ứng dụng đặc biệt phổ biến hoặc có chiến lược tiếp thị hoặc phát triển đặc biệt.

1. **Tương quan số lượt cài đặt**

Cơ sở người dùng là số liệu thống kê quan trọng nhất của một ứng dụng. Do đó, chúng tôi muốn xem sự tương quan của số lượt cài đặt với các đặc trưng khác trong bộ dữ liệu.

|  |  |
| --- | --- |
| Đặc trưng tương quan | Giá trị tương quan |
| Install | 1.000000 |
| Rating | 0.008998 |
| Reviews | 0.645807 |
| Content\_rating | 0.019408 |
| Cost\_lable | -0.005661 |
| Size | -0.027470 |
| Success | -0.007195 |

1. Mối tương quan giữa số lượt cài đặt và các đặc trưng số khác



1. Bản đồ nhiệt của mối tương quan giữa các đặc trưng

Bảng 2 và Hình 12 cho thấy mối tương quan giữa số lượt cài đặt với các đặc trưng số khác dưới dạng giá trị và bản đồ nhiệt tương ứng. Từ đó, chúng ta có thể kết luận rằng các đặc trưng số như xếp hạng tổng thể, chi phí, xếp hạng nội dung và kích thước đều không tương quan với lượt cài đặt. Tuy vậy, đặc trưng đáng chú ý nhất là số lượt đánh giá (reviews). Ta thấy có 1 mối quan hệ tương quan mạnh mẽ và tích cực giữa installs và reviews. Khi một ứng dụng được cài đặt nhiều lần hơn, có khả năng cao hơn rằng nhiều người dùng sẽ sử dụng ứng dụng đó và sau đó đánh giá nó trên Google Play Store.

1. **Tên ứng dụng**

Trong phần này, chúng tôi xác định xem có bất kỳ mối tương quan nào giữa số lượng từ trong tên ứng dụng và số lượt cài đặt bằng cách tạo một cột mới trong tập dữ liệu có tên là “app\_name\_length”. Sau đó, chúng tôi kiểm tra xem các ứng dụng thỏa ngưỡng về số lượng từ nhất định trong tên có khác biệt như thế nào về số lượt cài đặt so với những ứng dụng không thỏa ngưỡng này.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động

1. Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 2 từ

Ảnh có chứa văn bản, ảnh chụp màn hình, Hình chữ nhật, biểu đồ

Mô tả được tạo tự động

1. Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 3 từ

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động

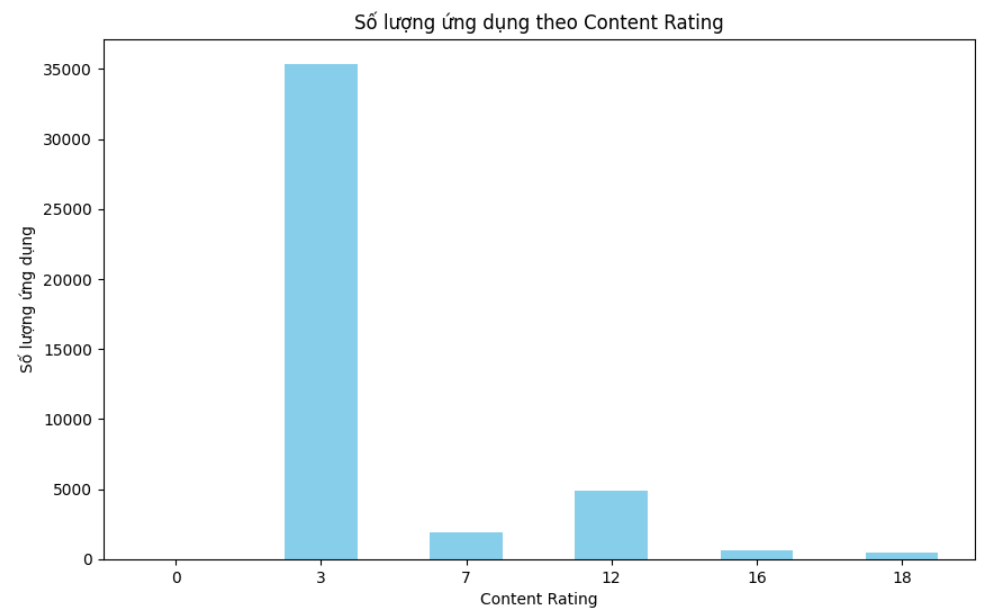
1. Biểu đồ tổng số lượt cài đặt của các ứng dụng ngưỡng 4 từ

Hình 13 cho thấy các ứng dụng có ít hơn hoặc bằng hai từ trong tên chiếm gần một phần ba số lượt cài đặt. Khi ngưỡng tăng lên 3 từ, chúng ta thấy lượt cài đặt tăng đáng kể trong Hình 14 (hơn 1 nửa tổng số cài đặt). Và để khẳng định, hình 15 cho ta thấy sự chiếm ưu thế rõ rệt của các ứng dụng có <= 4 từ, khi mà lượt cài đặt chiếm 2/3 tổng số cài đặt. Có vẻ như hầu hết lượt cài đặt trong Google Play Store đều được đóng góp bởi các ứng dụng có ít hơn hoặc bằng 4 từ trong tên của chúng. Do đó, tốt nhất là các nhà phát triển nên sử dụng số lượng từ ngắn gọn để đặt tên cho ứng dụng của mình.

Điều này khá dễ hiểu, 1 ứng dụng có tên ngắn gọn sẽ giúp người dùng dễ nhớ và dễ dàng tìm kiếm ứng dụng đó trên Google Play Store thay vì 1 ứng dụng có tên quá dài. Ngoài ra, sử dụng tên ngắn gọn làm cho nó dễ dàng hiển thị đầy đủ tên khi trưng bày ở trên Google Play Store.

1. **Xếp hạng nội dung**

Xếp hạng nội dung cho biết lứa tuổi phù hợp để sử dụng ứng dụng. Ở mục 1.1 và 1.2 chúng ta đã lập biểu đồ Số lượng ứng dụng theo thể loại và biểu đồ Lượt cài đặt theo thể loại, nhưng có thể thấy rằng các kết quả không giống nhau. Tuy nhiên, ở phần này, ta có thể thấy được rằng sự tương đồng giữa Số lượng ứng dụng theo Xếp hạng nội dung và Số lượt tải theo Xếp hạng nội dung (Hình 16, Hình 17).



1. Biểu đồ số lượng ứng dụng theo xếp hạng nội dung

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

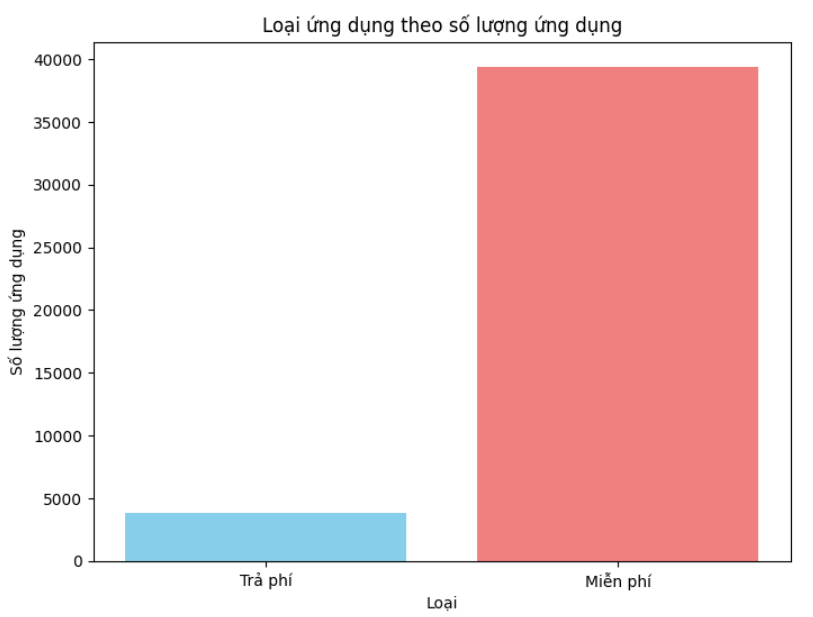
1. Biểu đồ số lượt cài đặt theo xếp hạng nội dung

Nhìn vào 2 biểu đồ có thể thấy được ứng dụng dành cho độ tuổi 3+ là cao nhất. Các ứng dụng chưa được xếp hạng thì chiếm thấp nhất trong cả 2 biểu đồ.

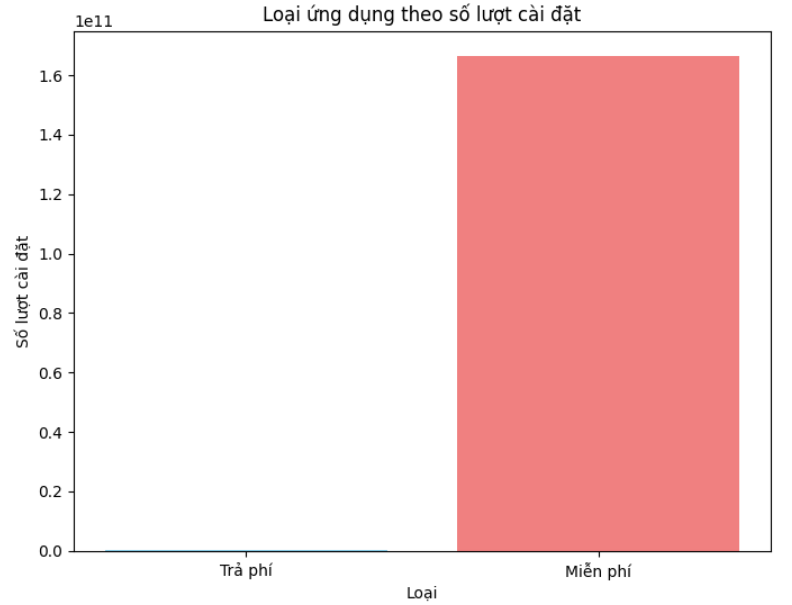
Điều này cho thấy rằng sự chú trọng của các nhà phát triển vào việc phát triển các ứng dụng phù hợp với mọi độ tuổi, và nỗ lực của họ trong việc mở rộng cơ sở người dùng cũng như thu hút sự chú ý trong một môi trường cạnh tranh.

