TRƯỜNG ĐẠI HỌC QUY NHƠN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP LỚN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

ĐỀ TÀI:

PHÁT HIỆN TIN TỨC GIẢ MẠO

GIẢNG VIÊN: TS. LÊ QUANG HÙNG
SINH VIÊN THỰC HIỆN: LÊ THỊ NHƯ TRÂM
LỚP: CNTT K41C

Quy Nhơn, 11/2021

MỤC LỤC

1. Mở đầu

Với sự phát triển của công nghệ thông tin, mạng Internet đã lan rộng và phủ sóng toàn cầu. Bên cạnh những lợi ích to lớn mà mạng xã hội mang lại, chúng ta đang đối mặt với nhiều nguy cơ, thách thức không nhỏ, thậm chí đe dọa đến an ninh quốc gia và trật tự an toàn xã hội. Trong đó, phải kể đến những ảnh hưởng tiêu cực từ các thông tin xấu, độc được lan truyền trên mạng xã hội cũng như vấn nạn tin giả - Fake News. Hệ lụy của việc lan truyền "tin giả" không chỉ dừng lại ở những cá nhân đơn lẻ, những nhóm người ở từng địa phương nhất định mà còn có tác động rộng lớn hơn, đe dọa trực tiếp tới an ninh quốc gia.

Tin giả lan tràn như con virus, dịch bệnh gây ra rất nhiều tổn thất không những với cá nhân mà với cả các tổ chức kinh tế. Như việc tung tin thất thiệt về dịch tả lợn châu Phi của một tài khoản Facebook khiến dư luận hoang mang, tẩy chay thịt lợn, ảnh hưởng tới chăn nuôi của người nông dân, khiến nhiều người hoảng loạn, mất phương hướng trong cuộc sống, hay mất lòng tin vào những chỉ đạo của các cơ quan quản lý nhà nước. Ngoài những thiệt hại về kinh tế, một trong những hậu quả nghiêm trọng nhất mà tin giả gây ra là làm suy giảm niềm tin của công chúng vào truyền thông nói chung và báo chí chủ lưu nói riêng. Tin giả khiến công chúng không xác định được đâu là những nguồn tin đáng tin cậy để tiếp nhận, luôn ở trạng thái ngờ vực, tham khảo cả những nguồn tin không chính thống dẫn đến bị nhiễu loạn thông tin.

2. Giới thiệu chung

<u>Tổng quan:</u>

Vài năm trở lại đây, đặc biệt là từ cuối năm 2016 đến nay, thế giới đề cập nhiều về tin tức giả, từ định nghĩa, hiện trạng tin tức giả đến các biện pháp ứng phó. Tuy nhiên, đến nay vẫn chưa có định nghĩa rõ ràng, thống nhất về tin tức giả. Theo định nghĩa của từ điển Collins, tin tức giả là "những thông tin sai, thường là giật gân, được phát tán dưới vỏ bọc tin tức. Trong khi đó một số hãng tin tức định nghĩa, tin tức giả là những tin tức hoặc câu chuyện trên internet không đúng sự thật.

Định nghĩa của từ điển Collins sát nhất với nghĩa của từ "fake news" hiện đang được đề cập nhiều trên truyền thông trong khi định nghĩa còn lại bao hàm rộng hơn, ngoài những thông tin sai dưới vỏ bọc tin tức còn có những thông tin, câu chuyện không đúng sự thật được lan truyền trên internet. Theo các định nghĩa kể trên, có thể phân loại tin tức giả thành hai loại.

- Loại thứ nhất là những thông tin hoàn toàn không chính xác (bao gồm cả những thông tin thông thường và những thông tin được trình bày giống như một tin báo chí) được cố tình đăng tải, lan truyền vì một mục đích nào đó.

- Loại thứ hai là những thông tin có thể có một phần sự thật nhưng không hoàn toàn chính xác do người viết chúng không kiểm chứng toàn bộ sự thật trước khi đăng tải chia sẻ hoặc có thể họ phóng đại một phần của câu chuyện đỏ. Trên thế giới và Việt Nam đều có những trường hợp xảy ra đối với hai loại tin tức giả này.

Ngoài tạo ra các tin tức sai sự thật, còn một hình thức giả mạo khác được các đối tượng sử dụng đó là mạo danh các tổ chức, công ty, các tờ báo lớn, chính thống hay các cá nhân là người nổi tiếng là lãnh đạo, nguyên thủ quốc gia để đưa tin theo chủ đích của chúng. Nhiều lãnh đạo Đảng, Nhà nước tại Việt Nam cũng bị mạo danh đặt tên là các trang tin tổng hợp hay các tài khoản, fanpage trên mạng xã hội.

