ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN: MÔN KTDL

PHÂN LOẠI CUNG BẬC CẢM XÚC TRÊN MẠNG XÃ HỘI

GVHD: Thầy Mai Xuân Hùng

Nhóm sinh viên thực hiện:

1. Nguyễn Đức Hưng 16520478
2. Nguyễn Thanh Bảo 16520086

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN: MÔN KTDL

PHÂN LOẠI CUNG BẬC CẢM XÚC TRÊN MẠNG XÃ HỘI

GVHD: Thầy Mai Xuân Hùng

Nhóm sinh viên thực hiện:

1. Nguyễn Đức Hưng 16520478
2. Nguyễn Thanh Bảo 16520086

MỤC LỤC

I. GIỚI THIỆU

- Bài toán khai phá quan điểm (hay còn gọi la "phân tích cảm xúc") có nhiệm vụ xác định các yếu tố tình cảm, các đánh giá đến các đối tượng sản phẩm, dịch vụ, nhân vật hay các chủ đề qua ba mức: văn bản, câu, và khía cạnh.
- Hiện nay đã có nhiều công trình làm về phân tích cảm xúc cho các ngôn ngữ phổ biến như tiếng Anh, tiếng Hoa và tiếng Nhật nhưng với tiếng Việt, đây vẫn là một chủ đề "nóng" cần giải quyết.
- Đây là chủ đề hấp dẫn để nghiên cứu và phân tích về cảm xúc của người dùng thông qua các phương tiện mạng xã hội. Việc phân tích sentiment, giúp chúng ta lắng nghe và thấu hiểu những gì đang được nói trên các phương tiện thông tin đại chúng. Chủ đề nào được đề cập tới, nói như thế nào, tốt hay xấu, tốt về mặt nào và xấu về mặt nào và trạng thái của mọi người như thế nào: thích thú, buồn bã, giận dữ, bất ngờ, sợ hãi hoặc ghê tởm ?
- Trong những năm gần đây, việc Phân tích cảm xúc ngày càng trở nên phổ biến hơn do những ứng dụng thực tế mà nó mang lại: kinh doanh, tâm lý học, bảo mật, trí tuệ nhân tạo, …
- Mục đích của đề tài này nhằm nhận biết cảm xúc của mọi người khi bình luận trên mạng xã hội, cụ thể input của bài toán là những bình luận bằng tiếng Việt được tổng hợp trên mạng xã hội, output là một trong những label: thích thú, buồn bã, giận dữ, bất ngờ, sợ hãi, ghê tởm hoặc khác.
- Cụ thể với bộ dữ liệu từ kho văn bản UIT-VSMEC (gồm 6927 câu bình luận và được gắn nhãn cảm xúc), sử dụng 2 mô hình máy học Support Vector Machine(SVM) và Rừng ngẫu nhiên (Random Forest) cùng với 1 mô hình học sâu Bộ nhớ ngắn hạn dài (Long Short Term Memory - LSTM).

II.CÔNG TRÌNH LIÊN QUAN

- Năm 2011, Johan Bollen, Huina Mao và Alberto Pepe [1] đã viết một bài báo Mô hình hóa tâm trạng và cảm xúc với dữ liệu được tổng hợp từ Twitter được gắn 6 nhãn trạng thái: căng thẳng, trầm cảm, giận dữ, mạnh mẽ, mệt mỏi, nhầm lẫn.
- Năm 2018, Anshul Mitta và Arpit Goel [2] đã phát triển ra công trình Dự đoán chứng khoán bằng cách sử dụng phân tích tình cảm Twitter, bao gồm 4 nhãn: Calmm Happym Alert và Kind.
- Về thuật toán, năm 2017 Mohammad Rezwanul Huq, Ahmad Ali và Anika Rahman [3] đã kiểm tra mô hình máy học SVM kết hợp với chuẩn hóa và từ khóa bằng mô hình tìm kiếm dạng lưới (SVM with normalization and keyword base (5 features) with grid search) đã cho ra độ chính xác lên đến 77.97%.
- Năm 2018, YassineAl Amrani [4] đã xây dựng mô hình máy học SVM và Random Forest để Phân tích ý kiến, kết quả thu được mô hình SVM 82.4%, RF với 81%.
- Kratzwald đã kiểm tra hiệu quả của máy học-thuật toán (Rừng ngẫu nhiên và SVM) và thuật toán học sâu (Bộ nhớ ngắn dài hạn (LSTM) và Bộ nhớ ngắn hạn hai chiều (BiL-STM)) kết hợp với các pre-trained được đào tạo trước trên nhiều cảm xúc thực vật vào năm 2018 [5]. Ngoài ra, BiLSTM kết hợp với các các pre-trained đạt kết quả cao nhất với 58,2% số điểm F1 so với Rừng ngẫu nhiên và SVM với 52,6% và 54,2% tương ứng trên kho văn bản Tweets chung.

