

Fake News Detection

Võ Nhật Cường, Nguyễn Duy Mẫn, Huỳnh Minh Tuấn

¹ University of Information Technology, Ho Chi Minh City, Vietnam

² Vietnam National University, Ho Chi Minh City, Vietnam

Tóm tắt nội dung Tin tức giả đang lan truyền trên các nền tảng xã hội, báo đài...ngày càng rộng rãi. Đó là một vấn đề đáng lo ngại, vì các tác động tiêu cực của nó gây ra nhiều thiệt hại cho quốc gia và xã hội. Trên thế giới cũng đã có rất nhiều công trình nghiên cứu liên quan đến phát hiện các tin tức giả. Trong đồ án lần này, chúng em kết hợp với những kiến thức đã được học trong môn Học máy và sự tìm hiểu từ các nguồn tài liệu, tiến hành xây dựng 1 mô hình máy học có thể đưa ra kết quả phân loại tin tức thật, giả với các độ đo tốt nhất. Bằng cách sử dụng các công cụ như python scikit-learning, NLP để phân tích văn bản, thực hiện mã hóa và trích xuất các tính năng từ dữ liệu văn bản. Sau đó, sẽ thực hiện lựa chọn các Hyperparameter phù hợp với từng mô hình được sử dụng để đạt được độ chính xác cao nhất. Ở đây em sử dụng 2 độ đo cho các mô hình của mình accuracy và F1-macro. Kết quả tốt nhất đạt được là mô hình Passive Aggressive với trích xuất đặc trưng Tf-idf F1-macro : 0.95 và độ đo Accuracy: 0.94.

1 Giới thiệu

Mặc dù vấn đề tin tức giả mạo không phải là một vấn đề mới, nhưng việc phát hiện tin tức giả mạo là được cho là một nhiệm vụ phức tạp do con người có xu hướng tin vào những thông tin sai lệch và việc thiếu kiểm soát việc phát tán nội dung giả mạo. Tin tức giả đã được chú ý nhiều hơn trong vài năm gần đây, đặc biệt là kể từ cuộc bầu cử Hoa Kỳ năm 2016. Con người rất khó phát hiện ra tin tức giả. Có thể lập luận rằng cách duy nhất để một người có thể tự xác định tin tức giả là có một kiến thức rộng lớn về chủ đề được đề cập. Ngay cả khi có hiểu biết, rất khó để xác định thành công thông tin trong bài báo là thật hay giả. Bản chất mở của web và phương tiện truyền thông xã hội cùng với tiến bộ gần đây của khoa học máy tính giúp đơn giản hóa quá trình tạo và phát tán tin tức giả mạo. Việc phát hiện tin tức giả được cho là một nhiệm vụ phức tạp và khó hơn nhiều so với việc phát hiện các đánh giá sản phẩm giả do chúng lan truyền dễ dàng bằng cách sử dụng phương tiện truyền thông xã hội và truyền miệng.

Dựa trên những vấn đề trên, đã tạo động lực cho nhóm chúng em tiến hành tìm hiểu bài toán Fake news detection trên văn bản tiếng anh, để tích lũy cho mình những kinh nghiệm thu được trong quá trình xây dựng các thuật toán máy học. Để sau này có thể xây dựng các mô hình tương tự với độ chính xác cao cho bài toán Fake news detection trên văn bản tiếng Việt.

2 Các dự án liên quan

Fake news Detetion là một chủ đề phổ biến và nổi bật trong lĩnh vực xử lý ngôn ngữ tự nhiên trong máy học. Ở tiếng Anh thì đã có nhiều bài báo khoa học về chủ đề này như bài báo của Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”.Hoặc bài báo của "Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques.".

Trong bài báo của mình, Ahmed H và các cộng sự đã xây dựng N-gram model với ý tưởng là tạo ra các bộ cấu hình tần số n-gram khác nhau từ dữ liệu đào tạo để đại diện cho các bài báo giả mạo và trung thực. Họ đã sử dụng một số đặc trưng n-gram cơ bản dựa trên các từ và kiểm tra ảnh hưởng của độ dài n-gram đến độ chính xác của các thuật toán phân loại khác nhau. Ở đây họ điều tra và so sánh hai kỹ thuật trích xuất các tính năng khác nhau và sáu kỹ thuật phân loại máy học khác nhau. Đánh giá thử nghiệm mang lại hiệu suất tốt nhất bằng cách sử dụng Tần số tài liệu đảo ngược tần số kỳ hạn (TF-IDF) làm kỹ thuật trích xuất tính năng và Máy vectơ hỗ trợ tuyến tính (LSVM) làm bộ phân loại, với độ chính xác 0.92..