Có thể nói dù cố ý hay vô ý các tin tức giả đã, đang và sẽ xuất hiện trên các phương tiện truyền thông từ báo chí chủ lưu đến các loại hình truyền thông xã hội. Những câu chuyện giả mạo được chia sẻ rộng rãi trên truyền thông xã hội và sau đó được Google và các công cụ tìm kiếm khác xếp hạng cao giúp chúng được tìm thấy dễ dàng hơn và gia tăng cảm giác tin tưởng của người đọc đối với chúng. Hiện Google và Facebook nằm trong số những nền tảng phân phối tin tức giả lớn nhất.

<u>Khó khăn trong ngăn chặn tin tức giả mạo:</u>

- **Lợi ích lớn và sự đa dạng của các tác nhân tham gia**.

Do mức độ lan truyền của tin giả trên mạng xã hội trong một thời gian ngắn, tin giả tạo ra mối đe dọa đối với các nguồn thông tin truyền thống, chẳng hạn như báo chí truyền thống. Việc lan truyền tin tức giả mạo xảy ra như một sự kiện được phân phối và liên quan đến nhiều thực thể và nền tảng công nghệ. Do đó, ngày càng có nhiều khó khăn trong việc nghiên cứu và thiết kế các chiến lược tính toán, công nghệ và kinh doanh để chống lại tin tức giả mạo mà không ảnh hưởng đến tốc độ và khả năng truy cập cộng tác vào thông tin chất lượng cao.

- **Ý đồ xấu của đối thủ**.

Nội dung tin tức giả được thiết kế để khiến con người khó xác định tin tức giả mạo, khai thác kỹ năng nhận thức, cảm xúc và định kiến tư tưởng của chúng ta. Hơn nữa, các phương pháp tính toán để phát hiện tin tức giả là một thách thức, vì cách thức trình bày tin tức giả tương tự như tin tức thật và đôi khi tin tức giả sử dụng các chứng cứ để gây khó khăn cho việc xác định nguồn tin hoặc làm sai lệch nguồn tin tức thực sự.

- **Tính nhạy cảm và thiếu nhận thức của cộng đồng**

Người sử dụng mạng xã hội phải hứng chịu một lượng lớn thông tin có nguồn gốc không rõ ràng, từ thông tin có tính chất hài hước, như châm biếm, đến thông tin nhằm đánh lừa người tiêu dùng về thông tin được coi là tin tức hợp pháp. Tuy nhiên, người sử dụng

mạng xã hội không thể phân biệt tin tức giả mạo với tin tức hợp pháp chỉ bằng nội dung. Người dùng không có thông tin về độ tin cậy của nguồn hoặc các hình thức lan truyền tin tức trên mạng. Do đó, để nâng cao nhận thức của công chúng, một số bài báo và chiến dịch quảng cáo được chạy để cung cấp các mẹo về cách phân biệt giữa tin tức sai và tin tức hợp pháp.

- **Động lực lan truyền**

Việc lan truyền tin tức giả trên mạng xã hội làm phức tạp việc phát hiện và giảm thiểu, vì thông tin giả có thể dễ dàng tiếp cận và ảnh hưởng đến số lượng lớn người dùng trong thời gian ngắn. Thông tin được truyền đi một cách nhanh chóng và dễ dàng, ngay cả khi tính xác thực của nó còn bị nghi ngờ. Việc xác minh tính xác thực phải được thực hiện một cách nhanh chóng, nhưng nó cũng phải xem xét các mô hình lan truyền thông tin trong toàn mạng.

- **Thay đổi liên tục trong các đặc điểm của tin tức giả mạo**

Sự phát triển trong việc xác định tự động tin tức giả mạo cũng thúc đẩy sự thích nghi của việc tạo ra nội dung thông tin sai lệch mới để tránh bị phân loại như vậy. Việc phát hiện tin tức giả dựa trên phong cách viết, phân biệt tin tức giả và tin tức hợp pháp bằng phân tích dựa trên Xử lý ngôn ngữ tự nhiên, là một trong những giải pháp thay thế được sử dụng nhiều nhất do những thách thức chưa được giải quyết trong việc xác minh thực tế tự động từ các cơ sở kiến thức được xác định trước. Do đó, các phương pháp tiếp cận hiện tại để xác định tin tức giả dựa trên nội dung tập trung vào việc trích xuất sự kiện trực tiếp từ nội dung tin tức và xác minh sự thật sau đó dựa trên cơ sở tri thức.