III.BỘ DỮ LIỆU

– Bộ dữ liệu lấy từ kho văn bản UIT-VSMEC, được thầy Nguyễn Văn Kiệt cung cấp. Cụ thể bộ dữ liệu bao gồm 3 tập:

1. Train: train_nor_811.xlsx
2. Test: test_nor_811. xlsx
3. Valid: valid_nor_811.xlsx

	Train		**Test**		**Valid**		**Total**	
Enjoyment	1558	28.08%	193	27.85%	214	31.20%	1965	28.37%
Disgust	1071	19.30%	132	19.05%	135	19.68%	1338	19.32%
Sadness	947	17.07%	116	16.74%	86	12.54%	1149	16.59%
Anger	391	7.05%	40	5.77%	49	7.14%	480	6.93%
Fear	318	5.73%	46	6.64%	31	4.52%	395	5.70%
Surprise	242	4.36%	37	5.34%	30	4.37%	309	4.46%
Other	1021	18.40%	129	18.61%	141	20.55%	1291	18.64%
Total	**5548**	100%	**693**	100%	**686**	100%	**6927**	100%

IV.CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

- Tách từ
    - Sử dụng thư viện dụng thư viện Untherthesea để tách từ tiếng Việt trong câu mà không làm mất đi nghĩa của nó.
    - VD: Hôm nay trời mưa to , sẽ tách thành : [“Hôm_nay”, ”trời”, ”mưa”, ”to”]
- Chuẩn hóa
    - Chuyển chữ hoa thành chữ thường
    - Chuyển label của 3 tập train, test, dev sang numeric cho các thuật toán SVM và Random Forest
    - Chuyển đổi các câu ở cột “Sentence” ra vector bằng phương pháp Word2vec

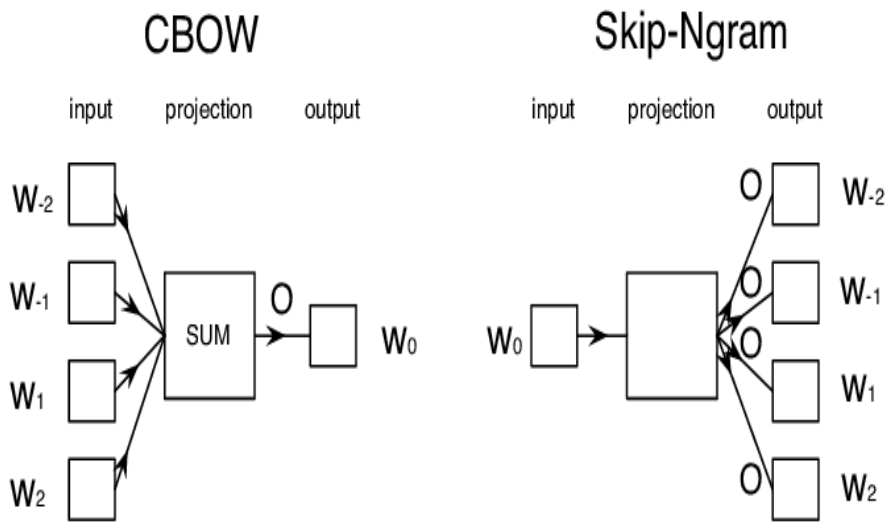
- Loại bỏ ký tự
    - Loại bỏ dấu câu, các ký tự đặc biệt, khoảng trắng, xóa các từ chỉ có một ký tự, và xóa các con số không cần thiết trong câu
    - Xóa emoji trong câu
    - Do stopword làm giảm độ chính xác của thuật toán nên tụi em không xóa stopword
- Tạo Vector
    - Chuyển đổi từ, câu sang vector bằng phương pháp Word2vec
    - Bộ pre-trained word embedding được tụi em sử dụng lấy từ bộ từ điển có sẵn được train từ BaoMoi