3 Phương Pháp được Đề Xuất

Trong đề án này nhóm chúng em sẽ trình bày 3 mô hình máy học Naive Bayes Classifier, Passive Aggressive Classifier, Logistic Regression với 2 phương pháp Feature Extractions : TF-IDF và Countvectorizer. Kết hợp với việc điều chỉnh các Hyperparameter để chọn ra các mô hình có độ chính xác cao nhất.

3.1 Data Pre-processing

Trước khi biểu diễn dữ liệu bằng cách sử dụng n-gram và mô hình dựa trên vectơ, dữ liệu cần phải tuân theo một số quy định nhất định như loại bỏ từ dừng, mã hóa, viết hoa thường, phân đoạn câu và loại bỏ dấu câu. Điều này sẽ giúp chúng tôi giảm kích thước của dữ liệu thực tế bằng cách loại bỏ thông tin không liên quan tồn tại trong dữ liệu.

Chúng em tạo một hàm để tiến hành xử lý các kí tự viết hoa, xóa bỏ các kí tự đặc biệt, khoảng trắng, để loại bỏ những kí tự nhiễu gây ảnh hưởng đến kết quả dự đoán của mô hình.

Stop Word Removal

Stop words là các từ không có nghĩa trong một ngôn ngữ hoàn toàn không có ý nghĩa khi được sử dụng làm các tính năng trong phân loại văn bản. Đây là những từ thường được sử dụng rất nhiều trong câu để giúp kết nối tư tưởng hoặc hỗ trợ trong cấu trúc câu. Các mạo từ, giới từ và liên từ và một số đại từ được coi là từ dừng. Chúng em đã xóa các từ phổ biến như, a, about, an, is, as, at, be, by, for, from, how, in, is, of, on, or, that, the, these, this, too, was , cái gì, khi

nào, ở đâu, ai, sẽ, v.v. Những từ đó đã được xóa khỏi mỗi tài liệu, và các tài liệu đã xử lý được lưu trữ và chuyển sang bước tiếp theo.

Lemmatization

Sau khi mã hóa dữ liệu, bước tiếp theo là chuyển đổi mã thông báo thành dạng chuẩn. Quá trình lemmatization trong xử lý ngôn ngữ tự nhiên liên quan đến việc làm việc với các từ theo các thành phần từ vựng gốc của chúng và giảm số lượng các loại hoặc lớp từ trong dữ liệu. Ví dụ: các từ “Running”, “Ran” và “Runner” sẽ được rút gọn thành từ “run”.

3.2 Features Extraction

Một trong những thách thức của phân loại văn bản là Khi số lượng đối tượng trong tập dữ liệu vượt quá số lượng quan sát, chúng ta sẽ không bao giờ có câu trả lời xác định cho vấn đề đang muốn quan sát. Có một số lượng lớn các thuật ngữ, từ và cụm từ trong tài liệu dẫn đến gánh nặng tính toán cao cho quá trình học tập. Hơn nữa, các tính năng không liên quan và dư thừa có thể ảnh hưởng đến độ chính xác và hiệu suất của bộ phân loại. Do đó, cách tốt nhất là thực hiện giảm đối tượng để giảm kích thước đối tượng văn bản và tránh kích thước không gian đối tượng lớn. Trong đề án này, chúng em đã nghiên cứu hai phương pháp lựa chọn đối tượng khác nhau, đó là Term Frequency-Inverted Document Frequency (TF-IDF) và Bag of Words bằng hàm countvectorizer. Các phương pháp này được mô tả như sau.

Countvectorizer

CountVectorizer tokenizes (tokenization có nghĩa là chia nhỏ một câu hoặc đoạn văn hoặc bất kỳ văn bản nào thành các từ) cùng với việc thực hiện xử lý trước rất cơ bản như loại bỏ các dấu câu, chuyển tất cả các từ thành chữ thường, v.v. để mã hóa văn bản không nhìn thấy sau này. Một vectơ được mã hóa được trả về với độ dài của toàn bộ từ vựng và số nguyên cho số lần mỗi từ xuất hiện trong tài liệu.