- **Tấn công vào việc học ngôn ngữ tự nhiên**

Ta có thể xác định được ba cuộc tấn công: sự bóp méo sự thật, sự trao đổi giữa chủ thể và khách thể, và sự nhầm lẫn về nguyên nhân. Sự bóp méo, trên thực tế, là để phóng đại hoặc sửa đổi một số từ. Các yếu tố văn bản, chẳng hạn như ký tự và thời gian, có thể bị bóp méo để dẫn đến giải thích sai. Sự trao đổi giữa chủ thể và khách thể nhằm mục đích khiến người đọc nhầm lẫn giữa những người thực hành và những người phải chịu đựng hành động được báo cáo. Sự tấn công của sự nhầm lẫn nguyên nhân bao gồm việc tạo ra các mối quan hệ nhân quả không tồn tại giữa hai sự kiện độc lập hoặc cắt các phần của một câu chuyện, chỉ để lại những phần mà kẻ tấn công muốn trình bày cho người đọc.

Các cơ hội nghiên cứu để xác định và giảm thiểu tin tức giả tập trung vào việc phát hiện nguồn tin nhanh chóng hoặc theo thời gian thực, kiểm soát sự lan truyền của thông tin sai lệch và giảm tác động của tin tức giả đối với xã hội. Tập dữ liệu được thu thập trong thời gian thực, tự động phát hiện tin đồn và vị trí của nguồn là những câu hỏi nghiên cứu đầy thách thức.

3. Định nghĩa bài toán

Phát hiện tin giả dựa trên nội dung

Dữ liệu đầu vào là văn bản, mẫu tin ngắn đến toàn bộ bài báo (gọi chung là bài báo). Gọi $A = \{a_1, a_2,\ldots,a_n\}$ là tập hợp gồm n bài báo. Giả sử bài toán cần xác minh a có thể được biểu diễn dưới dạng véc-tơ đặc trưng $v \in R^k$ (k là số chiều của dữ liệu). Nhiệm vụ xác minh bài báo a dựa trên nội dung là xác định hàm f, sao cho:

$$f : v \xrightarrow{D} \hat{y}$$

Trong đó $\hat{y} \in \{0$ (tin đúng), 1(tin giả)$\}$ là nhãn của bài báo được dự đoán và

$D = \{(v_i, y_i)|v_i \in R^k, y_i \in \{0, 1\}, i = 1\ldots m\}$ là dữ liệu huấn luyện. Dữ liệu huấn luyện D (bao gồm m bài báo, mỗi bài báo $a_i \in D$ được biểu diễn bởi véc-tơ đặc trưng v_i với nhãn y_i) giúp ước lượng các tham số trong hàm f.