1. **Loại ứng dụng**

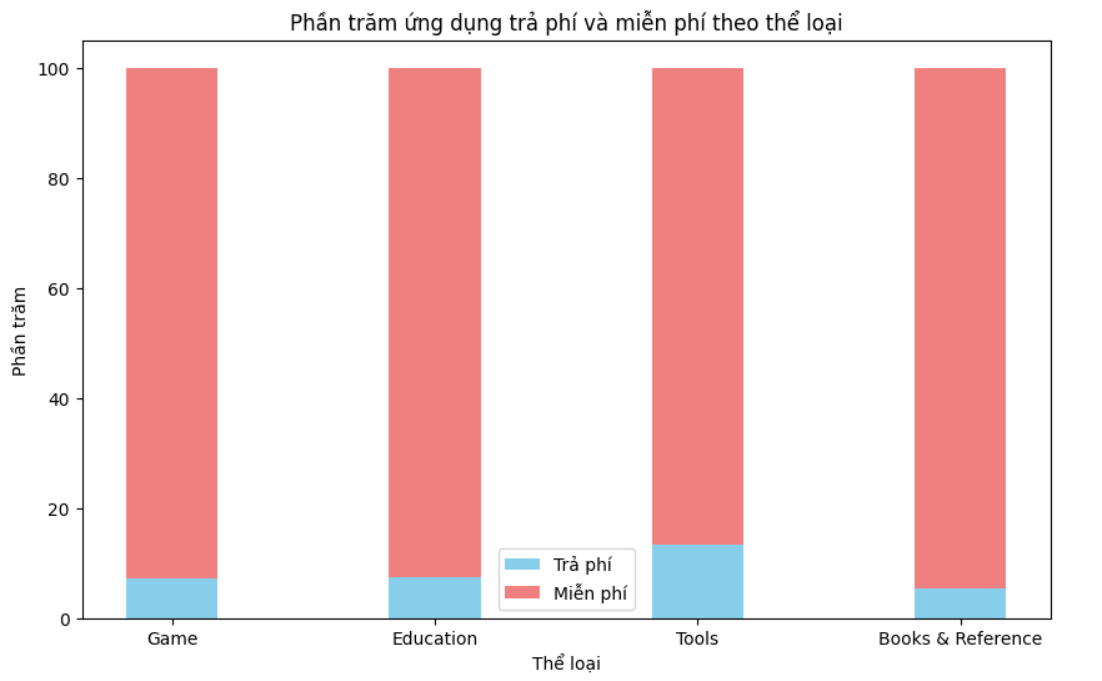
Trong Hình 18 và Hình 19, thể hiện sự thống trị của các ứng dụng “miễn phí” khi mà ứng dụng miễn phí lấn át ứng dụng trả phí, cả về số lượng ứng dụng và lượt cài đặt. Tuy nhiên, không phải tất cả các ứng dụng này đều thực sự miễn phí. Một số ứng dụng miễn phí khi tải xuống nhưng phải tốn phí đăng ký tài khoản để sử dụng ví dụ như NetFlix hay phải tốn phí nâng cấp lên bản premium để xem video không bị quảng cáo như Youtube...



1. Biểu đồ loại ứng dụng theo số lượng ứng dụng



1. Biểu đồ loại ứng dụng theo số lượt cài đặt



1. Biểu đồ loại ứng dụng của 4 thể loại phổ biến nhất

Ảnh có chứa văn bản, vòng tròn, biểu đồ, ảnh chụp màn hình

Mô tả được tạo tự động

1. Biểu đồ tròn về tỉ lệ của 4 thể loại phổ biến nhất

Trong Hình 20 và 21, chúng ta có thể thấy rằng Game, Education và Books & Reference - tất cả đều có một phần nhỏ ứng dụng phải trả phí và chiếm < 10%, riêng Tools thì gần 20%. Điều này có thể cho thấy rằng các ứng dụng trong thể loại Tools thường cung cấp các tính năng hoặc dịch vụ đặc biệt, chất lượng hoặc phức tạp hơn, và do đó có 1 số ứng dụng yêu cầu trả phí để duy trì và phát triển.

1. **Lượt cài đặt và nhà phát triển**

Trong hình 22, chúng ta được thấy top 10 nhà phát triển có tổng số lượt cài đặt nhiều nhất. Trong đó nổi bật hơn cả là Google LLC, lí do cũng khá dễ hiểu vì đây là 1 trong những công ty công nghệ hàng đầu thế giới. Điều này tạo nên niềm tin từ phía người dùng. Không những thế, những sản phẩm của họ đều là những ứng dụng thiết yếu mà người dùng quan tâm.

Xếp thứ 2 là Samsung Electronics và Facebook, họ vẫn có tổng số lượt cài đặt khá cao và xếp sau Google LLC. Đây là những công ty công nghệ hàng đầu thế giới và nhận được sự tin tưởng khá cao của người dùng.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

1. Biểu đồ top 10 nhà phát triển ứng dụng có tổng lượt cài đặt cao nhất
2. **Ứng dụng và nhà phát triển**

Nhìn vào hình 23, chúng ta có thể thấy số lượng ứng dụng của top 10 nhà phát triển có số lượt cài đặt cao nhất. Đáng chú ý là số lượng ứng dụng của Google LLC – nhà phát triển có tổng số lượt cài đặt cao nhất lại chỉ có > 20 ứng dụng, trong khi đó Samsung Electronics lại có số lượng ứng dụng cao gấp đôi (khoảng 40 ứng dụng) mà chỉ xếp thứ 2 về tổng số cài đặt. Điều này càng chứng minh rằng, số lượng ứng dụng chỉ ảnh hưởng 1 phần vào tổng số lượt cài đặt của 1 nhà phát triển. Điều làm người dùng quan tâm và tin tưởng ứng dụng là danh tiếng và uy tín của nhà phát triển ứng dụng đó.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

1. Biểu đồ số lượng ứng dụng của top 10 nhà phát triển có lượt cài đặt cao nhất

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

1. Biểu đồ top 10 nhà phát triển có số lượng ứng dụng nhiều nhất

Có rất nhiều nhà phát triển có số lượng ứng dụng trên Google Play Store nhưng khi nhìn vào tổng số lượt cài đặt thì không thấy tên họ ở đâu trên bảng xếp hạng (Hình 23).

1. **Thể loại và rating**

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, Song song

Mô tả được tạo tự động

1. Biểu đồ boxplot của rating theo thể loại

Trong hình 25, là biểu đồ boxplot của rating theo từng thể loại. Có thể thấy được rằng, rating trung bình của những ứng dụng nằm trong khoảng từ 4 – 4,5 sao. Riêng có thể loại Dating là có rating trung bình thấp hơn cả. Điều này có 1 số nguyên nhân khách quan và chủ quan như sau:

* Người dùng cảm thấy các app Dating không thực sự đáp ứng mong đợi của họ về đối tượng được kết nối.
* Nhiều người sử dụng app Dating để trục lợi và lợi dụng đối phương.
* Sự rò rỉ các hình ảnh nhạy cảm khi sử dụng app Dating.
* Quấy rối, làm phiền nặc danh.
  + Một cuộc khảo sát gần đây của Pew Research cho thấy:
* 38% người hẹn hò trực tuyến nhận được tin nhắn khiêu dâm không mong muốn.
* 30% người hẹn hò trực tuyến tiếp tục nhận được tin nhắn từ người lạ kể cả khi họ đã nói rằng họ không quan tâm.
* 24% người hẹn hò trực tuyến bị gọi bằng cái tên xúc phạm.
* 6% người dùng gặp phải mối đe dọa tổn hại về thể chất.
* 52% người dùng tin rằng họ đã gặp phải ít nhất một vụ lừa đảo, 63% nam giới dưới 50 tuổi và 44% phụ nữ ở mọi lứa tuổi.

1. **Kiểm định giả thuyết**

***Giả thuyết "Số lượt tải xuống của một ứng dụng phụ thuộc vào thể loại của ứng dụng đó"***

Đặt giả thuyết null : "Số lượt tải xuống của một ứng dụng không phụ thuộc vào thể loại của ứng dụng đó".

Xét giá trị α = 0.05

Sử dụng thống kê ANOVA, tính được p-value = 6.32  α => Bác bỏ giả thuyết . Tức là số lượt tải xuống của một ứng dụng phụ thuộc vào thể loại của ứng dụng đó.

***Giả thuyết "Số lượt tải xuống của một ứng dụng phụ thuộc vào nhà phát triển của ứng dụng đó"***

Đặt giả thuyết null : "Số lượt tải xuống của một ứng dụng không phụ thuộc vào nhà phát triển của ứng dụng đó".

Xét giá trị α = 0.05

Sử dụng thống kê ANOVA, tính được p-value = 0.99 > α => Chấp nhận giả thuyết . Tức là số lượt tải xuống của một ứng dụng không phụ thuộc vào nhà phát triển của ứng dụng đó.

***Giả thuyết "Đánh giá của người dùng (rating) phụ thuộc vào số lượt tải xuống của ứng dụng"***

Đặt giả thuyết null : "Đánh giá của người dùng (rating) không phụ thuộc vào số lượt tải xuống của ứng dụng."

Xét giá trị α = 0.05

Sử dụng thống kê ANOVA, tính được p-value = 0 < α => Bác bỏ giả thuyết . Tức là đánh giá của người dùng phụ thuộc vào số lượt tải xuống của ứng dụng.

Trong 3 giả thuyết được đặt ra, có 2 giả thuyết bị bác bỏ, 1 giả thuyết chấp nhận. Số lượt tải xuống phụ thuộc vào thể loại của ứng dụng đó. Có thể thấy rằng, thể loại trò chơi và giải trí đang được ưa chuộng bởi vì ảnh hưởng của Covid-19, hay thể loại Tools cũng được sử dụng nhiều vì nó rất hữu ích.