TF-IDF

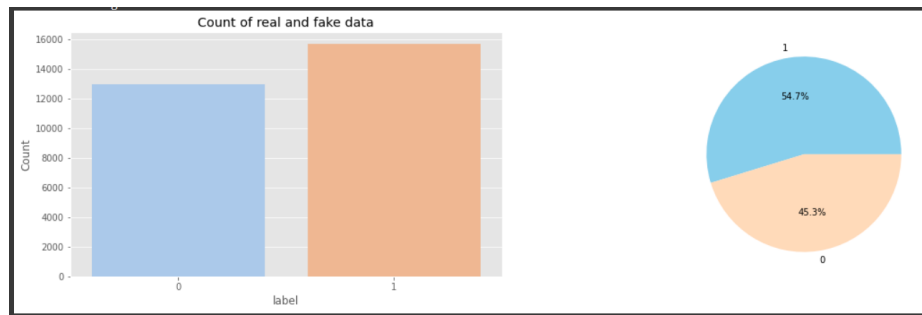
Term Frequency-Inverted Document Frequency (TF-IDF) là một thước đo trọng số thường được sử dụng trong truy xuất thông tin và xử lý ngôn ngữ tự nhiên. Nó là một số liệu thống kê được sử dụng để đo lường mức độ quan trọng của một thuật ngữ đối với một tài liệu trong tập dữ liệu. Tầm quan trọng của thuật ngữ tăng lên theo số lần một từ xuất hiện trong tài liệu, tuy nhiên, điều này bị ảnh hưởng bởi tần suất xuất hiện của từ trong ngữ liệu.

4 Quy trình thực hiện

4.1 Bộ dữ liệu

Trong mô hình máy học này nhóm sẽ sử dụng tập 2 tập dữ liệu: "Getting Real about Fake News" và "Gathering real news for Oct-Dec 2016" có sẵn từ Kaggle. Tập dữ liệu "Getting Real about Fake News" chứa văn bản và siêu dữ liệu từ 244 trang web và đại diện cho tổng số 12.999 bài đăng và 20 cột dữ liệu. Tập dữ liệu "Gathering real news for Oct-Dec 2016" chứa 15712 bài đăng gồm 11 cột dữ liệu. Bộ dữ liệu cuối cùng được sử dụng cho các mô hình trong báo cáo này là sự kết hợp của 2 bộ dữ liệu kể trên, sau khi đã tiến hành loại bỏ các cột thừa và hợp nhất các cột có cùng trường dữ liệu. Đối với mỗi bài báo đều có các thông tin sau:

- Article Text
- Article Title
- Article label (0 or 1)



Hình 1: Tỷ lệ các nhãn trong tập dữ liệu.

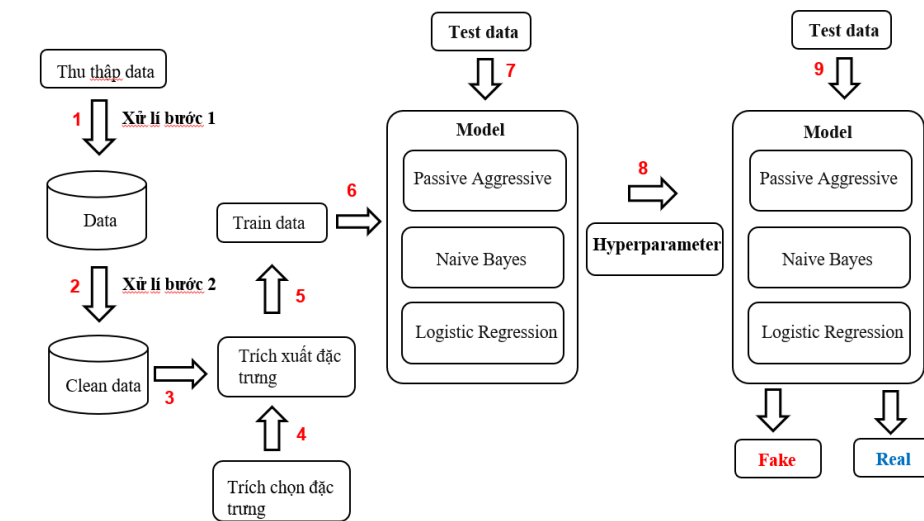
	title		text	label
0	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...		0
1	Re: Why Did Attorney General Loretta Lynch Ple...	Why Did Attorney General Loretta Lynch Plead T...		0
2	BREAKING: Weiner Cooperating With FBI On Hilla...	Red State : \nFox News Sunday reported this mo...		0
3	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	Email Kayla Mueller was a prisoner and torture...		0
4	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...		0
...
15707	An eavesdropping Uber driver saved his 16-year...	Uber driver Keith Avila picked up a p...		1
15708	Plane carrying six people returning from a Cav...	Crews on Friday continued to search L...		1
15709	After helping a fraction of homeowners expecte...	When the Obama administration announced a...		1
15710	Yes, this is real: Michigan just banned bannin...	This story has been updated. A new law in...		1
15711	What happened in Washington state after voters...	The nation's first recreational marijuana...		1