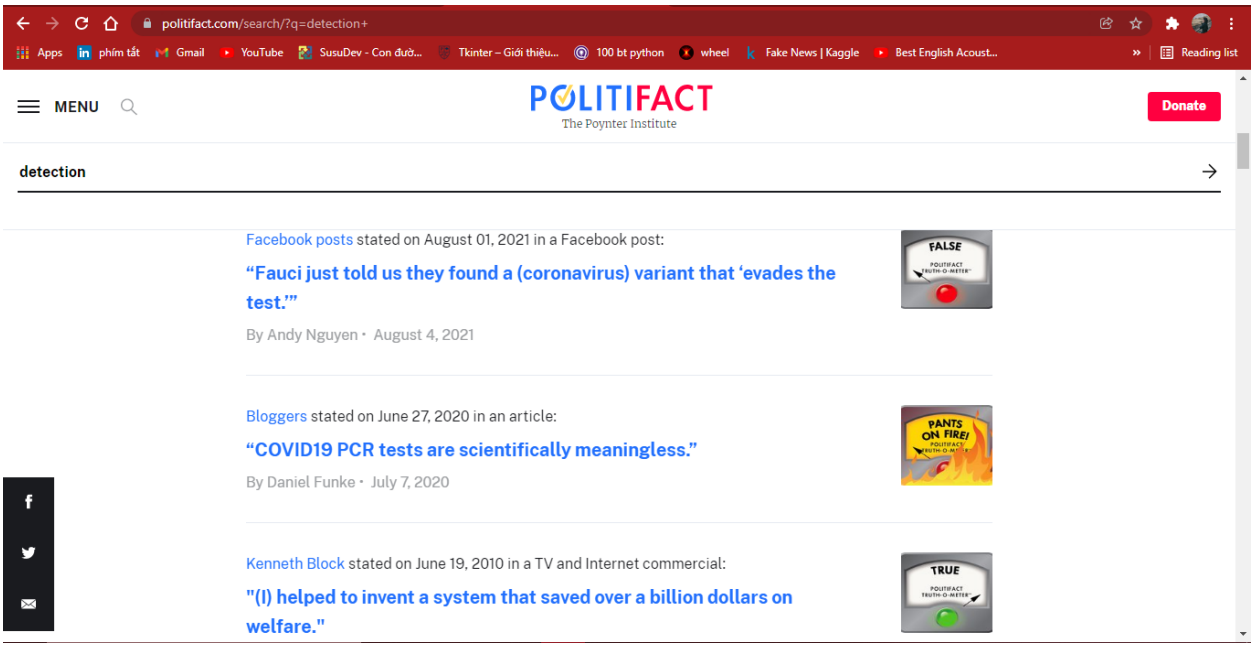
Phát hiện tin tức giả dựa trên URL

Xem xét một tập hợp T các URL, $\{(u_1 , y_1), ..., (u_T , y_T)\}$, trong đó u_t với $t = 1, ...,T$ đại diện cho một URL và $y_t \in \{0, 1\}$ biểu thị nhãn của URL, với $y_t = 1$ tương ứng "bad" là URL độc hại và $y_t = 0$ tương ứng "good" là một URL không độc hại. Bước đầu tiên trong quy trình phân loại là biểu diễn véc tơ đặc trưng $u_t \rightarrow x_t$ trong đó $x_t \in R^n$ là véc tơ đặc trưng n chiều đại diện cho URL u_t . Bước tiếp theo là học một hàm dự đoán $f: R^n \rightarrow R$ là điểm dự đoán việc gán lớp cho một cá thể URL x.

4. Một số hệ thống phát hiện tin tức giả mạo

Politifact.com

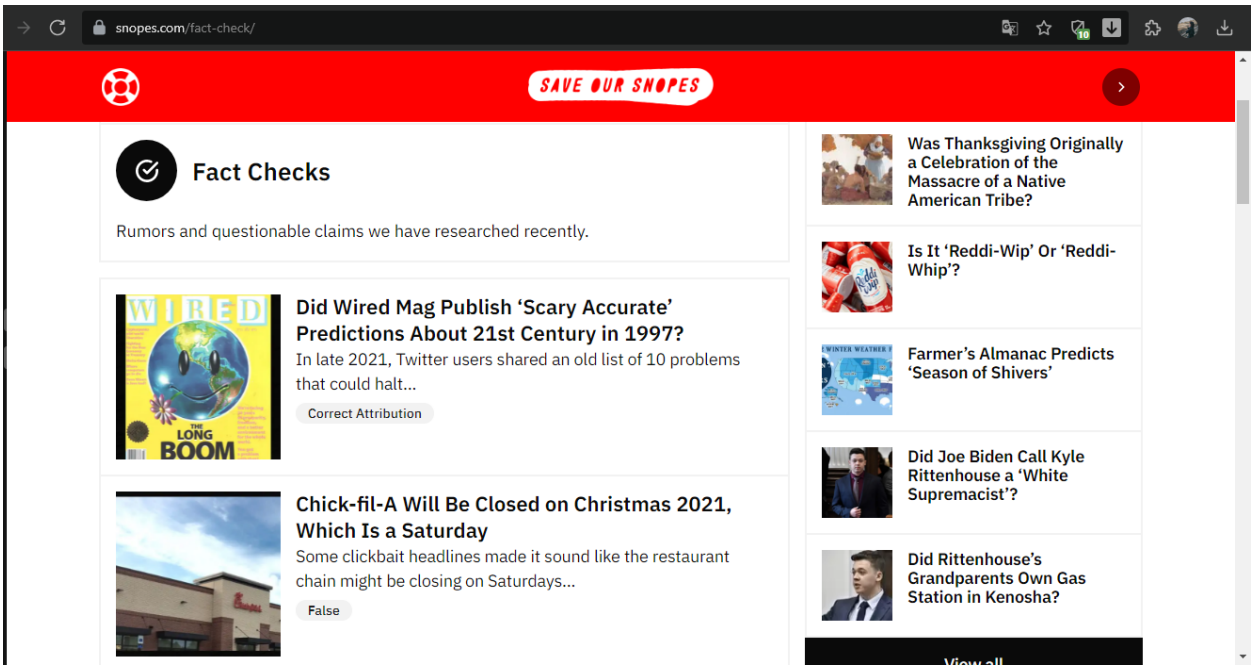
PolitiFact.com là một dự án phi lợi nhuận của Mỹ được điều hành bởi Viện Poynter ở Washington, DC. Nó bắt đầu vào năm 2007 như một dự án của Tampa Bay Times (sau đó là St. Petersburg Times), với các phóng viên và biên tập viên của tờ báo và các đối tác truyền thông liên kết của nó báo cáo về tính chính xác của các tuyên bố của các quan chức và những người khác có liên quan đến chính trị Hoa Kỳ. Các nhà báo đánh giá các tuyên bố ban đầu và xuất bản phát hiện của họ trên trang web PolitiFact.com, nơi mỗi tuyên bố nhận được xếp hạng "Truth-O-Meter". Xếp hạng từ "True" cho các tuyên bố mà các nhà báo cho là chính xác đến "Pants on fire" cho các tuyên bố mà các nhà báo cho là sai hoặc lố bịch và "False" cho các tuyên bố sai sự thật.