Hay số lượt tải xuống cũng ảnh hưởng tích cực đến đánh giá của người dùng bởi vì tâm lý người dùng thường tin tưởng vào các ứng dụng đã được tải xuống nhiều lần vì họ cho rằng nó phải có một giá trị hoặc chất lượng đủ để thu hút mọi người. Do đó, các ứng dụng có số lượt tải xuống cao thường có xu hướng nhận được các đánh giá tích cực.

1. **Phân tích Review và rating**

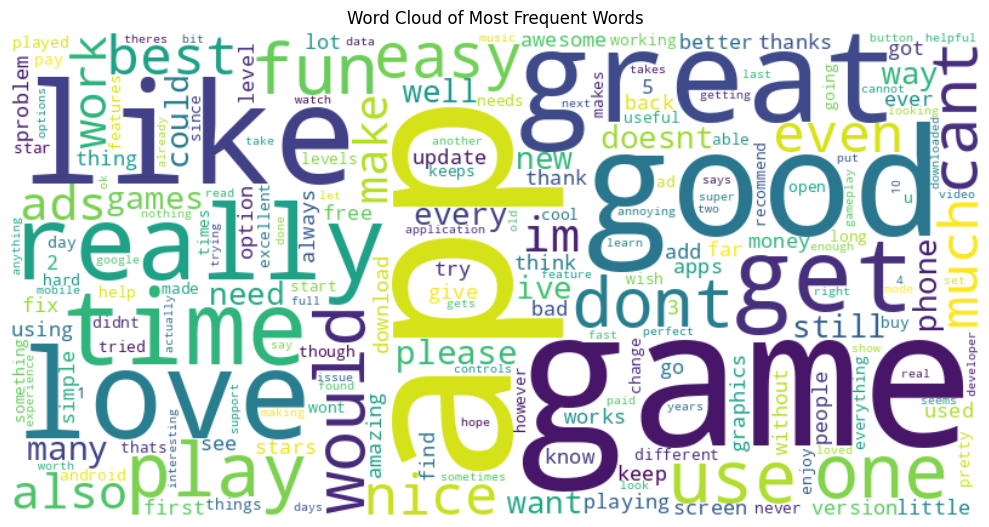
Phần này chúng tôi tập trung phân tích nội dung của reviews và rating tương ứng để khám phá các câu chuyện từ cảm xúc của người dùng về ứng dụng. Bộ dữ liệu được sử dụng cho phân tích chứa trong file comment.csv có 46706 apps và 979433 comments với khoảng 21 comments/app.

1. **Từ điển từ**

Đây là từ điển chứa các từ xuất hiện trong các reviews và tần suất xuất hiện của chúng. Để tạo một từ điển từ, ban đầu, chuyển tất cả các từ trong mỗi review về dạng in thường để tránh trường hợp 2 từ giống nhau nhưng lại được xem là khác nhau bởi chúng khác nhau về ký tự in hoa – in thường ví dụ: Awesome và AWESOME. Tiếp theo, loại bỏ hết các dấu câu trong mỗi review, rồi tạo một danh sách từ đại diện cho mỗi review và lưu tất cả các từ vào danh sách đó, sau đó loại bỏ sự trùng lặp từ và loại bỏ các từ dừng ( là các từ xuất hiện nhiều trong câu nhưng không mang nhiều ý nghĩa ví dụ: a, an, the, and, or, …) trong danh sách. Cuối cùng với mỗi từ trong danh sách đó, nếu chúng chưa có trong từ điển thì thêm chúng vào từ điển và gán giá trị tần suất bằng 1 hoặc tăng tần suất của từ đó lên 1 nếu chúng đã tồn tại trong từ điển.

Thực hiện tạo từ điển từ năm lần cho năm loại rating khác nhau (từ 1\* đến 5\*) và có được năm từ điển từ khác nhau tương ứng với mỗi rating, mỗi từ điển biểu thị các từ xuất hiện và số reviews của một rating tương ứng mà có chứa một từ cụ thể. Dựa trên các từ điển từ này, chúng tôi vẽ các biểu đồ cột về phân phối tần suất xuất hiện của các từ để phân tích mối liên hệ giữa chúng với rating tương ứng.

1. **Phân tích từ điển từ**



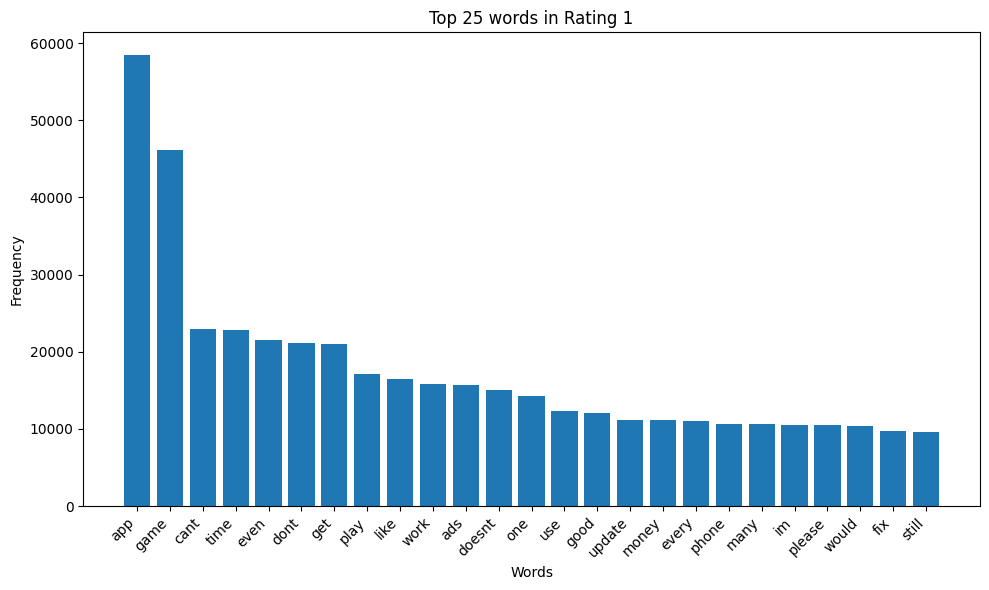
1. Biểu đồ word cloud về các từ xuất hiện nhiều nhất trong các reviews

Dựa vào biểu đồ word cloud ở trên có thể nhận thấy “game” và “app” là hai từ được sử dụng thường xuyên nhất trong các reviews, điều này hợp lý vì 1/5 bộ dữ liệu của chúng tôi chứa các app thuộc thể loại game vàtheo thống kê mới nhất của googplay store năm 2024 (Hình 27) thìgame là thể loại app phổ biến nhất trên GOOGLE PLAY STORE (chiếm 12.64% tổng số apps).Với số lượng apps nhiều hơn dẫn đến thể loại game có thể nhận được nhiều reviews hơn so với các thể loại khác.

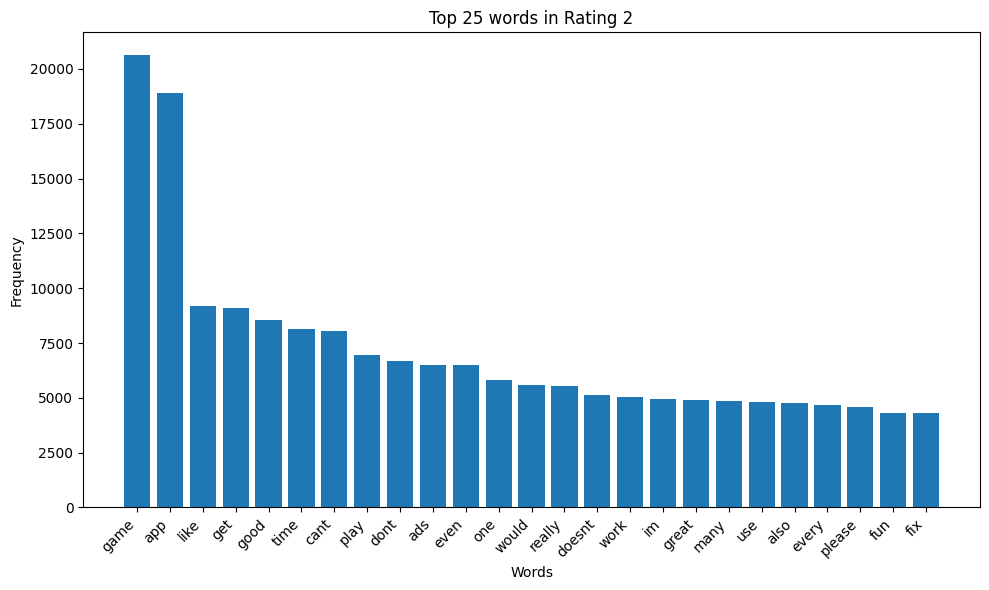
Ngoài ra, hai từ “game” và “app” không chỉ liên quan đến một rating cụ thể. Vì khi quan sát các biểu đồ ở (Hình 28, 29, 30, 31, 32), các từ này luôn đứng đầu về tần suất xuất hiện ở tất cả các rating từ 1\* đến 5\*.