28711 rows x 3 columns

Hình 2: Sơ lược bộ dữ liệu.

Nhận xét:

- Đây là một tập dữ liệu thô chưa được xử lý còn chứa nhiều dữ liệu gây nhiễu.
- Bộ dữ liệu có 28711 hàng và 3 cột với tỉ lệ các nhãn fake(0)/real(1) lần lượt là 54,7 và 45,3.
- Tỉ lệ của các nhãn trong tập dữ liệu khi chưa xử lý đang ở mức cân bằng.

4.2 Giải pháp

Hình 3: Tổng quan quy trình và giải pháp .

Hình 3 mô tả tổng quát quy trình và giải pháp cho mô hình được sử dụng trong đề án này. Dựa vào sơ đồ ta có thể chia giải pháp thành 3 bước chính như sau

Bước 1:

- Thu thập data và tiến hành tiền xử lý dữ liệu. Các bước tiền xử lý dữ liệu sẽ được mô tả kĩ ở phần sau

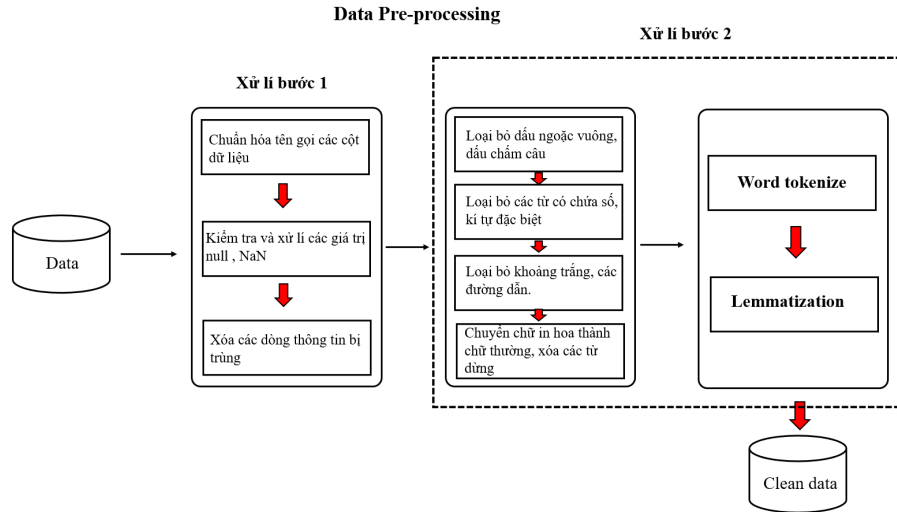
Bước 2:

- Tiến hành trích chọn và trích xuất các đặc trưng phù hợp cho mô hình

Bước 3:

- Tiến hành train, tinh chỉnh các siêu tham số của các mô hình dựa trên tập dữ liệu train data, và được kiểm chứng trên tập test data được chia từ bộ dữ liệu với tỉ lệ 80/20.

4.3 Tiền xử lí dữ liệu



Hình 4: Quy trình tiền xử lí dữ liệu.