V. PHƯƠNG PHÁP

V.1. Word2vec

Ý tưởng:

- Word2Vec là một trong những mô hình về Word Embedding sử dụng mạng neural, có khả năng vector hóa từng từ dựa trên tập các từ chính và các từ văn cảnh... Về mặt toán học, Word2Vec là việc ánh xạ từ từ 1 tập các từ (vocabulary) sang 1 không gian vector, mỗi vector được biểu diễn bởi n số thực. Mỗi từ ứng với 1 vector cố định. Sau quá trình huấn luyện mô hình bằng thuật toán backprobagation, trọng số các vector của từng từ được cập nhật liên tục. Từ đó, ta có thể thực hiện tính toán bằng các khoảng cách quen thuộc như euclide, cosine, mahattan, ,..., những từ càng "gần" nhau về mặt khoảng cách thường là các từ hay xuất hiện cùng nhau trong văn cảnh, các từ đồng nghĩa, các từ cùng 1 trường từ vừng, ...
- Word2Vec bao gồm 2 cách tiếp cận chính, bao gồm: Cbow, Skip-gram
- Cả hai đều dựa trên một mạng neural network

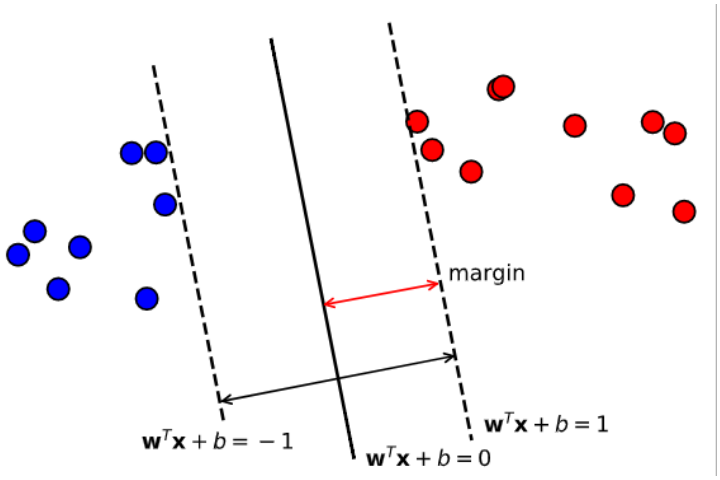


- CBOW model: ý tưởng chính của CBOW là dựa vào các context word (hay các từ xung quanh) để dự đoán center word (từ ở giữa). CBOW có điểm thuận lợi là training mô hình nhanh hơn so với mô hình skip-gram, thường cho kết quả tốt hơn với frequence words (hay các từ thường xuất hiện trong văn cảnh).
- Còn skip-gram thì ngược lại với CBOW, dùng target word để dự đoán các từ xung quanh. Skip-gram huấn luyện chậm hơn. Thường làm việc khá tốt với các tập data nhỏ, đặc biệt do đặc trưng của mô hình nên khả năng vector hóa cho các từ ít xuất hiện tốt hơn CBOW.

