DỰ ĐOÁN CẢM XÚC TỪ TWITTER VÀ KHẢ NĂNG MẮC BỆNH TIỂU ĐƯỜNG BẰNG MÔ HÌNH HỌC MÁY

242_CS320_02_FinalProject_Group5

Nguyễn Minh Đức
A36214
TI32h1
ThangLong University

Vũ Kim An
A40076
TT33h6
ThangLong University

Đinh Xuân Hùng
A42568
TT34e1
ThangLong University

Phạm Văn Tài
A42661
TI34h2
ThangLong University

Trần Huyền Trang
A46350
TA35CL01
ThangLong University

***Abstract*— Với sự bùng nổ của mạng xã hội, Twitter đã trở thành nguồn dữ liệu quý giá cho việc phân tích cảm xúc, đặc biệt trong các lĩnh vực như tiếp thị, chính trị và xã hội. Cùng lúc đó, bệnh tiểu đường là một mối quan tâm y tế quan trọng, gây ra nhiều biến chứng nguy hiểm. Nghiên cứu này khám phá hai lĩnh vực chính: phân tích cảm xúc Twitter và dự đoán bệnh tiểu đường. Trong phân tích cảm xúc, chúng tôi sử dụng bộ dữ liệu 1,6 triệu tweet để đánh giá cảm nhận của công chúng, trong khi ở dự đoán bệnh tiểu đường, chúng tôi phân tích dữ liệu từ 768 bệnh nhân để dự đoán nguy cơ mắc bệnh tiểu đường. Các bước tiền xử lý dữ liệu như lemmatization, tokenization và xử lý ngoại lệ đã được áp dụng cho cả hai bộ dữ liệu. Chúng tôi đánh giá nhiều mô hình cho cả hai tác vụ, bao gồm Naive Bayes và Deep Learning (LSTM) cho phân tích cảm xúc, và Logistic Regression, Support Vector Machine (SVM), và Decision Tree cho dự đoán bệnh tiểu đường. Kết quả cho thấy, Deep Learning vượt trội hơn Naive Bayes trong phân tích cảm xúc, đạt Accuracy 78,00% và F1-Score là 78,71%. Đối với dự đoán bệnh tiểu đường, Logistic Regression đạt độ Accuracy 78,00% và AUC-ROC là 0,85, là mô hình hiệu quả nhất trong ứng dụng y tế. Mặc dù mỗi mô hình có điểm mạnh riêng, nhưng cũng tồn tại những hạn chế về độ phức tạp tính toán và khả năng xử lý dữ liệu nhiễu. Tổng kết, Deep Learning là phương pháp hiệu quả nhất cho phân tích cảm xúc phức tạp, trong khi Logistic Regression được khuyến nghị trong chẩn đoán y tế nhờ vào tính ổn định và dễ hiểu.**

***Keywords—Twitter, bệnh tiểu đường, Naive Bayes, LSTM, Logistic Regression, SVM,* Decision Tree**

I. Introduction

A. Dự đoán bệnh tiểu đường

Bệnh đái tháo đường (hay còn gọi là bệnh tiểu đường) là một tình trạng bệnh lý rối loạn chuyển hóa không đồng nhất, có đặc điểm tăng lượng đường huyết trong cơ thể. Nguyên nhân thường là do nồng độ insulin trong cơ thể không ổn định (có thể thiếu thậm chí thừa). Nếu bị đái tháo đường mà bạn kiểm soát được lượng đường trong máu và thường xuyên theo dõi tốt thì chắc chắn lượng đường nằm trong mức an toàn gần như người bình thường.