Ở đây chúng em đi xây dựng quy trình tiền xử lí dữ liệu với 2 bước chính.

Xử lí bước 1:

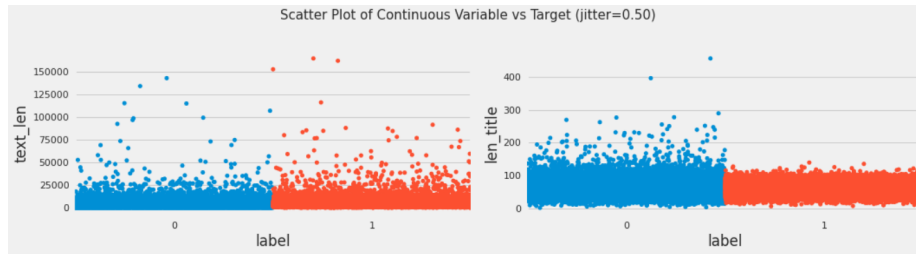
- Vì bộ dữ liệu được kết hợp từ 2 nguồn khác nhau nên sẽ tồn tại các tên gọi không đồng nhất giữa các cột dữ liệu với nhau, để thuận tiện trong quá trình xử lí tiếp theo, ở đây em tiến hành chuẩn hóa tên gọi các cột dữ liệu
- Tiếp theo là sẽ xử lí các dữ liệu mang giá trị Null và NaN trong tập dữ liệu
- Bởi vì đây là các tin tức được thu thập 1 cách chủ động bằng các API nên việc có các dòng dữ liệu bị trùng là có khả năng, để giảm độ phức tạp cho mô hình thì cần loại bỏ các hàng dữ liệu bị trùng

Xử lí bước 2:

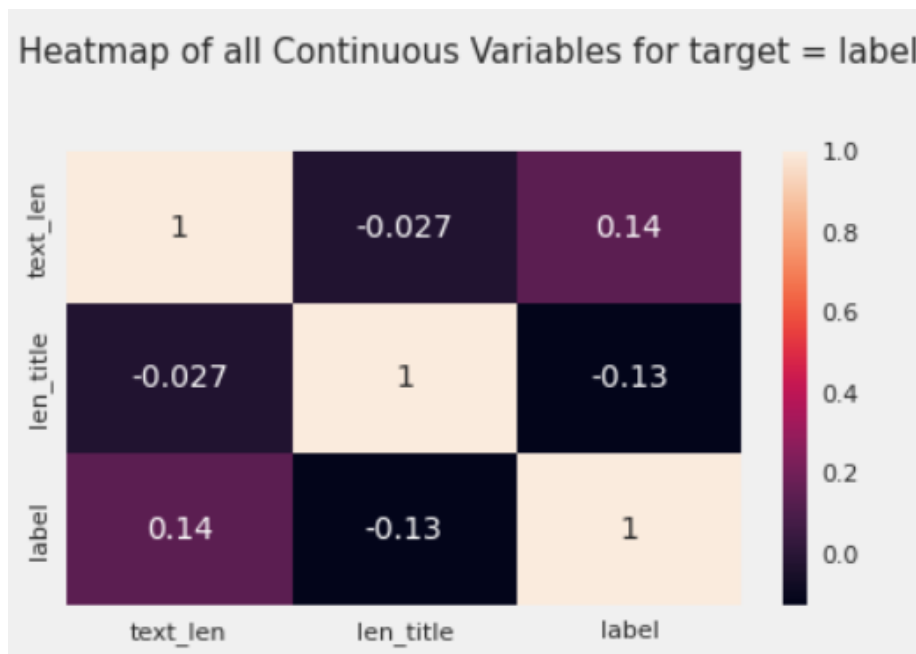
- Vì dữ liệu này được trích xuất từ các bài báo khác nhau nên có tồn tại nhiều dữ liệu nhiễu gây ảnh hưởng đến quá trình xây dựng các mô hình máy học, vì vậy ta cần tiến hành loại bỏ các dữ liệu nhiễu đó như: kí tự đặc biệt, chuyển chữ thường thành chữ hoa, loại bỏ các từ dừng.v.v
- Bước tiếp theo là tiến hành tokenize các đoạn văn bản và lenmatization để chuẩn hóa các từ thành dạng gốc của nó

4.4 Trích chọn đặc trưng phù hợp với mô hình

Vì trong tập dữ liệu này có 2 cột dữ liệu là title và text, để chọn được cột thuộc tính phù hợp cho các model chúng em tiến hành phân tích mối tương quan giữa chiều dài của các văn bản trong title, text và các nhãn.