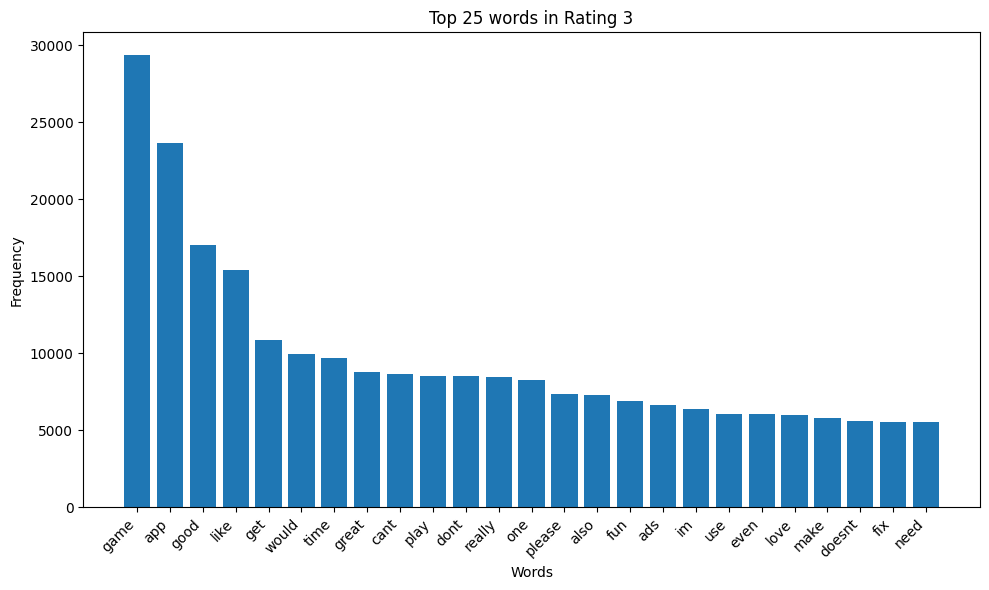
1. Thống kê của GOOGLE PLAY STORE về tỷ lệ số lượng ứng dụng thuộc thể loại game cập nhật đến ngày 24/4/2024



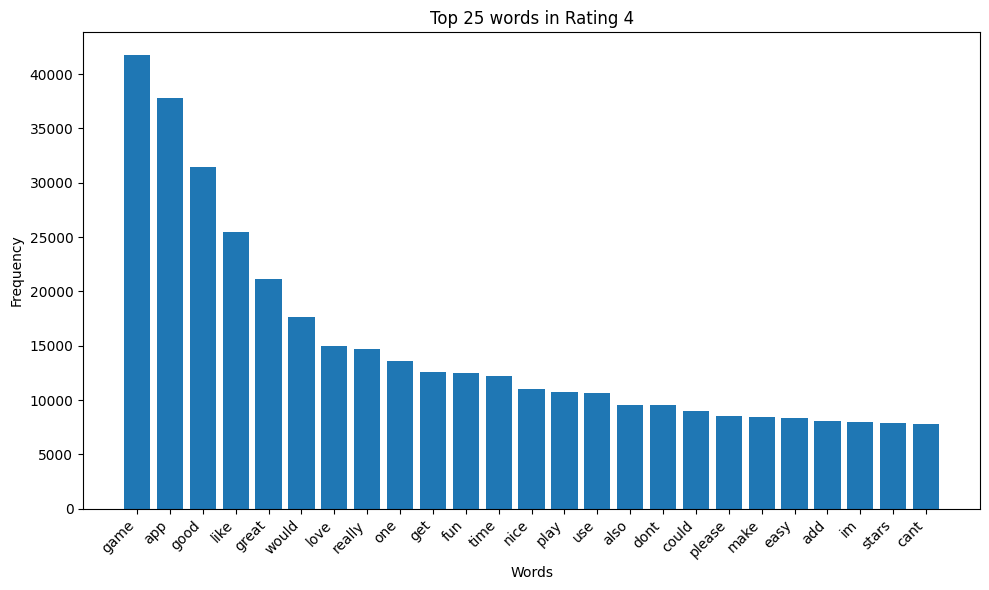
1. Phân phối tần suất của các từ ứng với rating 1



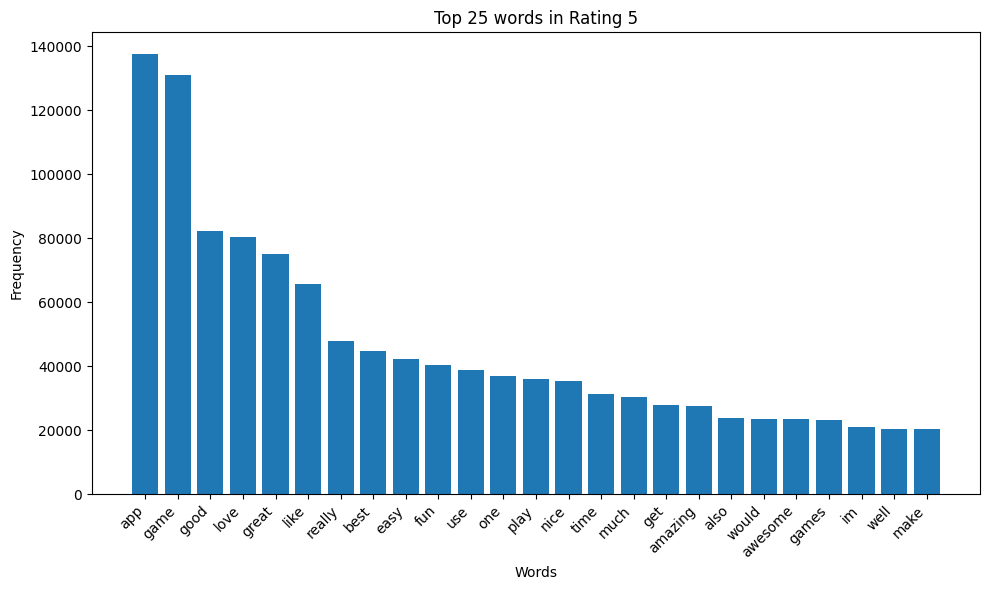
1. Phân phối tần suất của các từ ứng với rating 2



1. Phân phối tần suất của các từ ứng với rating 3



1. Phân phối tần suất của các từ ứng với rating 4



1. Phân phối tần suất của các từ ứng với rating 5

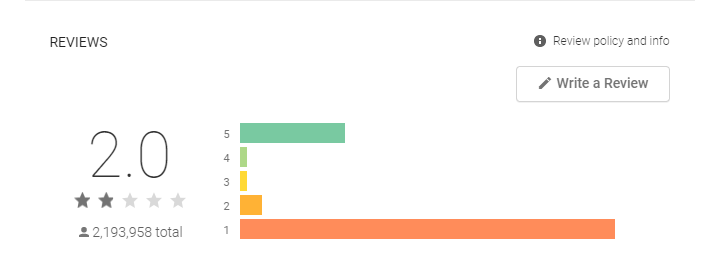
Trong các biểu đồ (Hình 28, 29, 30, 31, 32) có thể thấy rằng có một số từ trùng lặp giữa các ratings khác nhau. Ví dụ, từ "good" thường xuất hiện trong các đánh giá có 3\*, 4\* và 5\*. Tuy nhiên, ngay cả với đánh giá 1\* và 2\*, chúng ta cũng thấy từ này xuất hiện trong danh sách 25 từ được sử dụng nhiều nhất. Nguyên nhân có thể là do người dùng sử dụng cụm từ tiêu cực chứa “good” ví dụ như “not good” để thể hiện sự không hài lòng về ứng dụng, tuy nhiên để rõ hơn thì chúng tôi thực hiện truy vấn các reviews có từ "good" nhưng có rating 2\*, 1\* để tìm thêm các nguyên nhân khác. Một phần kết quả truy vấn được trình bày trong các bảng 1

|  |  |  |
| --- | --- | --- |
| Rating | Sentiment | Review |
| 2\* | -0.25 | The game is really good...but after the recent update I cant even open the game. It freezes then returns back to home page. |
| 2\* | -0.1 | WORST UPDATE EVER! It is a really good game but this update really sucks. You'll need fast internet just to play. You can't even edit your avatar without the internet. |
| 2\* | -0.35 | App is very good in general. But after last update tap to translate bubble is not appearing/disappearing automatically. This degrades the usability nearly to zero. |
| 1\* | -0.15 | It's a good game but the new update is faulty. I have never been fps dropped in this game before until the new update. And now I can not access to the game despite many different attempts |
| 1\* | -0.25 | Worst update ever!! I enjoyed playing the older version which had more missions and had a good graphics too. |
| 1\* | -0.2 | It was pretty good but in the updates it getting worse there's a mission streetrace4.0 it saids there are a spring but there's none. |

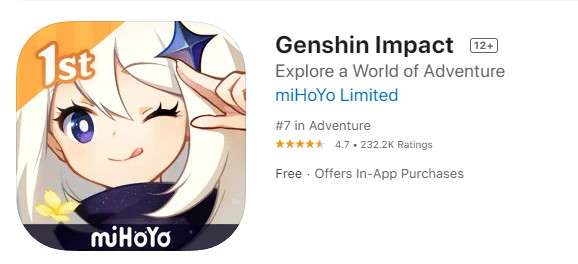
1. Những reviews phản ánh sự hài lòng trong quá khứ

Bảng trên cho ta biết rằng một số người dùng sử dụng những từ tích cực như “good” để mô tả phiên bản trước đó của ứng dụng mà họ hài lòng. Những từ tích cực như vậy xuất hiện trong các reviews có rating thấp cũng có thể cho thấy rằng phiên phản mới cập nhật của ứng dụng chưa thực sự được đón nhận bởi người dùng.

Khi người dùng gặp sự cố với phiên bản mới, thì họ có thể thay đổi lại rating với số sao thấp hơn trong quá khứ làm cho rating trung bình ứng dụng tụt dốc nhanh chóng. Ví dụ như trường hợp của game Gensin Impact vào năm 2021, sau bản cập nhật mới làm người dùng không hài lòng, điều này đã tạo ra hiện tượng bão 1\* trên trang đánh giá ứng dụng khi mà người dùng ồ ạt rating lại 1\* để phản đối nhà phát triển (Hình 32). Ban đầu rating trung bình trong trang review của Genshin Impact là 4.7 / 5. Nhưng chỉ trong vòng một ngày, rating trung bình trong trang review đã giảm xuống chỉ còn 2.0.



1. Hiện tượng bão 1\* trên trang đánh giá của game Gensin Impact



1. Game Genshin Impact ở trang sản phẩm của GOOGLE PLAY STORE

Ta thấy rằng mặc dù rating trung bình trong trang review của Gensin Impact là 2.0, tuy nhiên ở trang sản phẩm thì rating trung bình vẫn giữ ở mức 4.7 (Hình 34). Điều này là do GOOGLE PLAY STORE thực thi một chế độ mà không cập nhật rating trung bình trên trang sản phẩm ngay lập tức, thay vào đó nếu có một tỉ lệ phần trăm cụ thể người dùng thực hiện đánh giá lại với rating thấp hơn thì GOOGLE PLAY STORE sẽ thông báo cho nhà phát triển ứng dụng biết và cho họ 1 tuần để thực hiện những điều chỉnh nhằm giúp tăng lại rating trung bình trước khi rating trung bình được cập nhật trên trang sản phẩm.Vì vậy khoảng thời gian này giúp cho người dùng tiềm năng của ứng dụng sẽ không bị ảnh hưởng bởi rating trung bình thấp khi đưa ra quyết định cài đặt ứng dụng.