V.2. SVM

Ý tưởng:

- Kẻ đường thẳng để phân chia dữ liệu thành 2 phần sao cho khoảng các từ điểm gần nhất gần nhất của dữ liệu đến đường thẳng là lớn nhất. Khoảng cách đó là margin.
- Bài toán tối ưu SVM là bài toán đi tìm mặt phân cách sao cho *margin* tìm được là lớn nhất



Một số kernel thường dùng:

Kiểu hàm	Công thức
Linear Kernel	K(x,y) = x.y
Polynominal kernel	$K(x,y) = (x.y + 1)^d$
Radial basis function (Gaussian) kernel	$K(x,y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}}$
Hyperbolic tangent kernel	K(x,y) = tanh(a.x.y - b)

Ưu điểm:

- Được sử dụng rộng rãi vì cho ra kết quả tốt đối với những dữ liệu linearly separable.
- Xử lý trên không gian số chiều cao, tiết kiệm bộ nhớ
- Đối với những bài toán nonlinear separable thì có thể sử dụng kernel để đưa về miền dữ liệu mới.

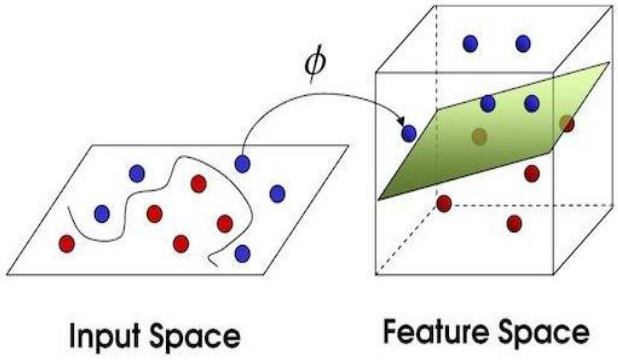


Figure 1 Thể hiện việc áp dụng kernel (nguồn: https://stats.stackexchange.com/questions/18030/how-to-select-kernel-for-svm)

Nhược điểm:

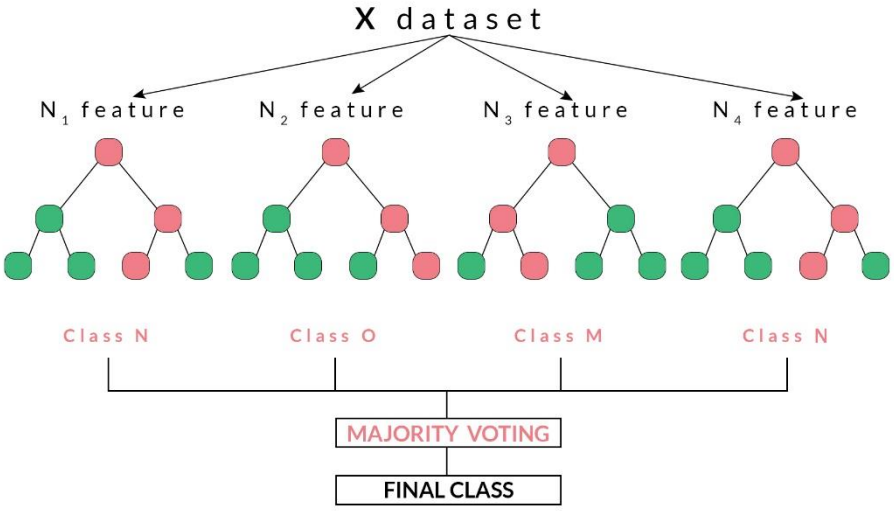
- Đối với SVM không sử dụng kernel hoạt động không hiệu quả đối với dữ liệu nhiễu hoặc dữ liệu gần linearly separable.
- Điểm yếu của SVM so với các thuật toán khác đặc biệt là deep learning đó là để SVM hiệu quả thì dữ liệu cần phải được trích-chọn các thuộc tính phù hợp, SVM không thể lựa chọn thuộc tính được nên các bạn phải tự lựa chọn thuộc tính hoặc dùng các thuật toán khác.