Dựa vào đặc điểm và diễn biến của bệnh chia ra có các loại đái tháo đường: typ1, typ2, đái tháo đường thứ phát và đái tháo đường thai kỳ[1]...Mục tiêu nhóm tôi nghiên cứu lần này là dự đoán khả năng mắc bệnh tiểu đường của bệnh nhân thông qua việc phân tích các dữ liệu y tế. Tiểu đường làm tăng nguy cơ biến chứng tim mạch, làm giảm chức năng thận, dẫn đến mất khả năng lọc chất thải từ máu.Bên cạnh đó, nếu bạn bị rối loạn tiểu đường, khả năng bạn bị đột quỵ cao gấp 1,5 người không mắc. Vì vậy, việc phát hiện sớm bệnh tiểu đường có vai trò rất quan trọng. Chúng tôi sẽ dùng các mô hình Học Máy để xây dựng mô hình dự đoán, như: Logistic Regression, SVM, Decision Tree.

B. Dự đoán cảm xúc từ Twitter

Twitter, là một phương tiện truyền thông mạng xã hội và dịch vụ mạng xã hội trực tuyến được điều hành bởi X Corp., công ty kế thừa của Twitter, Inc.

Twitter cho phép người sử dụng đọc, nhắn và cập nhật các mẩu tin nhỏ gọi là tweets, một dạng tiểu blog. Những mẩu tweet được giới hạn tối đa 280 ký tự được lan truyền nhanh chóng trong phạm vi nhóm bạn của người nhắn hoặc có thể được trưng rộng rãi cho mọi người. Thành lập từ năm 2006, Twitter đã trở thành một hiện tượng phổ biến toàn cầu. Những tweet có thể chỉ là dòng tin vặt cá nhân cho đến những cập nhật thời sự tại chỗ kịp thời và nhanh chóng hơn cả truyền thông chính thống. Dữ liệu này đã trở thành nguồn tài nguyên quý giá, không chỉ phản ánh xu hướng dư luận mà còn hỗ trợ trong việc ra quyết định của các tổ chức và doanh nghiệp.

II. Background

A. Dự đoán bệnh tiểu đường

Bệnh tiểu đường là một vấn đề y tế toàn cầu, có thể gây các biến chứng nghiêm trọng nếu không được chẩn đoán sớm. Tính đến thời điểm hiện tại, tiểu đường là căn bệnh nguy hiểm xếp thứ 3, chỉ sau các bệnh lý tim mạch và ung thư. Tiểu đường biến chứng gây ảnh hưởng nhiều đến chất lượng cuộc sống và sức khỏe của người bệnh. Những người mắc bệnh tiểu đường phụ thuộc vào insulin. Insulin giúp glucose đi vào các tế bào để cung cấp năng lượng và người bị bệnh tiểu đường thường phải sử dụng nó để kiểm soát lượng đường trong máu. Có nghiên cứu cho rằng, “giá trung bình hàng năm cho một lọ insulin 10ml ở các nước thu nhập thấp là khoảng 12-20 USD, trong khi ở các nước thu nhập cao, giá này thấp hơn đáng kể. ” [4]. Qua đó, nhóm tôi thấy việc nghiên cứu dự đoán bệnh nhân có nguy cơ mắc bệnh tiểu đường có thể giúp cho bác sĩ có những can thiệp kịp

thời với các biện pháp phòng ngừa hoặc điều trị sớm. Bên cạnh đó, còn giúp giảm thiểu các tác động tiêu cực đến sức khỏe và nâng cao chất lượng cuộc sống cho bệnh nhân.

Mục tiêu nghiên cứu của nhóm tôi cho mô hình này là chỉ số Recall >0.7 và AUC-ROC > 0.8. Mục tiêu này được đưa ra bởi sự đánh giá của cá nhân. Và huấn luyện với 3 mô hình Học máy phân loại phổ biến: Logistic Regression, SVM, Decision Tree.

Nguồn dữ liệu được lấy từ Kaggle. Bao gồm hơn 700 dữ liệu và 9 cột giá trị về các chỉ số y tế của bệnh nhân.