Snope.com

Snopes, trước đây gọi là trang tham khảo Urban Legends , là một thực tế kiểm tra trang web. Nó đã được mô tả là một "tài liệu tham khảo được đánh giá cao để phân loại các tin đồn" trên Internet. Trang web cũng được coi là một nguồn để xác thực và bóc tách các truyền thuyết đô thị cũng như những câu chuyện tương tự trong văn hóa đại chúng Hoa Kỳ.

Snopes là một bách khoa toàn thư trực tuyến chủ yếu trình bày kết quả tìm kiếm về các cuộc thảo luận của người dùng. Trang web đã phát triển để bao gồm nhiều đối tượng và trở thành một nguồn tài nguyên mà người dùng Internet bắt đầu gửi những bức ảnh và câu chuyện đáng nghi vấn về tính xác thực.



5. Cách tiếp cận

Xử lý ngôn ngữ tự nhiên:

- Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.
- Trong bài báo cáo này kỹ thuật làm sạch và định hình dữ liệu được áp dụng bao gồm mã hóa, loại bỏ dấu chấm câu và các ký tự đặc biệt, loại bỏ các từ dừng.
- Công nghệ sử dụng:
    - Mô đun re: Mô-đun này cung cấp các phép toán so khớp biểu thức chính quy.
    - Bộ công cụ nltk: là một thư viện Python mã nguồn mở để Xử lý Ngôn ngữ Tự nhiên. Cụ thể là sử dụng lớp porterstemmer và stopword.

Học máy:

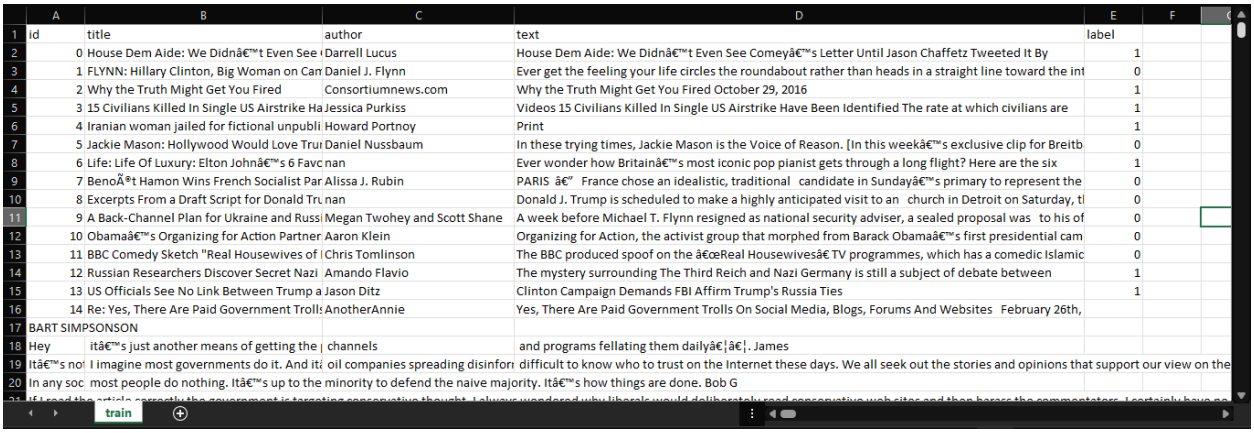
- Học máy là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.
- Công nghệ sử dụng:

- tfidfVectorizer: Chuyển đổi một bộ sưu tập các tài liệu thô sang một ma trận TF-IDF để mã hóa và đếm số lần xuất hiện từ của một kho tài liệu văn bản tối giản.
- LogisticRegression: Hồi quy logistic là một mô hình thống kê ở dạng cơ bản của mô hình logistic, sử dụng một hàm logistic để mô hình hóa một biến phụ thuộc nhị phân.