V.3. Random Forest

Ý tưởng:

- Random Forest là một tập hợp mô hình (ensemble). Mô hình Random Forest rất hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thể mạnh yếu khác nhau, nhưng theo nguyên tắc « wisdom of the crowd » (wisdom of the crowd là một ý tưởng cho rằng một nhóm lớn nói chung thông minh hơn một chuyên gia đơn lẻ khi đề cập đến việc giải quyết vấn đề, ra quyết định, đổi mới và dự đoán), ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

- Như tên gọi của nó, Random Forest (RF) dựa trên cơ sở : Random = Tính ngẫu nhiên, Forest = nhiều cây quyết định (decision tree)
- Đơn vị của RF là thuật toán cây quyết định, với số lượng hàng trăm. Mỗi cây quyết định được tạo ra một cách ngẫu nhiên từ việc : Tái chọn mẫu (bootstrap, random sampling) và chỉ dùng một phần nhỏ tập biến ngẫu nhiên (random features) từ toàn bộ các biến trong dữ liệu. Ở trạng thái sau cùng, mô hình RF thường hoạt động rất chính xác, nhưng đổi lại, ta không thể nào hiểu được cơ chế hoạt động bên trong mô hình vì cấu trúc quá phức tạp. RF do đó là một trong số những mô hình hộp đen (black box).



Ưu điểm:

- Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting.
- Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy.
- Random forests cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu.

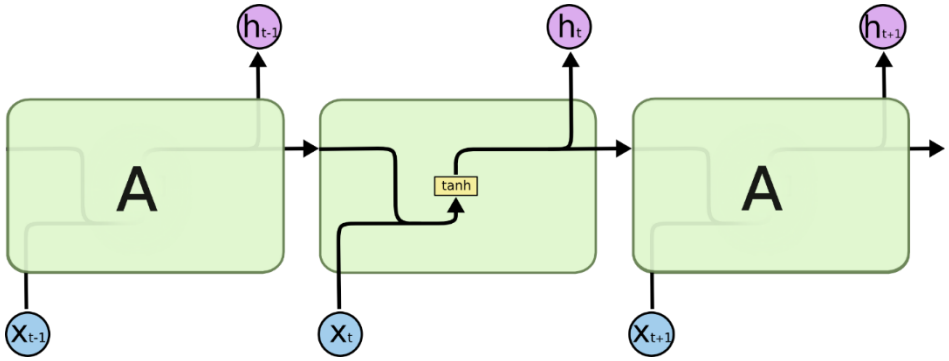
Nhược điểm

- Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào đưa ra dự đoán, tất cả các cây phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian.

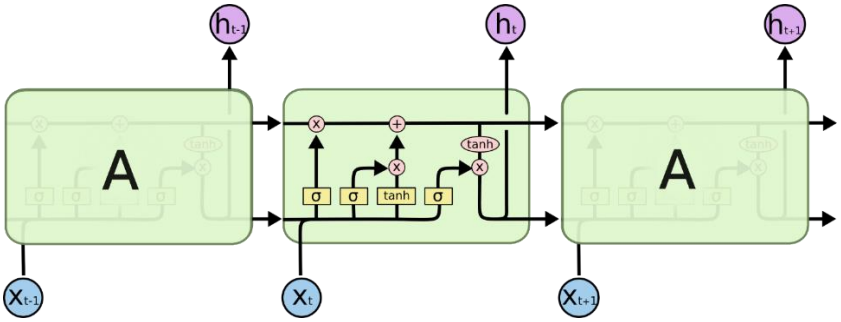
V.4. LSTM + Attention Layer

Ý tưởng:

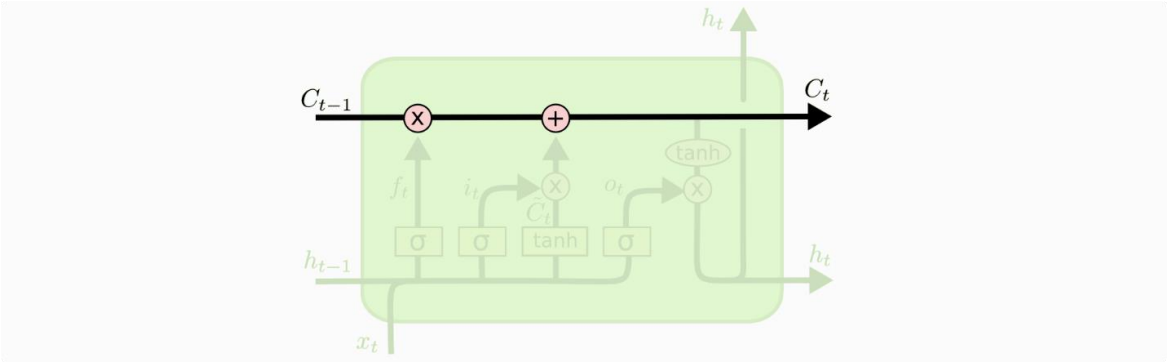
- Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.
- LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.
- Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng tanh *tanh*.



- LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

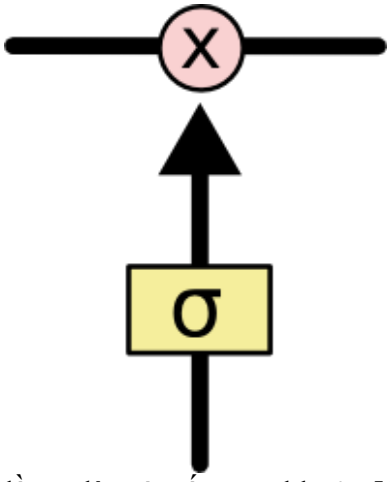


- Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính đường chạy thông ngang phía trên của sơ đồ hình vẽ.
- Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.



- LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho trạng thái tế báo, chúng được điều chỉnh cẩn thận bởi các nhóm được gọi là cổng (gate).

- Các cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân.



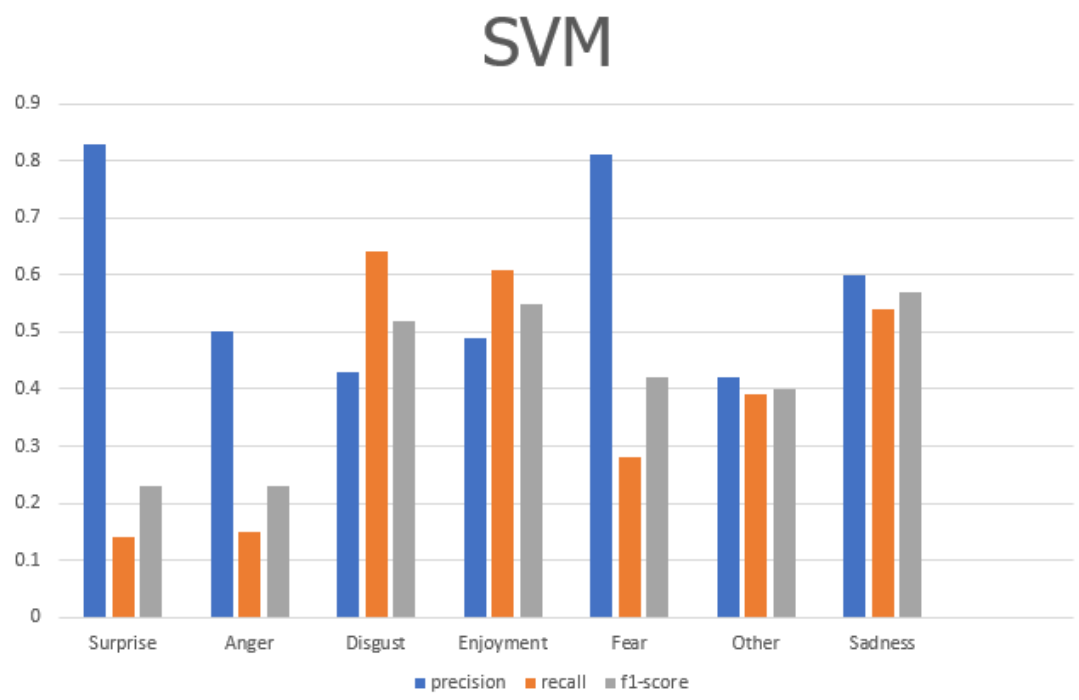
- Tầng sigmoid sẽ cho đầu ra là một số trong khoản [0, 1][0,1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 00 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 11 thì có nghĩa là cho tất cả các thông tin đi qua nó.
- Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.
- Ở bài toán này, sử dụng kết hợp LSTM + Attention Layer để ra độ chính xác cao hơn