B. *Dự đoán cảm xúc từ Twitter*

Với sự bùng nổ của các nền tảng mạng xã hội, đặc biệt là Twitter, hàng triệu bài đăng (tweets) được chia sẻ mỗi ngày, mang theo những ý kiến, cảm xúc và quan điểm về đa dạng chủ đề từ chính trị, xã hội, kinh tế cho đến các sản phẩm và dịch vụ thương mại. Phân tích cảm xúc từ các tweet mang lại nhiều ứng dụng thực tế, như: phát hiện xu hướng xã hội, các tổ chức có thể theo dõi ý kiến của cộng đồng về các vấn đề thời sự, giúp đưa ra quyết định phù hợp và cải thiện chính sách công. Bên cạnh đó, việc phân tích giúp đánh giá thương hiệu các công ty, từ đó hiểu cảm nhận của khách hàng về sản phẩm hoặc dịch vụ của mình, và cải thiện trải nghiệm khách hàng. Đặc biệt là sẽ có ích cho việc phân tích rủi ro, các lĩnh vực như tài chính và bảo hiểm có thể sử dụng dữ liệu từ Twitter để phát hiện sớm các nguy cơ hoặc cơ hội dựa trên tâm lý chung của cộng đồng.

Mục tiêu nghiên cứu của nhóm tôi cho mô hình là phân tích để phân loại và dự đoán các 'tweet' chia thành 0- cảm xúc tích cực và 4 - cảm xúc tiêu cực. Huấn luyện với 2 mô hình là: Naive Bayes và Deep Learning (LTSM).

Nguồn dữ liệu được lấy từ Kaggle. Bộ dữ liệu chứa 1.6 triệu tweets đã được gán nhãn tự động dựa trên emoticons. Các tweets được đăng bởi các người dùng khác nhau trong khoảng thời gian từ ngày 6 tháng 4 đến ngày 29 tháng 5 (năm 2009).

III. Approach

Với mỗi mô hình khi huấn luyện, ta đều phải kiểm tra và xử lý dữ liệu để có thể huấn luyện ra được một mô hình tối ưu nhất và chính xác nhất.

A. *Dự đoán mắc bệnh tiểu đường*

Trước khi xử lý dữ liệu, chúng tôi đã thử huấn luyện với tập dữ liệu ban đầu và kết quả là Recall ~0.6 và AUC-ROC ~0.6. Các chỉ số này khá ổn, chưa được tốt nên chúng tôi tiến hành kiểm tra và xử lý một số vấn đề mà dữ liệu gặp phải.

1) *Phát hiện và xử lý một số giá trị 0 bất thường*

Glucose (min = 0): Glucose không thể bằng 0 trong thực tế. Đây có thể là dữ liệu bị thiếu (missing data) được thay bằng 0.

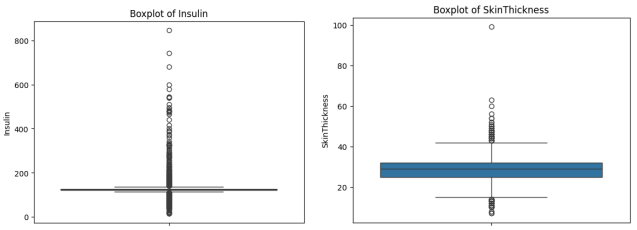
Blood Pressure (min = 0): Huyết áp cũng không thể bằng 0. Đây cũng là trường hợp dữ liệu bị thiếu.

Skin Thickness (min = 0): Độ dày da không thể bằng 0. Đây cũng là giá trị bất hợp lý.

Insulin (min = 0): Giá trị insulin bằng 0 có thể đại diện cho dữ liệu bị thiếu, vì giá trị này thường có ý nghĩa y tế.

Xử lý các biến bất thường 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI' bằng cách tính toán median cho từng cột, chỉ sử dụng các giá trị khác 0, rồi Thay thế giá trị 0 trong các cột bằng median.