6. Demo chương trình

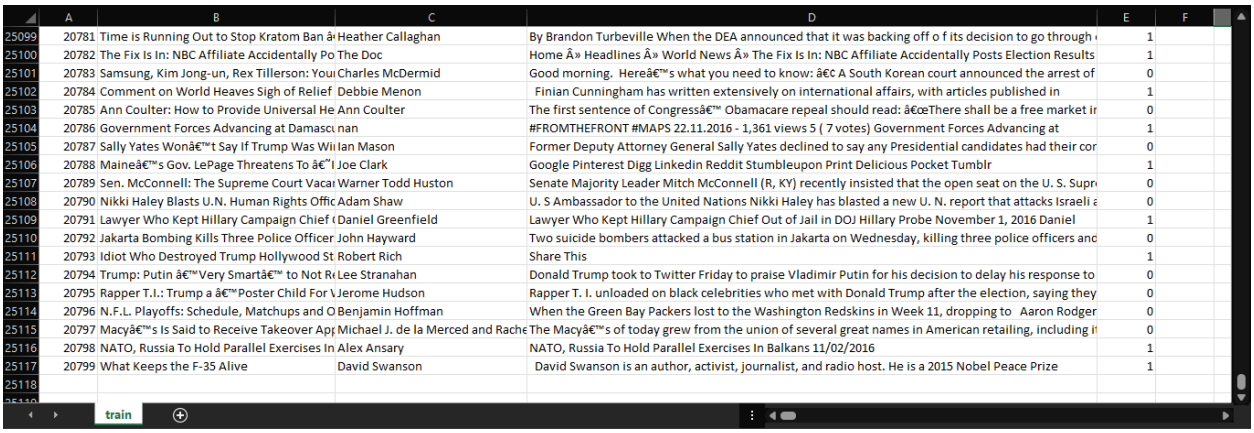
Phát hiện tin giả dựa trên nội dung

- Bộ dữ liệu:



	A	B	C	D	E	F
1	id	title	author	text	label	
2	0	House Dem Aide: We Didnâ€™t Even See (	Darrell Lucus	House Dem Aide: We Didnâ€™t Even See Comeyâ€™s Letter Until Jason Chaffetz Tweeted It By	1	
3	1	FLYNN: Hillary Clinton, Big Woman on Can	Daniel J. Flynn	Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the int	0	
4	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, 2016	1	
5	3	15 Civilians Killed In Single US Airstrike Ha	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstrike Have Been Identified The rate at which civilians are	1	
6	4	Iranian woman jailed for fictional unpubli	Howard Portnoy	Print	1	
7	5	Jackie Mason: Hollywood Would Love Trur	Daniel Nussbaum	In these trying times, Jackie Mason is the Voice of Reason. [In this weekâ€™s exclusive clip for Breitb	0	
8	6	Life: Life Of Luxury: Elton Johnâ€™s 6 Favc	nan	Ever wonder how Britainâ€™s most iconic pop pianist gets through a long flight? Here are the six	1	
9	7	BenoÃ®t Hamon Wins French Socialist Par	Alissa J. Rubin	PARIS â€" France chose an idealistic, traditional candidate in Sundayâ€™s primary to represent the	0	
10	8	Excerpts From a Draft Script for Donald Tru	nan	Donald J. Trump is scheduled to make a highly anticipated visit to an church in Detroit on Saturday, tl	0	
11	9	A Back-Channel Plan for Ukraine and Russi	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as national security adviser, a sealed proposal was to his of	0	
12	10	Obamaâ€™s Organizing for Action Partner	Aaron Klein	Organizing for Action, the activist group that morphed from Barack Obamaâ€™s first presidential cam	0	
13	11	BBC Comedy Sketch "Real Housewives of	Chris Tomlinson	The BBC produced spoof on the â€œReal Housewivesâ€ TV programmes, which has a comedic Islamic	0	
14	12	Russian Researchers Discover Secret Nazi	Amando Flavio	The mystery surrounding The Third Reich and Nazi Germany is still a subject of debate between	1	
15	13	US Officials See No Link Between Trump a	Jason Ditz	Clinton Campaign Demands FBI Affirm Trump's Russia Ties	1	
16	14	Re: Yes, There Are Paid Government Trolls	AnotherAnnie	Yes, There Are Paid Government Trolls On Social Media, Blogs, Forums And Websites February 26th,		
17	BART SIMPSONSON					
18	Hey	itâ€™s just another means of getting the	channels	and programs fellating them dailyâ€¦â€¦. James		
19	Itâ€™s not	I imagine most governments do it. And it	oil companies spreading disinfor	difficult to know who to trust on the Internet these days. We all seek out the stories and opinions that support our view on the		
20	In any soc	most people do nothing. Itâ€™s up to the	minority to defend the naive majority. Itâ€™s how things are done. Bob G			