VI. THÍ NGHIỆM VÀ KẾT QUẢ

VI.1. Thực nghiệm SVM

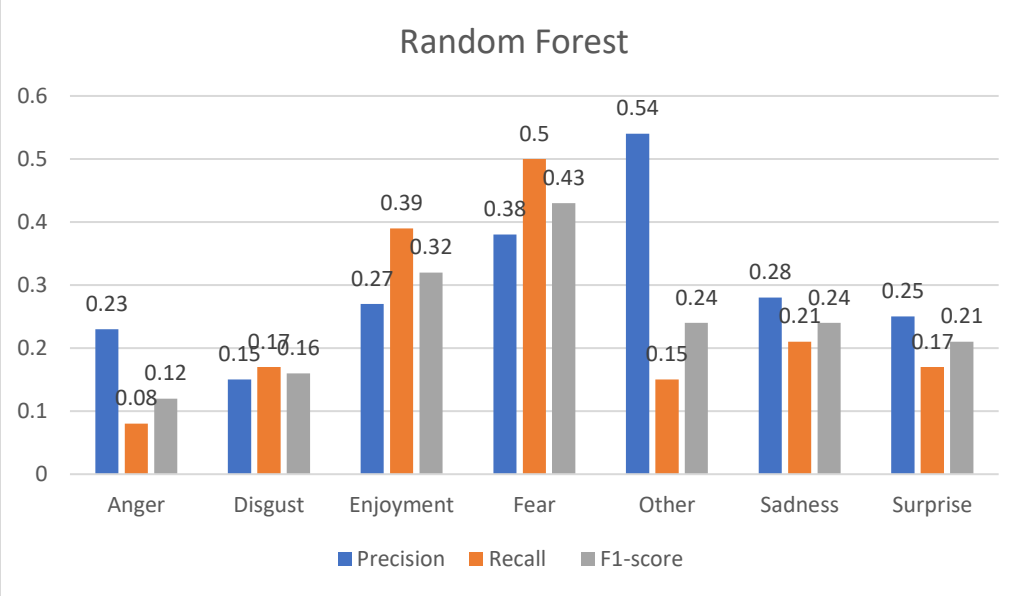
- Sử dụng Radial basis function (rbf): gamma =scale

	precision	recall	f1-score	support
Surprise	0.83	0.14	0.23	37
Anger	0.5	0.15	0.23	40
Disgust	0.43	0.64	0.52	132
Enjoyment	0.49	0.61	0.55	193
Fear	0.81	0.28	0.42	46
Other	0.42	0.39	0.40	129
Sadness	0.6	0.54	0.57	116
accuracy			0.49	693
micro avg	0.49	0.49	0.49	693
marco avg	0.58	0.39	0.42	693
weighted avg	0.53	0.49	0.47	693