2) *Phát hiện và xử lý ngoại lệ*



Các biến 'Insulin', 'Skin Thickness', 'Diabetes Pedigree Function' có nhiều ngoại lệ nên chỉ tập trung xử lý ngoại lệ với 3 cột dữ liệu này. Phương pháp sử dụng là phương pháp IQR. IQR = Q3-Q1, cận trên = Q3+ 1.5*IQR, cận dưới = Q1- 1.5*IQR. Tính toán cận trên, cận dưới để loại bỏ ngoại lê. Tiếp theo, dùng phương pháp cắt giới hạn (Clipping) và thay thế các giá trị ngoại lệ trong một cột của bộ dữ liệu bằng các giá trị cận dưới (lower bound) hoặc cận trên (upper bound) được tính từ IQR.Nếu giá trị nhỏ hơn cận dưới, nó sẽ được thay thế bằng cận dưới. Nếu giá trị lớn hơn cận trên, nó sẽ được thay thế bằng cận trên. Nếu giá trị nằm trong khoảng giữa cận dưới và cận trên, giá trị không bị thay đổi.

3) *Dữ liệu không cân bằng*

Outcome	count
0	500
1	268

Giá trị 0 chiếm 65%, cho thấy dữ liệu không cân bằng.Áp dụng phương pháp SMOTE để cân bằng dữ liệu. Phương pháp SMOTE được áp dụng để cân bằng dữ liệu bằng cách tạo thêm các mẫu nhân tạo từ lớp thiểu số (lớp mắc bệnh), giúp mô hình học tốt hơn và giảm nguy cơ overfitting.

4) *Huấn luyện mô hình*

4.1. Logistic Regression

Logistic Regression là một thuật toán học máy thuộc nhóm phân loại tuyến tính, sử dụng hồi quy tuyến tính để dự đoán xác suất và phân loại đầu ra.[5]
Dự đoán xác suất: Logistic Regression sử dụng hàm sigmoid để biến đầu ra của hồi quy tuyến tính thành giá trị xác suất. Công thức:

$$P(y=1|X) = \sigma(z) = \frac{1}{1+e^{-z}}$$

Hàm mất mát: Hàm Log-Loss, thể hiện sự chênh lệch giữa xác suất dự đoán và nhãn thực tế:

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \log(h(x_i)) + (1-y_i)\log(1-h(x_i))\right]$$

4.2. SVM

SVM là thuật toán phân loại mạnh mẽ, thuộc nhóm phân loại phi tuyến tính hoặc tuyến tính, hoạt động dựa trên việc tìm siêu phẳng (hyperplane) tối ưu để phân biệt các lớp.

Hàm quyết định: SVM tìm siêu phẳng $w \cdot x+b=0$ sao cho khoảng cách giữa siêu phẳng và các điểm gần nhất (support vectors) là lớn nhất. [6]

$$\text{Maximize: } \frac{2}{||w||}$$

Hàm mất mát:

$$\mathcal{L}(w,b) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i(w \cdot x_i + b))$$

4.3. Decision Tree

Decision Tree là thuật toán thuộc nhóm học có giám sát, dựa trên việc chia nhỏ dữ liệu thành các nhóm dựa trên tiêu chí tối ưu hóa thông tin. Tuned Decision Tree là phiên bản được điều chỉnh siêu tham số để cải thiện hiệu suất.[7]

Xây dựng cây: Cây quyết định được xây dựng bằng cách chia nhỏ dữ liệu dựa trên tiêu chí lựa chọn tốt nhất tại mỗi nút (node). Các tiêu chí phổ biến:
Gini Impurity:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Entropy (Information Gain):

$$Entropy = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

5) Quy trình huấn luyện

Bước 1: Chia tệp huấn luyện thành 2 phần là train và test theo tỉ lệ 70-30.
Bước 2: Chuẩn hóa dữ liệu với phương pháp Standardscaler
Bước 3: Huấn luyện với từng mô hình