train



	A	B	C	D	E	F
25099	20781	Time is Running Out to Stop Kratom Ban a	Heather Callaghan	By Brandon Turbeville When the DEA announced that it was backing off o f its decision to go through	1	
25100	20782	The Fix Is In: NBC Affiliate Accidentally Po	The Doc	Home Â» Headlines Â» World News Â» The Fix Is In: NBC Affiliate Accidentally Posts Election Results	1	
25101	20783	Samsung, Kim Jong-un, Rex Tillerson: Youi	Charles McDermid	Good morning. Hereâ€™s what you need to know: â€¢ A South Korean court announced the arrest of	0	
25102	20784	Comment on World Heaves Sigh of Relief	Debbie Menon	Finian Cunningham has written extensively on international affairs, with articles published in	1	
25103	20785	Ann Coulter: How to Provide Universal He	Ann Coulter	The first sentence of Congressâ€™ Obamacare repeal should read: â€œThere shall be a free market ir	0	
25104	20786	Government Forces Advancing at Damasc	nan	#FROMTHEFRONT #MAPS 22.11.2016 - 1,361 views 5 (7 votes) Government Forces Advancing at	1	
25105	20787	Sally Yates Wonâ€™t Say If Trump Was Wi	Ian Mason	Former Deputy Attorney General Sally Yates declined to say any Presidential candidates had their cor	0	
25106	20788	Maineâ€™s Gov. LePage Threatens To â€™I	Joe Clark	Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Pocket Tumblr	1	
25107	20789	Sen. McConnell: The Supreme Court Vacar	Warner Todd Huston	Senate Majority Leader Mitch McConnell (R, KY) recently insisted that the open seat on the U. S. Supr	0	
25108	20790	Nikki Haley Blasts U.N. Human Rights Offic	Adam Shaw	U. S Ambassador to the United Nations Nikki Haley has blasted a new U. N. report that attacks Israeli a	0	
25109	20791	Lawyer Who Kept Hillary Campaign Chief (	Daniel Greenfield	Lawyer Who Kept Hillary Campaign Chief Out of Jail in DOJ Hillary Probe November 1, 2016 Daniel	1	
25110	20792	Jakarta Bombing Kills Three Police Officer	John Hayward	Two suicide bombers attacked a bus station in Jakarta on Wednesday, killing three police officers and	0	
25111	20793	Idiot Who Destroyed Trump Hollywood St	Robert Rich	Share This	1	
25112	20794	Trump: Putin â€™Very Smartâ€™ to Not R	Lee Stranahan	Donald Trump took to Twitter Friday to praise Vladimir Putin for his decision to delay his response to	0	
25113	20795	Rapper T.I.: Trump a â€™Poster Child For V	Jerome Hudson	Rapper T. I. unloaded on black celebrities who met with Donald Trump after the election, saying they	0	
25114	20796	N.F.L. Playoffs: Schedule, Matchups and O	Benjamin Hoffman	When the Green Bay Packers lost to the Washington Redskins in Week 11, dropping to Aaron Rodger	0	
25115	20797	Macyâ€™s Is Said to Receive Takeover Ap	Michael J. de la Merced and Rache	The Macyâ€™s of today grew from the union of several great names in American retailing, including i	0	
25116	20798	NATO, Russia To Hold Parallel Exercises In	Alex Ansary	NATO, Russia To Hold Parallel Exercises In Balkans 11/02/2016	1	
25117	20799	What Keeps the F-35 Alive	David Swanson	David Swanson is an author, activist, journalist, and radio host. He is a 2015 Nobel Peace Prize	1	
25118						

train

- Nhập Packages cần thiết:

```
[1]  import numpy as np
     import pandas as pd
     import re
     from nltk.corpus import stopwords
     from nltk.stem.porter import PorterStemmer
     from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
     ✓ 4.3s
```

```
[2]  import nltk
     nltk.download('stopwords')
     ✓ 0.3s
...  [nltk_data] Downloading package stopwords to
     [nltk_data]     C:\Users\lethi\AppData\Roaming\nltk_data...
     [nltk_data]   Package stopwords is already up-to-date!
     True
```

- Đọc bộ dữ liệu:

```
[3]  dataset = pd.read_csv('train.csv')
     dataset.shape
     ✓ 2.4s
...  (20800, 5)
```

- Thay thế các giá trị null bằng chuỗi rỗng, đồng thời hợp nhất tên tác giả và tiêu đề tin tức

```
[6]  dataset = dataset.fillna('')
     ✓ 0.1s
```

```
[7]  dataset['content'] = dataset['author']+' '+ dataset['title']
     ✓ 0.1s
```

- Tách nhãn ra khỏi dữ liệu

```
[9]  X = dataset.drop(columns='label', axis=1)
     Y = dataset['label']
     ✓ 0.6s
```