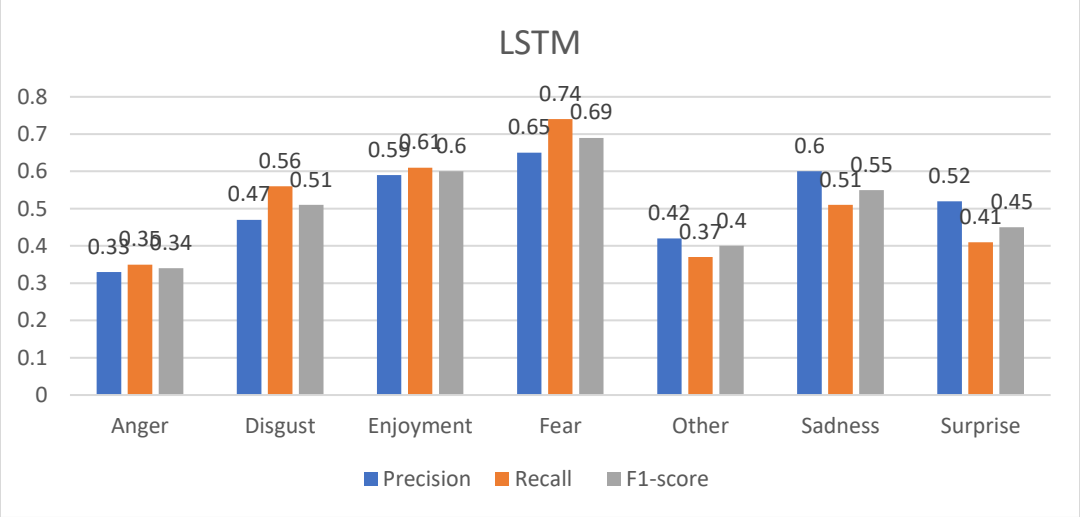
VI.2. Thực nghiệm Random Forest

	precision	recall	f1-score	support
Surprise	0.23	0.08	0.12	37
Anger	0.15	0.17	0.16	40
Disgust	0.27	0.39	0.32	132
Enjoyment	0.38	0.5	0.43	193
Fear	0.54	0.15	0.24	46
Other	0.28	0.21	0.24	129
Sadness	0.25	0.17	0.21	116
accuracy			0.3	693
marco avg	0.3	0.24	0.24	693
weighted avg	0.31	0.3	0.29	693



VI.3. Thực nghiệm LSTM

	precision	recall	f1-score	support
Surprise	0.33	0.35	0.34	40
Anger	0.47	0.56	0.51	132
Disgust	0.59	0.61	0.60	193
Enjoyment	0.65	0.74	0.69	46
Fear	0.42	0.37	0.40	129
Other	0.60	0.51	0.55	116
Sadness	0.52	0.41	0.45	37
accuracy			0.52	693
marco avg	0.51	0.51	0.51	693
weighted avg	0.52	0.52	0.52	693



VII. PHÂN TÍCH KẾT QUẢ THỰC NGHIỆM

- Trong các thuật toán SVM, Random Forest và LSTM thì LSTM cho các độ đo accuracy, f1-score cao nhất.
- Thuật toán deep learning ra độ chính xác cao hơn các thuật toán machine learning về bài toán phân loại cảm xúc mạng xã hội.

VIII. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

- Mô hình LSTM sử dụng phương pháp Attention Layer sẽ cho độ chính xác cao hơn do Attention giúp mô hình tập trung vào các từ chỉ cảm xúc.
- Các SVM, Random Forest ra độ chính xác không được cao bằng LSTM do dữ liệu mất cân bằng hoặc có thể do phương pháp chuyển vector chưa phù hợp
- Do trong quá trình thực hiện, tụi em chỉ chuyển vector bằng phương pháp Word2vec nên kết quả với phương pháp chưa tối ưu nhất
- Sử dụng các phương pháp preprocessing với các mô hình khác để tìm ra phương pháp thích hợp nhất cho bộ dữ liệu.
- Mô hình LSTM đạt được kết quả cao nhất trong các độ đo nên tụi em chọn LSTM.

IX. TÀI LIỆU THAM KHẢO

[1] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[2] A. Kirlić, Z. Orhan, A. Hasovic, and M. Kevser-Gokgol, "Stock market prediction using Twitter sentiment analysis," *Invent. J. Res. Technol. Eng. Manag.*, vol. 2, no. 1, pp. 1–4, 2018.

[3] M. R. Huq, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 19–25, 2017.

[4] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Comput. Sci.*, vol. 127, pp. 511–520, 2018.

[5] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, 2018.