1. **Hiện tượng không khớp giữa các trường dữ liệu**

|  |  |  |
| --- | --- | --- |
|  | Sentiment | Review |
| False Positive | 0.55 | It says free to play when you push the app, when you play it they ask for you to pay for it, not good. |
| 0.6 | Not working so good because many interfere off and on. |
| 0.7 | Not so good. |
| 0.7 | This app doesn't work with 60fps cameras. This is why the app is not so good. |
| 0.7 | graphics are not so good as the picture's |
| False Negative | -0.25 | Aaasome, I had never played like this game ,this game story and plot are so good.very peculiar game from other. The controls,map,bot stunts are very admiring me and it refresh the mind and activate the brain to think. I'm addicted to this game. |
| -0.325 | This game is so enjoyable,it's so good,I can't stop playing it,I played this everyday I just can't stop playing this game. |
| -0.4 | This so good.i like this game. |
| -0.4 | No bugs and no crashes at the game so good. |

1. Những giá trị sentiment không khớp với Review

Quan sát bảng 2, ta thấy câu review “Not so good” bị gán nhầm giá trị sentiment khá cao là 0.7 và câu review “This so good.i like this game.” cũng bị gán nhầm giá trị sentiment khá thấp là -0.4. Do đó, có thể thấy rằng TextBlob API không thực sự chính xác hoàn toàn khi mà có sự dự đoán sai một số reviews có nội dung là phàn nàn và đánh giá thấp ứng dụng nhưng giá trị sentiment lại cao (False Positive), ngược lại có những reviews mà nội dung tổng thể là khen và đánh giá cao ứng dụng nhưng giá trị sentiment lại rất thấp (False Negative).

Một trong những nguyên nhân dẫn đến việc dự đoán nhầm có thể là do nội dung tổng thể của câu review là tích cực/tiêu cực tuy nhiên trong câu lại xuất hiện những cụm từ mang tính ngược lại. Ví dụ nội dung tổng thể của câu review đầu tiên trong bảng 2 là tiêu cực, phàn nàn nhưng lại có cụm từ tích cực như “free to play”. Hoặc nội dung tổng thể của review thứ 6, 7 là tính cực, khen game hay nhưng lại xuất hiện những cụm từ không tích cực như “I'm addicted to this game.” hoặc “I just can't stop playing this game.”. Vấn đề sentiment không khớp với review có thể dẫn đến việc giá trị sentiment không phản ánh đúng được cảm nhận của người dùng về ứng dụng. Điều này có thể ảnh hưởng đến kết quả của mô hình dự đoán sau này khi mà mô hình có sử dụng trường sentiment trong bộ dữ liệu để sử dụng cho việc huấn luyện.



1. Biểu đồ box plot về phân phối của sentiment theo rating

Dựa vào biểu đồ (Hình 35) ta thấy 2 khu vực đóng khung đỏ là các giá trị ngoại lai bất thường của 2 nhóm: rating thấp (1\*, 2\*) và rating cao (4\*, 5\*). Rõ ràng nhóm rating cao (4\*, 5\*) lại có những ngoại lai với giá trị sentiment rất thấp và nhóm rating thấp (1\*, 2\*) lại có những ngoại lai với giá trị sentiment rất cao có thể cho biết được rằng có sự không khớp giữa sentiment và rating tương ứng trong bộ dữ liệu. Để rõ hơn, chúng tôi thực hiện truy xuất các reviews trong bộ dữ liệu (Bảng 3) và thấy rằng review đầu tiên có nội dung là thất vọng, phàn nàn về ứng dụng và giá trị sentiment thấp nhưng lại nhận được rating rất cao là 5\*. Điều này có thể là do có thể do người dùng rating nhầm. Hay trong bảng 4, review đầu tiên: “Very good app I like this app so good” có nội dung tích cực, đánh giá cao app và giá trị sentiment cao tuy nhiên chỉ nhận được rating 2\* mặc dù chúng nên được nhận rating 4\*, 5\*. Nguyên nhân có thể là nhà phát triển đã cập nhật thêm các tính năng mới mà người dùng mong đợi hoặc khắc phục các sự cố hiện có của ứng dụng và người dùng thể hiện sự hài lòng thông qua việc cập nhật lại review với nội dung tích cực hơn tuy nhiên họ lại quên không cập nhật lại rating mà trước đó họ đã rating thấp.

|  |  |  |
| --- | --- | --- |
| Rating | Sentiment | Review |
| 5\* | -0.3125 | I can't play even 5 min fix your game because I spent money on it the game is not good connection errors always |
| 5\* | -0.35 | I try this box video coz some app not good and always error |
| 5\* | -0.35 | The navigation and location is not good. Introducing is not good. Updating is not good. It dosent have any options in update |
| 4\* | -0.35 | graphics are not good |
| 4\* | -0.258 | I don't like this game because the control is not proper at all. The game's graphics are also not good. My device was not working properly after installing this game. |
| 4\* | -0.25 | Getting black screens..and no its not good |

*Bảng 3: Các reviews tiêu cực nhưng có rating cao*

|  |  |  |
| --- | --- | --- |
| Rating | Sentiment | Review |
| 2\* | 0.7 | Very good app I like this app so good |
| 2\* | 0.7 | Also good |
| 2\* | 0.7 | Its so good. |
| 1\* | 0.65 | This app is so good.it makes me all the time happy when using it...I love so must this app |
| 1\* | 0.6 | This camera is so good you can love it it |
| 1\* | 0.6 | I am not from punjab but it's look this app so good. My wish govt sindh follow and make sure to eassy for access of sindh people |

1. Các reviews tích cực nhưng có rating thấp

Vấn đề rating không khớp với review có thể dẫn đến sự không hợp lý của rating trung bình, khi mà rating trung bình không thực sự phản ánh đúng được đánh giá chung của người dùng về ứng dụng. Điều này có thể ảnh hưởng đến việc tải xuống của người dùng bởi thông thường người dùng sẽ nhìn vào rating trung bình để đưa ra quyết định có tải xuống ứng dụng hay không.

1. **Kiểm định giả thiết**

Phần này đưa ra các giả thuyết và thực hiện kiểm định để kiểm tra xem liệu một số từ/ký hiệu xuất hiện trong review có thể dẫn đến rating cao hơn/ thấp hơn so với những reviews không chứa những từ/ký tự đó hay không.

|  |  |
| --- | --- |
| Rating | Review |
| 5\* | CUTE!!! I LOVE THIS GAME |
| 5\* | VERY INTERESTING AND EDUCATIONAL |
| 5\* | I HIGHLY RECAMEND THIS APP TO ANYONE. |
| 4\* | ITS FUN TO PLAY. SIMPLE. I THINK THIS ONE HAS TOO MANY ADS. BUT ITS FREE! |
| 4\* | THIS IS NICE GAME .I PLAY THIS FOR A LONG TIME |
| 4\* | OMG NICE |

1. Các reviews có nội dung chứa toàn chữ in hoa

|  |  |
| --- | --- |
| Rating | Review |
| 2\* | Now what's wrong with this game ? It kept logout in the middle game, fix it or I uninstall the game |
| 2\* | Its good but where the hell is the endless mode and new characters? Like just update it. |
| 2\* | why it doesn't work on tablet? F ix. it please and the altered beast! |
| 1\* | I am using a premium v. So Why delay accepting finger print, also still adds there ?? |
| 1\* | I can't use apps sorry I do not need to pay for meet anyone.. why I should to pay? |
| 1\* | how can it used that's is not log in |

1. Các reviews có chứa câu hỏi

|  |  |
| --- | --- |
| Rating | Review |
| 5\* | Very good. You get alert whenever a new mail enters. Great app! |
| 5\* | great! I use it for more than 4 years. |
| 5\* | no delays, no hassle and fast! |
| 4\* | Kids love it! Just wish we didnt have to pay to progress to boss levels |
| 4\* | This game is awesome! If your a killing or action person this is exacly for you! |
| 4\* | Superb! Controls are awesome ! |

1. Các reviews có chứa dấu chấm than “!”

Khi quan sát một vài review trong bộ dữ liệu, chúng tôi thấy rằng:

* + Nếu review mà tất cả nội dung được viết in hoa thì điều đó thể hiện sự hài lòng, dẫn đến rating cao hơn. (Bảng 5)
  + Nếu review là để hỏi về một điều gì đó thì thể hiện sự không hài lòng, dẫn đến rating thấp hơn. Review này được xác định qua các từ để hỏi như Why, When, Where, What hoặc có dấu “?” (Bảng 6)
  + Nếu review chứa dấu “!” thì thể hiện sự hài lòng và dẫn đến rating cao hơn. (Bảng 7)

Để kiểm tra xem các giả thiết ở trên có đúng hay không thì chúng tôi tiến hành xác định các giả thuyết null tương ứng và kiểm định từng giả thuyết null đó.

***Đối với giả thiết 1:***

Gọi là rating trung bình của các review chứa toàn chữ in hoa trên GOOGLE PLAY STORE

là rating trung bình của các review còn lại

Ta cần kiểm định giả thiết:

Dựa trên mẫu từ bộ dữ liệu, tính toán ta thu được giá trị t của phân phối t-Student:

t = 30.135

p-value = P( t > 30.135) = 0 < 0.05

Suy ra bác bỏ , chấp nhận .

Vậy những reviews chứa toàn bộ chữ in hoa có xu hướng có rating trung bình cao hơn so với các reviews còn lại.

***Đối với giả thiết 2:***

Gọi là rating trung bình của các review có chứa câu hỏi trên GOOGLE PLAY STORE

là rating trung bình của các review còn lại

Ta cần kiểm định giả thiết:

Dựa trên mẫu từ bộ dữ liệu, tính toán ta thu được giá trị t của phân phối t-Student:

t = -124.6291

p-value = P( t < -124.6291) = 0 < 0.05.