Hình 5: Biểu đồ phân tán thể hiện mối tương quan giữa len-title, len-text và label.



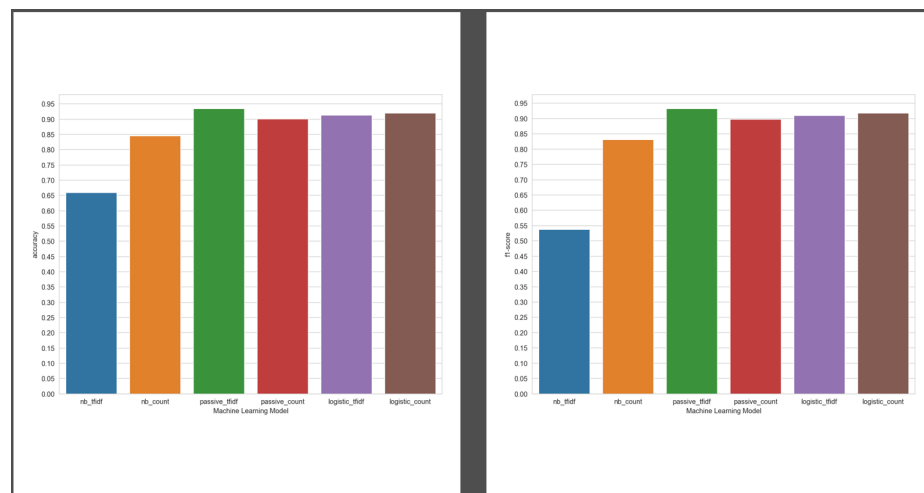
Hình 6: Heatmap của tất các các biến liên tục.

Dựa vào kết quả từ hình ảnh trên ta có thể thấy được thuộc tính len-text có mối tương quan với các nhãn lớn hơn len-title. Và ta có thể chọn đặc trưng text cho các mô hình máy học.

4.5 Kết quả

	Tf-idf vector		Count vectors	
	F1	Accuracy	F1	Accuracy
Naive Bayes	0.54	0.54	0.83	0.85
Passive Aggressive	0.93	0.93	0.90	0.90
Logistic Regression	0.91	0.91	0.92	0.92

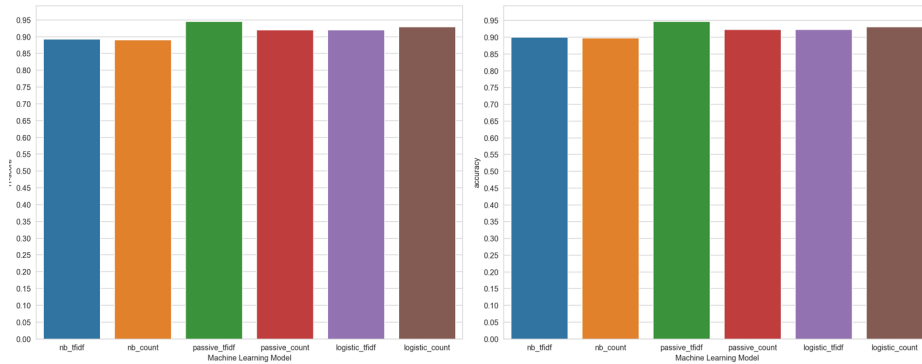
Hình 7: Kết quả của các mô hình ở chế độ mặc định.



Hình 8: Sơ đồ hóa kết quả của các mô hình ở chế độ mặc định.

	Tf-idf vector		Count vectors	
	F1	Accuracy	F1	Accuracy
Naive Bayes	0.89	0.89	0.90	0.90
Passive Aggressive	0.95	0.94	0.92	0.92
Logistic Regression	0.92	0.92	0.93	0.93

Hình 9: Kết quả của các mô hình sau khi hiệu chỉnh các siêu tham số.



Hình 10: Sơ đồ hóa kết quả của các mô hình sau khi hiệu chỉnh các siêu tham số.