B. Dự đoán cảm xúc bài Twitter

1) Lọc cột giá trị

Bộ dữ liệu có 6 cột, như: 'label', 'date', 'time', 'query', 'username', 'text'. Chúng ta chỉ quan tâm đến cột 'text' và 'label', nên đầu tiên sẽ loại bỏ các cột không cần dùng đến (time, date, query, username). Vì bộ dữ liệu có tổng cộng 1,600,000 mẫu, đây là một tập dữ liệu lớn, đảm bảo đủ thông tin để mô hình machine learning hoặc deep learning đạt được hiệu suất tốt. Số lượng mẫu lớn cũng góp phần giảm nguy cơ overfitting, từ đó giúp mô hình tổng quát hóa tốt hơn khi áp dụng vào các tập dữ liệu thực tế. Vì vậy chúng tôi giảm dữ liệu xuống ¼ để xử lý nhanh hơn. Điều đó giúp giữ cân bằng giữa positive và negative samples, cũng như là tăng tốc độ xử lý mà vẫn đảm bảo đủ dữ liệu để train mô hình hiệu quả. Mô hình đã cân bằng dữ liệu:

label	count
0	800000
4	800000

2) Tokenization và xử lý text

TweetTokenizer là công cụ chuyên biệt để xử lý tweets, giúp giảm độ dài của các từ lặp lại và phân tách ra từng từ. Ví dụ: "Hellooooooo" -> "Hellooo", chuyển text và label thành lists riêng biệt, Tokenize mỗi tweet thành list các từ, chuyển label 4 thành 1 (positive) và 0 giữ nguyên (negative).

3) Chuẩn hóa từ

Chúng tôi sử dụng Lemmatization là một kỹ thuật trong xử lý ngôn ngữ tự nhiên (NLP). Nó là quá trình chuyển đổi các dạng khác nhau của từ về dạng gốc của nó, được gọi là "lemma". Ví dụ, các từ như "running", "ran", và "runs" đều được chuyển về dạng gốc là "run". Tương tự, "better" sẽ được chuyển về "good".

4) Clean data

Các bước làm sạch: Loại bỏ stopword (là những từ được sử dụng thường xuyên (chẳng hạn như "the", "a", "an", "in") không có bất kỳ ý nghĩa nào hữu ích để trích xuất cảm xúc.)
- Chuyển text về lowercase
- Loại bỏ URLs và mentions (@username)
- Loại bỏ dấu câu
- Loại bỏ từ có độ dài ≤ 2

Viết hàm cleaner để : Xử lý các trường hợp đặc biệt, ví dụ: "u" -> "you", "r" -> "are", chuẩn hóa các từ viết tắt phổ biến. Sau đó, sử dụng để chuyển đổi danh sách các token đã được làm sạch thành một cấu trúc từ điển (dict). Cuối cùng loại bỏ "noise" (tạp âm) từ danh sách các tokens. Tạp âm có thể bao gồm các ký tự không mong muốn, các từ không có ý nghĩa (stop words) hoặc các yếu tố khác làm giảm chất lượng dữ liệu.

```
[({'love': True, 'guy': True, 'best': True}, 1),
 ({'meet': True,
   'one': True,
   'besties': True,
   'tonight': True,
   'cant': True,
   'wait': True,
   'girl': True,
   'talk': True}.
```

5) Trực quan hóa dữ liệu

Sử dụng công cụ WordCloud. Là một công cụ trực quan hóa dữ liệu, hiển thị tần suất xuất hiện của các từ trong văn bản. Các từ xuất hiện thường xuyên hơn sẽ có kích thước lớn hơn trong hình ảnh.

Positive words.