Suy ra bác bỏ , chấp nhận .

Vậy những reviews có chứa câu hỏi có xu hướng có rating trung bình thấp hơn so với các reviews còn lại.

***Đối với giả thiết 3:***

Gọi là rating trung bình của các review có chứa dấu chấm than trên GOOGLE PLAY STORE

là rating trung bình của các review còn lại

Ta cần kiểm định giả thiết:

Dựa trên mẫu từ bộ dữ liệu, tính toán ta thu được giá trị t của phân phối t-Student:

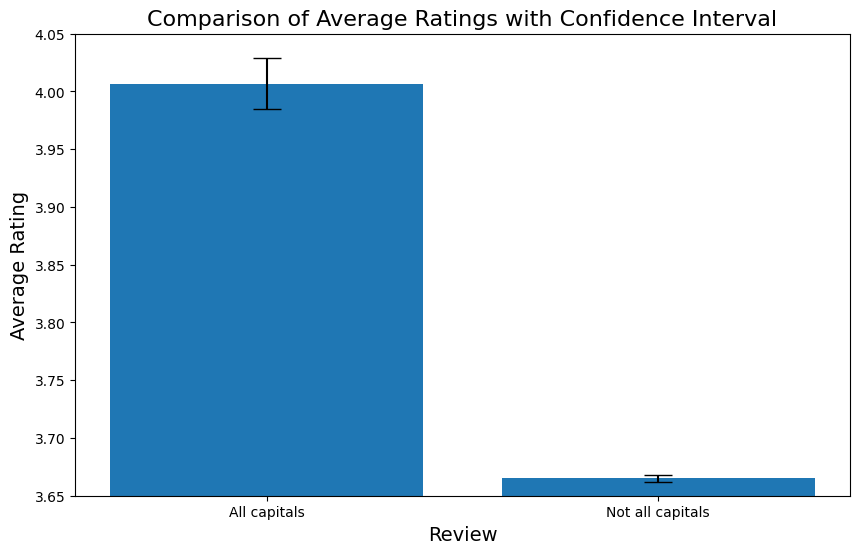
t = 77.77

p-value = P( t > 77.77) = 0 < 0.05.

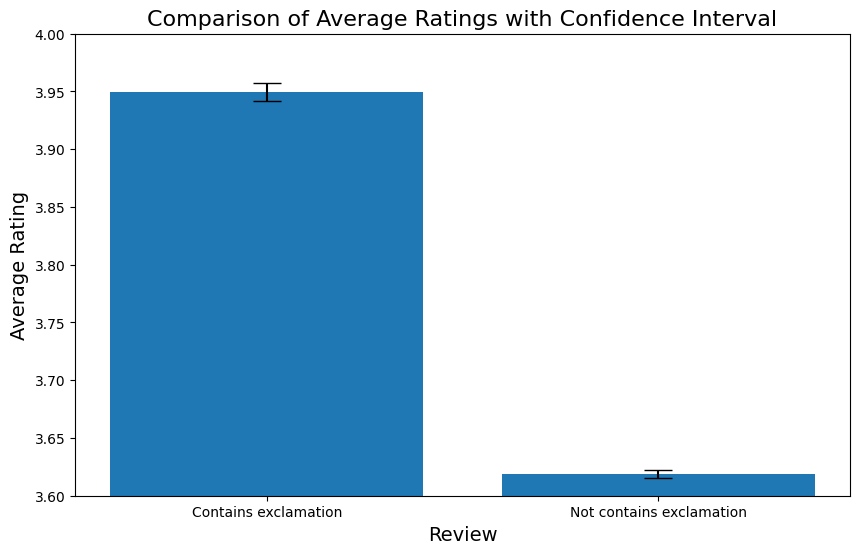
Suy ra bác bỏ , chấp nhận .

Vậy những reviews có chứa dấu chấm than có xu hướng có rating trung bình cao hơn so với các reviews còn lại.

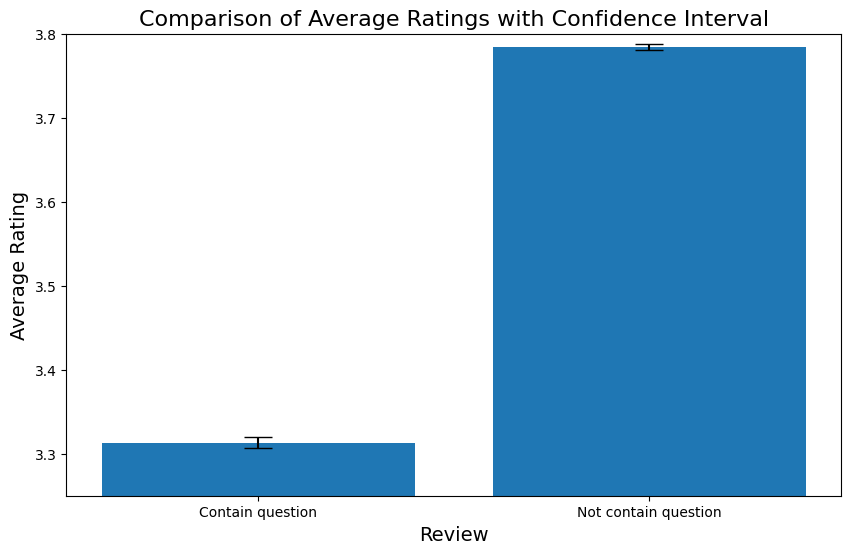
Để chắc chắn hơn về kết quả của kiểm định thì chúng tôi thực hiện tính toán khoảng tin cậy cho các rating trung bình và trực quan hóa ở các biểu đồ (Hình 35, 36, 37). Nhìn vào 3 biểu đồ ta thấy các kết quả nhất quán với kết quả của kiểm định đó là: khoảng tin cậy rating trung bình của review chứa toàn ký tự hoa nằm trên hoàn toàn so với khoảng tin cậy rating trung bình của các review không chứa các ký tự hoa, tương tự với trường hợp của review chứa dấu chấm than. Trường hợp review chứa câu hỏi thì khoảng tin cậy nằm dưới hoàn toàn so với khoảng tin cậy của các review không chứa câu hỏi.



1. Rating trung bình của các review chứa toàn ký tự in hoa so với các review còn lại



1. Rating trung bình của các review chứa dấu chấm than so với các review còn lại



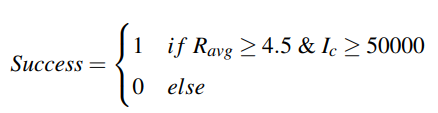
1. Rating trung bình của các review chứa câu hỏi so với các review còn lại

Thông qua kiểm định và ước lượng khoảng tin cậy, ta kết luận được rằng: những review mà toàn bộ nội dung được in hoa hoặc có chứa dấu chấm than thì thường sẽ nhận được rating cao hơn, điều này là do việc sử dụng các từ in hoa thường là để nhấn mạnh một điều gì đó tích cực và sử dụng dấu chấm than là để bày tỏ cảm xúc hứng thú, hài lòng của người dùng về ứng dụng. Ngược lại những review có nội dung là để hỏi thì thường sẽ nhận được rating thấp hơn, có thể là do người dùng đang gặp sự cố/vấn đề khi sử dụng ứng dụng và họ đặt câu hỏi để tìm giải đáp hoặc yêu cầu nhà phát triển giải quyết sự cố hiện có của ứng dụng.

# **PHẦN 5: MÔ HÌNH DỰ ĐOÁN SỰ THÀNH CÔNG CỦA ỨNG DỤNG**

1. **Định nghĩa sự thành công của ứng dụng.**

Để chuẩn bị dữ liệu huấn luyện mô hình dự đoán, ta cần một tiêu chí để đánh giá sự thành công cho các ứng dụng nhằm gán nhãn cho dữ liệu. Có nhiều cách đánh giá một ứng dụng được xem là thành công hay không, cụ thể như [6]:



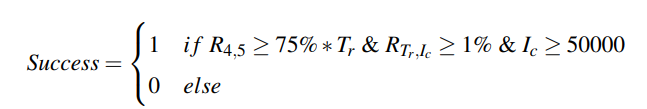
1. Công thức đánh nhãn sự thành công (1)

Trong công thức này có một vấn đề là đối với những ứng dụng có rất ít số lượng đánh giá, thì những đánh giá sau có thể ảnh hưởng rất lớn đến độ thành công. Cụ thể với ứng dụng có 100 đánh giá, và đạt được đánh giá trung bình là 4.6 thì chỉ cần những đánh giá 1\*, 2\* xuất hiện thì đánh giá trung bình sẽ tụt không phanh. Cụ thể như các ứng dụng sau:



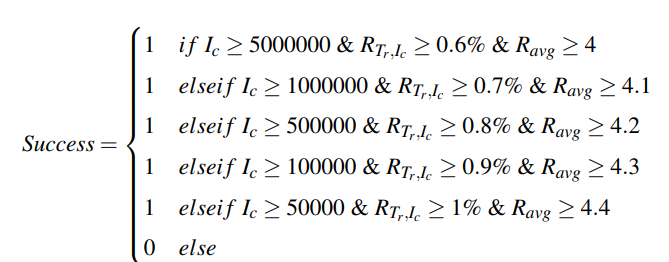
1. Các trường hợp ngoại lệ cho công thức (1)

Để cải thiện việc này thì có thể sử dụng công thức sau [3]:



1. *Công thức đánh nhãn sự thành công (2)*

Tuy nhiên với công thức này thì những ứng dụng có lượt tải về quá lớn thì ta không thể yêu cầu tỉ lệ trên ≥ 1%. Nên ta có thể điều chỉnh công thức với những mức tải về khác nhau. Nhóm có tham khảo công thức được đề xuất trong [3]:



1. *Công thức cho đô*

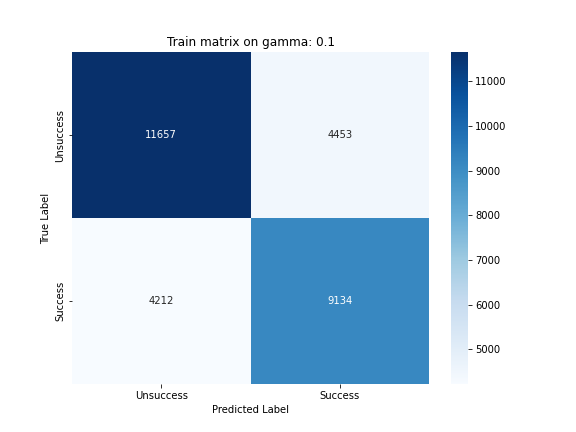
Công thức này có thể điều chỉnh để phù hợp hơn. Hoặc có thể có những cách khác tốt hơn, nhưng nhóm quyết định sử dụng công thức này cho bài toán này. Bằng công thức này chúng tôi thu được 19009 apps được gán nhãn thành công và 23072 apps được gán nhãn không thành công.