Quy trình

- Đầu tiên chúng em tiến hành train các thuật toán Naive Bayes, Passive Aggressive, Logistic Regression với tập dữ liệu train-data với các siêu tham số đều ở chế độ mặc định.
- Sau đó sẽ tiến hành dự đoán các mô hình với tập test-data để đưa ra 2 độ đo F1-macro và accuracy để đánh giá mô hình. Kết quả đạt được như hình 7, 8.
- Tiếp theo sẽ tiến hành điều chỉnh các siêu tham số của từng mô hình bằng Grid Search model để chọn ra các siêu tham số tối ưu nhất cho


ra độ đo F1-macro và accuracy cao nhất cho từng mô hình. Kết quả model Passive Aggressive với tính năng TF-IDF đạt kết quả tốt nhất với F1-macro: 0.95 và accuracy: 0.94.

4.6 Đánh giá kết quả

Table 2. LSVM Accuracy results. The second row corresponds to the features size. Accuracy values are in %.

N-Gram Size	TF-IDF				TF			
	1000	5000	10,000	50,000	1000	5000	10,000	50,000
Uni-gram	89.0	89.0	89.0	92.0	87.0	87.0	87.0	87.0
Bi-gram	87.0	87.0	88.0	89.0	86.0	83.0	82.0	82.0
Tri-gram	84.0	85.0	86.0	87.0	86.0	84.0	84.0	79.0
Four-gram	71.0	76.0	76.0	81.0	70.0	70.0	70.0	61.0

	TF-IDF vector		Count vectors	
	F1	Accuracy	F1	Accuracy
Naive Bayes	0.89	0.89	0.90	0.90
Passive Aggressive	0.95	0.94	0.92	0.92
Logistic Regression	0.92	0.92	0.93	0.93



Passive Aggressive_Tf-idf có F1_macro : 95%
 Passive Aggressive_Tf-idf có Accuracy: 94%

Hình 11: So sánh kết quả.

Kết quả tốt nhất của chúng em đạt được với model Passive Aggressive và kỹ thuật trích xuất tính năng TF-IDF cho các độ đo F1-macro và accuracy lần lượt là 0.95 ,0.94 đã đạt được độ chính xác cao hơn khi so sánh với kết quả tốt nhất trong model N-gram của tác giả Ahmed H với mô hình máy học LSVM và tính năng trích xuất đặc trưng TF-IDF có độ chính xác accuracy là 0.92 .

4.7 Đánh giá mô hình

1. Một số mô hình chưa đạt được độ chính xác như mong muốn, nguyên nhân là do quá trình tinh chỉnh các siêu tham số của các mô hình, chỉ mới dừng lại ở một số siêu tham số cơ bản, chưa khai thác hết các siêu tham số.
2. Mặc dù đã cải thiện được độ đo chính xác của bài toán nhưng chưa kiểm thử lại trong thực tế với các bộ dữ liệu khác nhau nhằm đánh giá chính xác mức độ khả dụng của các mô hình.

5 Kết luận

Trong bài báo cáo này chúng em đã có thể vận dụng được những kiến thức về máy học để có thể tự xây dựng cho bản thân một mô hình máy học với đầy đủ các bước, nhằm giải quyết bài toán fake news detection .

Trong đồ án này, mục tiêu của nhóm đề ra là sử dụng các mô hình máy học truyền thống bằng cách tinh chỉnh các siêu tham số để đạt được tỉ lệ các độ đo như F1-macro và accuracy cao nhất .Kết quả model Passive Aggressive với tính năng TF-IDF đạt kết quả tốt nhất với F1-macro: 0.95 và accuracy: 0.94.

6 Hướng phát triển trong tương lai

1. Xây dựng thêm các mô hình máy học để cải thiện độ chính xác cho bài toán fake news detection trên văn bản tiếng anh làm tiền đề xử lí cho bài toán trên văn bản tiếng Việt.
2. Dựa vào những kinh nghiệm trong bài toán Fake news detection trên văn bản tiếng anh có thể xây dựng 1 mô hình tương tự xử lí bài toán bằng ngôn ngữ tiếng Việt.

7 Danh mục tài liệu tham khảo

1. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification
2. Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138)