6) Huấn luyện mô hình

6.1. Naive Bayes

Naive Bayes là một mô hình học giám sát dựa trên xác suất rất phổ biến trong Machine Learning. Định lý Bayes được sử dụng để tính xác suất của một lớp dựa trên các đặc trưng đầu vào:
$P(C|X)=P(X|C)\cdot P(C) / P(X)$

Naive Bayes giả định rằng các đặc trưng là độc lập với nhau, tức là: $P(X \mid C) = P(x1|C).P(x2|C)\cdots P(xn|C)$. Ưu điểm của Naive Bayes là đơn giản và nhanh chóng, hiệu quả với dữ liệu lớn.

6.2. Deep Learning Model

Tạo một mô hình Deep Learning trong Keras, bao gồm các lớp như Embedding, LSTM và Dense, để dự đoán cảm xúc từ các bài đăng trên Twitter. LSTM là một loại mạng

nơ-ron hồi tiếp (Recurrent Neural Network - RNN) đặc biệt, được thiết kế để giải quyết vấn đề lãng quên (vanishing gradient problem) trong các mô hình RNN truyền thống. LSTM được sử dụng rộng rãi trong các bài toán học máy có dữ liệu chuỗi thời gian, như dự đoán chuỗi, nhận diện giọng nói, dịch ngôn ngữ tự động, và nhiều ứng dụng khác.

LSTM có một cấu trúc phức tạp hơn so với RNN truyền thống, với ba thành phần chính:

Cổng Quên (Forget Gate): Quyết định thông tin nào sẽ bị quên đi, tức là thông tin nào trong trạng thái nhớ (cell state) của mạng sẽ bị loại bỏ.

Cổng Nhập (Input Gate): Quyết định thông tin nào sẽ được thêm vào trạng thái nhớ, tức là các giá trị mới sẽ được cập nhật vào bộ nhớ của mạng.

Cổng Xuất (Output Gate): Quyết định thông tin nào sẽ được xuất ra ngoài, tức là thông tin nào trong trạng thái nhớ sẽ ảnh hưởng đến đầu ra của mạng trong thời điểm đó.

f_t, i_t, o_t tương ứng với forget gate, input gate và output gate.

- Forget gate: $f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f)$
- Input gate: $i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i)$
- Output gate: $o_t = \sigma(U_o * x_t + W_o * h_{t-1} + b_o)$

IV. Results

Sau khi thực hiện các bước xử lý dữ liệu và chạy với các mô hình huấn luyện, chúng tôi đã có những kết quả như sau:

A. Dự đoán mắc bệnh tiểu đường

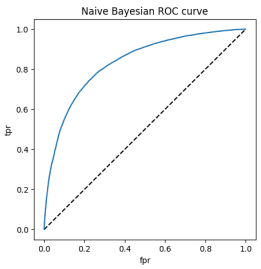
```
Logistic Regression:
  Accuracy: 0.7800
  Precision: 0.7673
  Recall: 0.8079
  F1-Score: 0.7871
  AUC-ROC: 0.8503
  Confusion Matrix:
[[112  37]
 [ 29 122]]

SVM:
  Accuracy: 0.7667
  Precision: 0.7647
  Recall: 0.7748
  F1-Score: 0.7697
  AUC-ROC: 0.8369
  Confusion Matrix:
[[113  36]
 [ 34 117]]

Tuned Decision Tree:
  Accuracy: 0.7700
  Precision: 0.7228
  Recall: 0.8808
  F1-Score: 0.7940
  AUC-ROC: 0.8029
  Confusion Matrix:
[[ 98  51]
 [ 18 133]]
```

Nhìn chung tất cả các mô hình đã đạt được mục tiêu chúng tôi đề ra (recall >0.7 và AUC-ROC > 0.8).. Logistic Regression có tổng thể hiệu suất tốt nhất, với độ chính xác, Precision, Recall, F1-Score và AUC-ROC cao nhất. Đây là mô hình khá cân bằng giữa Precision và Recall.SVM có hiệu suất thấp hơn một chút so với Logistic Regression, đặc biệt ở các chỉ số Precision, Recall và AUC-ROC. Decision Tree có Recall rất cao (80.8%), nhưng Precision thấp và AUC-ROC thấp, cho thấy mô hình này dễ bỏ sót các bệnh nhân mắc bệnh nhưng lại đưa ra nhiều dự đoán sai về những người không mắc bệnh.