Thực hiện thêm trường success vào bộ dữ liệu trong file train.csv (được tạo trong quá trình tiền xử lý dữ liệu). Vì trong công thức có sử dụng trực tiếp trường installs và trường rating (đánh giá trung bình) nên để tránh tình trạng mô hình tìm ra được quy luật do người lập trình xây dựng, dẫn đến mô hình không mang tính khách quan thì nhóm không đưa 2 trường này vào bộ dữ liệu huấn luyện. Cuối cùng, chúng tôi chia ngẫu nhiên bộ dữ liệu thành 2 phần training (80%) và test (20%) rồi thực hiện huấn luyện và đánh giá trên 3 mô hình: Super vector machine (SVM), Gradient Boosting, Random Forest.

1. **Super vector machine (SVM).**

Nhóm sử dụng SVM với kernal là ‘rbf’ và train thử trên các ‘gamma’ khác nhau (0.01, 0.1, 1), thì mô hình với ‘gamma’ là tốt nhất với 0.702%

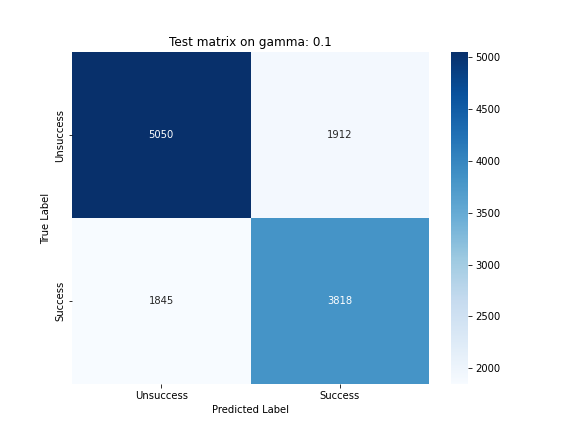
Với các confusion matrix cho việc dự đoán sự thành công của ứng dụng trên 2 tập train và test



1. Confusion matrix cho mô hình SVM trên tập train.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.74 | 0.72 | 0.73 | 16110 |
| Success | 0.67 | 0.68 | 0.67 | 13346 |
| Average/Total | 0.705 | 0.70 | 0.7 | 29456 |

1. Các giá trị đánh giá mô hình SVM trên tập train



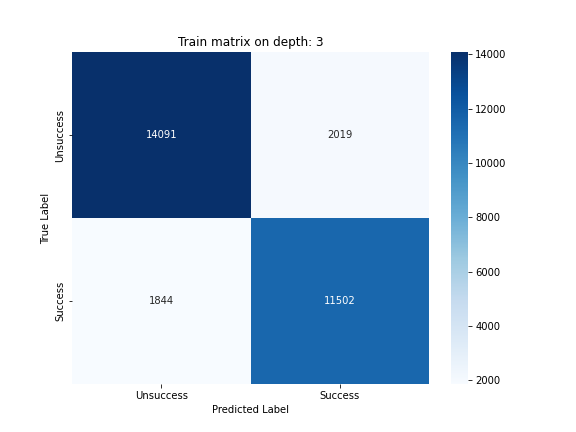
1. Confusion matrix cho mô hình SVM trên tập test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.73 | 0.73 | 0.73 | 6962 |
| Success | 0.67 | 0.67 | 0.67 | 5663 |
| Average/Total | 0.7 | 0.7 | 0.7 | 12625 |

1. Các giá trị đánh giá mô hình SVM trên tập test
2. **Gradient Boosting**

Nhóm sử dụng Gradient Boosting với các depth khác nhau (3, 5, 10) thì nhận thấy được rằng mô hình với depth là 3 thì có hiệu suất tốt nhất khi đánh giá trên tập test với độ chính xác 0.826%.

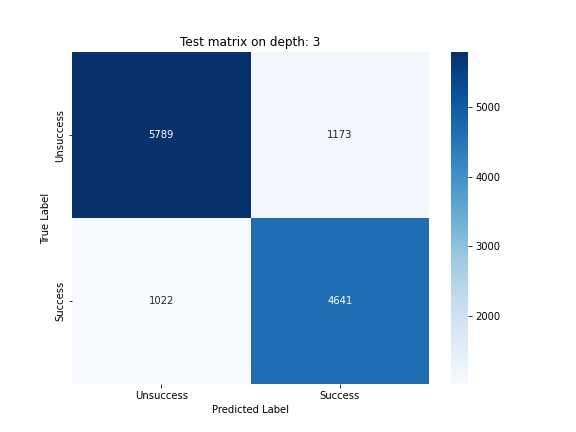
Với các confusion matrix cho việc dự đoán sự thành công của ứng dụng trên 2 tập train và test



1. Confusion matrix cho mô hình Gradient trên tập train.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.88 | 0.88 | 0.88 | 16110 |
| Success | 0.85 | 0.86 | 0.86 | 13346 |
| Average/Total | 0.87 | 0.87 | 0.87 | 29456 |

1. Các giá trị đánh giá mô hình Gradient Boosting trên tập train



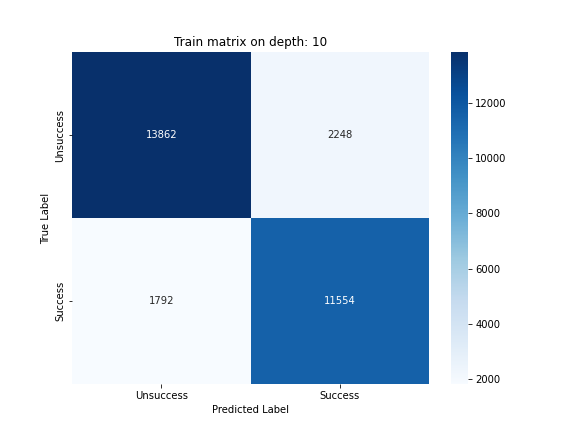
1. Confusion matrix cho mô hình Gradient Boosting trên tập test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.85 | 0.83 | 0.84 | 6962 |
| Success | 0.80 | 0.82 | 0.81 | 5663 |
| Average/Total | 0.83 | 0.83 | 0.83 | 12625 |

1. Các giá trị đánh giá mô hình Gradient Boosting trên tập test
2. **Random Forest.**

Nhóm sử dụng Random Forest với các depth khác nhau (3, 5, 10) thì nhận thấy được rằng mô hình với depth là 10 thì có hiệu suất tốt nhất khi đánh giá trên tập test với độ chính xác là 0.834%.

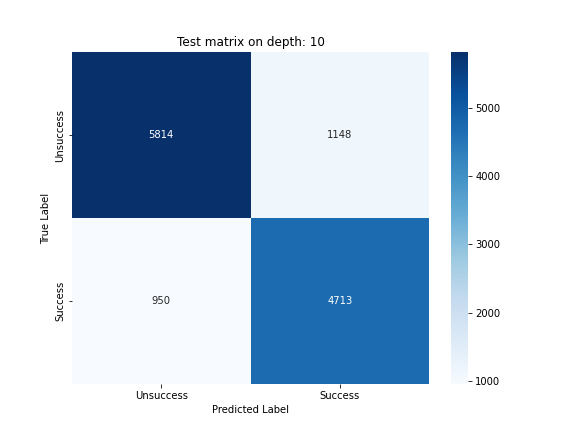
Với các confusion matrix cho việc dự đoán sự thành công của ứng dụng trên 2 tập train và test



1. Confusion matrix cho mô hình Random Forest trên tập train.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.89 | 0.86 | 0.85 | 16110 |
| Success | 0.84 | 0.87 | 0.86 | 13346 |
| Average/Total | 0.87 | 0.87 | 0.86 | 29456 |

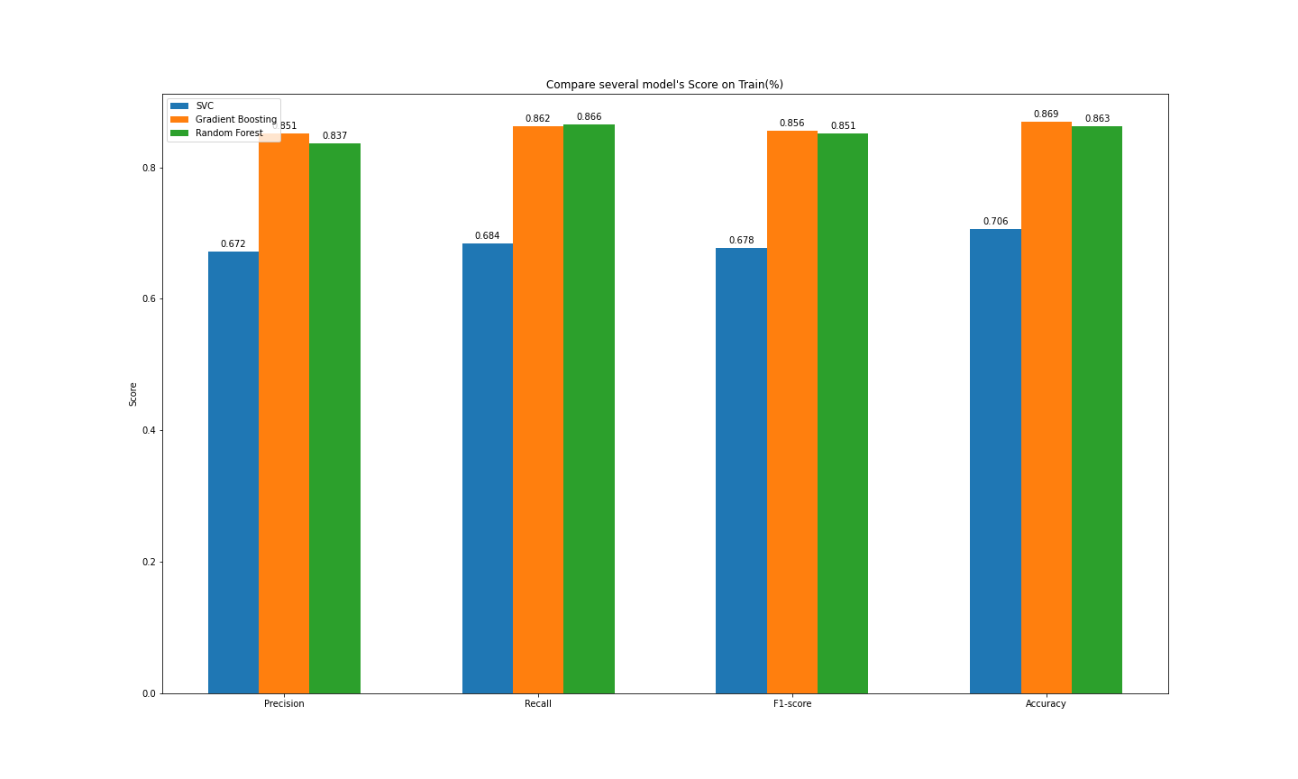
1. Các giá trị đánh giá mô hình Random Forest trên tập train



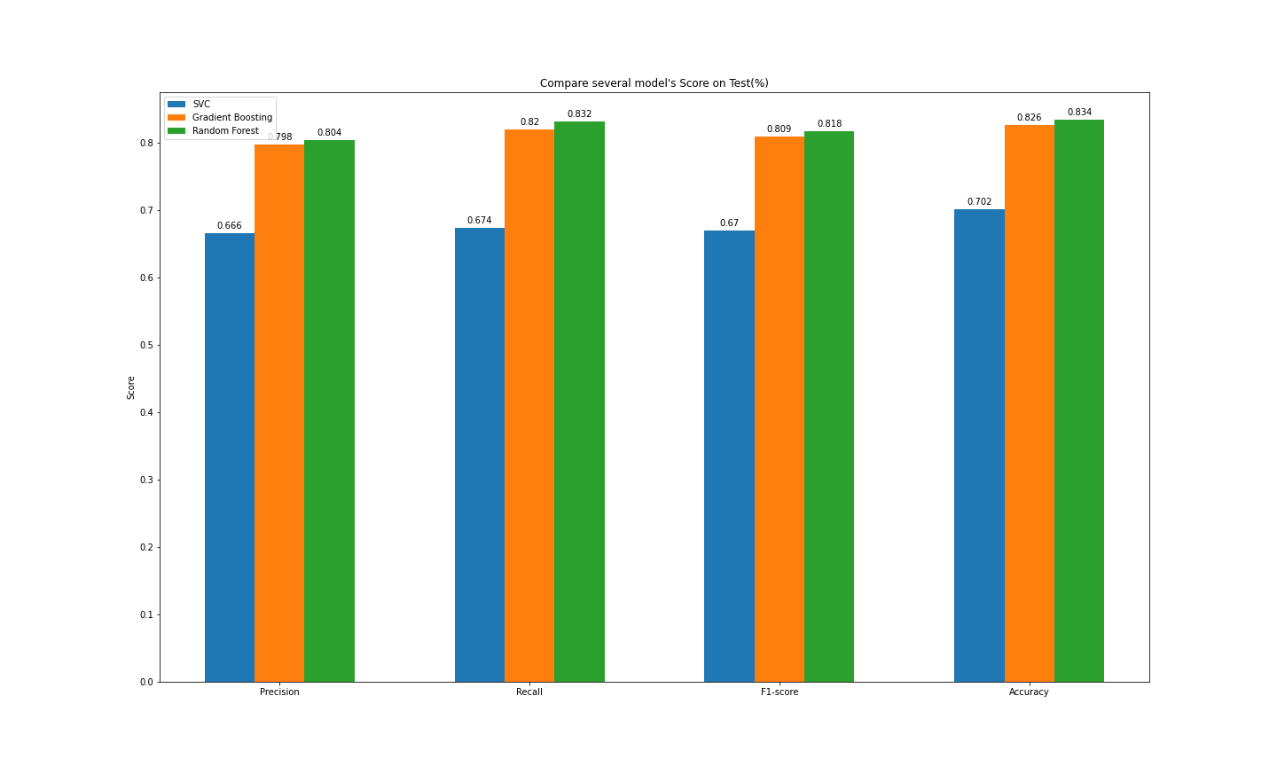
1. Confusion matrix cho mô hình Random Forest trên tập test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-Score | Support |
| Not Success | 0.86 | 0.84 | 0.85 | 6962 |
| Success | 0.80 | 0.83 | 0.82 | 5663 |
| Average/Total | 0.83 | 0.84 | 0.84 | 12625 |

1. Các giá trị đánh giá mô hình Random Forest trên tập test
2. **Đánh giá các mô hình và lựa chọn mô hình**



1. Đánh giá các mô hình trên tập train



1. Đánh giá các mô hình trên tập test

Với các biểu đồ trên, và các số liệu liên quan đến trên cả 4 độ đo precision, recall, F1-score và accuracy thì nhóm kêt luận rằng mô hình Gradient Boosting và Random Forest có độ chính xác khá tương đồng nhau (hơn 80%) và cao hơn nhiều so với mô hình SVM ( gần 70%) trên cả tập train và tập test . Có thể thấy khi đánh giá trên tập train thì mộ số chỉ số đánh giá trên mô hình Gradient Boosting tốt hơn Random Forest, tuy nhiên đưa vào thực nghiệm nhóm sẽ chọn mô hình Random Forest vì các thông số khi đánh giá trên tập test của Random Forest có phần tốt hơn so với Gradient Boosting, có nghĩa là mô hình này xử lý những dữ liệu mới tốt hơn dẫn đến khả năng tổng quát hóa của mô hình có thể tốt hơn.

# **PHẦN 6: KẾT LUẬN**

Qua quá trình phân tích, khám phá dữ liệu chúng tôi đề xuất các lời khuyên cho các nhà phát triển muốn xây dựng được một ứng dụng thành công đó là:  Nhà phát triển muốn thu hút lượng lớn người dùng thì nên chọn thể loại dựa trên số lượt cài đặt lớn thay vì dựa trên số lượng lớn ứng dụng trong thể loại đó. Ngoài ra cần phải chú trọng vào việc phát triển các ứng dụng phù hợp với mọi độ tuổi  và nên sử dụng số lượng từ ngắn gọn để đặt tên cho ứng dụng của mình. Hơn nữa họ cũng nên tập trung vào phát triển những ứng dụng chất lượng, thu hút người dùng thay vì tập trung quá nhiều vào việc phát triển quá nhiều ứng dụng cùng 1 lúc trên GPS. Thị trường luôn thay đổi không ngừng và điều này yêu cầu nhà phát triển phải luôn cập nhật các công nghệ mới nhất để không bị tụt lại. Cuối cùng, khi người dùng gặp sự cố với phiên bản mới, thì họ có thể thay đổi lại rating với số sao thấp hơn trong quá khứ làm cho rating trung bình ứng dụng tụt dốc nhanh chóng. Nhà phát triển cần phải nhanh chóng thực hiện những điều chỉnh theo phản hồi của người dùng nhằm khắc phục sự cố để giúp tăng lại rating trung bình của ứng dụng.

Để đánh giá được sự thành công của 1 ứng dụng, những số liệu về lượt cài đặt, đánh giá tích cực của người dùng là chưa đủ. Chúng ta cần xem xét thêm nhiều tiêu chí để đánh giá, chẳng hạn như: Lượt tương tác giữa người dùng và ứng dụng (tần suất, thời gian sử dụng), doanh thu và lợi nhuận của ứng dụng mang lại cho nhà phát triển, mức độ ảnh hưởng của ứng dụng đến người dùng và sự lan truyền tích cực trên các nền tảng mạng xã hội khác. Để phát triển thị trường sử dụng thì không nên gói gọn ứng dụng chỉ trên điện thoại mà phải mở rộng ra các nền tảng web, ứng dụng dành cho máy tính bảng, thiết bị đeo thông minh hoặc TV thông minh.

Trong báo cáo này, chúng tôi đã huấn luyện mô hình học máy như SVM, gradient boosting và random forest cho nhiệm vụ dự đoán sự thành công ứng dụng, kết quả mô hình random forest đạt được độ chính xác cao nhất (>80%). Nếu hiện tại ứng dụng chưa được thành công thì có thể sử dụng mô hình dự đoán này để dự đoán xem liệu trong tương lai ứng dụng đó có thể  thành công hay không.

# **TÀI LIỆU THAM KHẢO**

1. Sangeeta Rathod, Prajakta Ugalmugle, Ruchita Waghmare, Tabassum Maktum, An Approach To Predict Software Application Success Using Voting Ensemble Method, 2020.
2. Hadiqa Amanullah, Mishal Fatima, Umair Muneer, Sadaf Ilyas, Rana Abdul, Ibraheem Afzal, Causal Impact Analysis on Android Market, 2019.
3. Abdul Mueez, Khushba Ahmed, Tuba Islam, Waqqas Iqbal, Exploratory Data Analysis and Success Prediction of Google Play Store Apps, 2018.
4. Rimsha Maredia, Analysis of Google Play Store Data set and predict the popularity of an app on Google Play Store, 2020.
5. Aayush Gupta1, Deepali Kamthania, Study of sentiment on Google Play Store Applications, 2021.
6. Tuckerman. Predicting mobile application success. 2014

# **ZEPPELIN REPORT**