B. Dự đoán cảm xúc bài twitter

1) Mô hình Naive Bayes

Trước khi huấn luyện, chia thành 2 phần: 90% dùng để huấn luyện và 10% dùng để kiểm tra hiệu suất của mô hình. Mô hình Naive Bayes được huấn luyện trên tập dữ liệu huấn luyện, học cách phân loại các tweets dựa trên các từ và nhãn cảm xúc tương ứng. Accuracy trên tập huấn luyện (0.8108); Accuracy trên tập kiểm tra (0.7557). Mô hình có thể hoạt động khá tốt trên dữ liệu không thấy trước đó nhưng chưa hoàn hảo.



Predicted	0	1	All
Actual			
0	16532	3448	19980
1	6334	13686	20020
All	22866	17134	40000

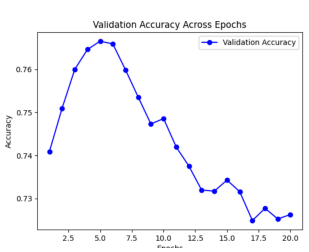
AUC = 0.834, đây là một kết quả khá tốt, cho thấy mô hình có khả năng phân biệt tốt giữa các classes.

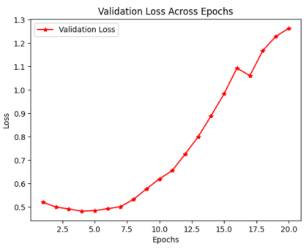
2) Mô hình Deep Learning (LTSM)

Bước 1: Lọc và chuyển đổi dữ liệu: Xử lý các từ không tìm thấy và chuyển đổi dữ liệu.

Bước 2: Tạo một lớp Embedding trong Keras và điền vào các trọng số của nó với ma trận embedding. Sau đó định nghĩa một mô hình Sequential trong Keras, bao gồm một lớp embedding, một cặp lớp Bidirectional LSTM và cuối cùng là 1 lớp sigmoid để tạo đầu ra từ 0 đến 1.

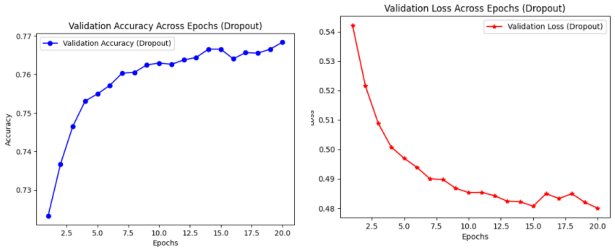
Bước 3: Sử dụng Adam optimizer và chỉ số (metric) để biên dịch mô hình với hàm mất mát binary cross-entropy.





Độ chính xác của mô hình trên tập huấn luyện đang tăng vọt, vượt quá 95% sau 20 epoch! Tuy nhiên, độ chính xác trên tập validation chỉ tăng nhẹ trong các epoch đầu tiên, đạt 78,4% vào epoch thứ 5, sau đó trải qua một quá trình giảm dần đều. Trong khoa học dữ liệu, chúng ta sẽ phân loại mô hình này là có độ biến động (variance) rất cao và độ lệch (bias) thấp. Đây cũng được gọi là "overfitting" (quá khớp).

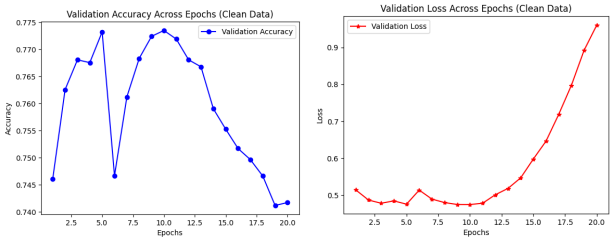
Bước 4: Sử dụng Dropout để giảm overfitting và cải thiện mô hình.



Mô hình đạt được 81,1% độ chính xác trên tập kiểm tra ở epoch thứ 5 chứng minh rằng chất lượng dữ liệu quyết định hiệu suất mô hình. Việc mô hình hoạt động tốt hơn sau khi làm sạch dữ liệu chứng tỏ rằng đầu vào đã được chuẩn bị tốt hơn sẽ giúp mô hình tổng quát tốt hơn.

Bước 5: Kiểm tra lại số lượng từ chưa biết và xử lý.

Khi kiểm tra lại, phát hiện 200 nghìn từ chưa biết, chiếm tương đương với 7% tổng số từ. Vì vậy cần xử lý lại dữ liệu. Giảm số lượng từ chưa biết từ 200K xuống 147K. Phần trăm từ chưa biết bây giờ, nó giảm đi 26%.



Mô hình đạt được 81,1% độ chính xác trên tập kiểm tra ở epoch thứ 5 chứng minh rằng chất lượng mô hình đã hoạt động tốt hơn sau khi làm sạch dữ liệu.

V. Conclusion

A. Dự đoán mắc bệnh tiểu đường

Với chủ đề "Dự đoán mắc bệnh tiểu đường", mô hình Logistic Regression (LR) là mô hình tốt nhất:

• Cân bằng giữa Precision và Recall => Mô hình có hiệu quả trong việc phân loại và giảm thiệt hại các trường hợp dự đoán sai

• Khả năng phân biệt lớp tốt (AUC = 0.8503) => Mô hình có khả năng phân biệt giữa các nhóm mắc bệnh và không mắc bệnh rất tốt.

Ưu điểm:

Phát hiện mắc bệnh hiệu quả: Ưu tiên recall giúp mô hình phát hiện được các bệnh nhân mắc bệnh tiểu đường.

Huấn luyện nhiều mô hình: giúp dễ dàng so sánh và chọn mô hình thích hợp nhất.

Nhược điểm:

Precision chưa cao: Dẫn đến nhiều dự đoán nhầm gây rắc rối cho bệnh nhân

Dữ liệu chưa đầy đủ: chưa bao quát đủ các yếu tố đặc trưng (như: mức độ stress, thói quen ăn uống,...) Và các mô hình chưa được tối ưu hóa hoàn toàn các tham số.

B. Dự đoán cảm xúc bài Twitter

Deep Learning là lựa chọn tối ưu với độ chính xác cao và khả năng xử lý ngữ cảnh.

• Naive Bayes đơn giản nhưng xử lý negative kém.

• Deep Learning tốt đối với ngữ cảnh phức tạp, nhưng dễ overfitting.

VI. Future Recommendation

A. Dự đoán mắc bệnh tiểu đường

• Thu thập thêm các đặc trưng như lịch sử gia đình, thói quen ăn uống để nâng cao hiệu quả dự đoán.

• Phát triển ứng dụng hỗ trợ chẩn đoán trên nền tảng web hoặc di động.

B. Dự đoán cảm xúc bài Twitter

• Kết hợp thêm dữ liệu từ các nền tảng mạng xã hội khác để nâng cao khả năng phân tích.

• Áp dụng các mô hình tiên tiến như Transformer (e.g., BERT) để cải thiện độ chính xác và khả năng hiểu ngữ cảnh phức tạp.

VII. Reference

[1] https://www.vinmec.com/vie/bai-viet/dau-hieu-som-bao-hieu-benh-tieu-duong-vi

[2] https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[4] https://iris.who.int/bitstream/handle/10665/204871/9789241565257_eng.pdf?sequence=1/

[5] Logistic Regression

[6] SVM

[7] Decision Tree

[8] Tokenization

[9] LSTM