- Loại bỏ các hậu tố hoặc tiền tố được sử dụng với một từ

```
[12] port_stem = PorterStemmer()
     ✓ 0.4s

[13] def stemming(content):
         stemmed_content = re.sub('[^a-zA-Z]',' ',content)
         stemmed_content = stemmed_content.lower()
         stemmed_content = stemmed_content.split()
         stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
         stemmed_content = ' '.join(stemmed_content)
         return stemmed_content
     ✓ 0.3s

[14] dataset['content'] = dataset['content'].apply(stemming)
     ✓ 2m 58.2s
```

- Trả dữ liệu về values

```
y = dataset["label"]
x = dataset["content"]
✓ 0.7s
```

- Chuyển đổi dữ liệu văn bản sang dữ liệu số

```
vectorizer = TfidfVectorizer(tokenizer=stemming)
X = vectorizer.fit_transform(x.values.astype('U'))
✓ 3m 0.1s
```

- Tách tập dữ liệu thành dữ liệu đào tạo và kiểm tra

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
✓ 0.7s
```

- Huấn luyện mô hình hồi quy Logistic

```
logit = LogisticRegression()
logit.fit(X_train, y_train)
✓ 0.4s
LogisticRegression()
```

- Đánh giá
    - Độ chính xác trên tập dữ liệu huấn luyện

```
print("Accuracy ",logit.score(X_train, y_train))
✓ 0.9s
Accuracy  0.8073918269230769
```

    - Độ chính xác trên tập dữ liệu kiểm tra

```
print("Accuracy ",logit.score(X_test, y_test))
✓ 0.7s
Accuracy  0.8069711538461538
```

- Kiểm tra Hệ thống Dự đoán

```
X_predict = ["""Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination? /
[Hillary Clinton remains the big woman on campus in leafy, liberal Wellesley, Massachusetts. Everywhere else votes her most likely to don her /
inauguration dress for the remainder of her days the way Miss Havisham forever wore that wedding dress.  Speaking of Great Expectations, /
Hillary Rodham overflowed with them 48 years ago when she first addressed a Wellesley graduating class. The president of the college informed those/
 gathered in 1969 that the students needed â€œno debate so far as I could ascertain as to who their spokesman was to beâ€ (kind of the like the /
 Democratic primaries in 2016 minus the    terms unknown then even at a Seven Sisters school). â€œI am very glad that Miss Adams made it clear that /
 what I am speaking for today is all of us â€”  the 400 of us,â€ Miss Rodham told her classmates. After appointing herself Edger Bergen to the Charlie/
  McCarthys and Mortimer Snerds in attendance, the     bespectacled in granny glasses (awarding her matronly wisdom â€”  or at least John Lennon wisdom) /
  took issue with the previous speaker. Despite becoming the first    to win election to a seat in the U. S. Senate since Reconstruction, Edward Brooke /
  came in for criticism for calling for â€œempathyâ€ for the goals of protestors as he criticized tactics. Though Clinton in her senior thesis on Saul /
  Alinsky lamented â€œBlack Power demagoguesâ€ and â€œelitist arrogance and repressive intoleranceâ€ within the New Left, similar words coming out of a /
  Republican necessitated a brief rebuttal. â€œTrust,â€ Rodham ironically observed in 1969, â€œthis is one word that when I asked the class at our rehearsal/
   what it was they wanted me to say for them, everyone came up to me and said â€˜Talk about trust, talk about the lack of trust both for us and the way /
   we feel about others. Talk about the trust bust.â€™ What can you say about it? What can you say about a feeling that permeates a generation and that /
   perhaps is not even understood by those who are distrusted?â€ The â€œtrust bustâ€ certainly busted Clintonâ€™s 2016 plans. She certainly did not /
   even understand that people distrusted her. After Whitewater, Travelgate, the vast    conspiracy, Benghazi, and the missing emails, Clinton found herself /
   the distrusted voice on Friday. There was a load of compromising on the road to the broadening of her political horizons. And distrust from the /
   American people â€”  Trump edged her 48 percent to 38 percent on the question immediately prior to Novemberâ€™s election â€”  stood as a major /
   reason for the closing of those horizons. Clinton described her vanquisher and his supporters as embracing a â€œlie,â€ a â€œcon,â€ â€œalternative /
   facts,â€ and â€œa    assault on truth and reason. â€ She failed to explain why the American people chose his lies over her truth. â€œAs the /
   history majors among you here today know all too well, when people in power invent their own facts and attack those who question them, /
   it can mark the beginning of the end of a free society,â€ she offered. â€œThat is not hyperbole. â€ Like so many people to emerge from the 1960s, /
   Hillary Clinton embarked upon a long, strange trip. From high school Goldwater Girl and Wellesley College Republican president to Democratic politician, /
   Clinton drank in the times and the place that gave her a degree. More significantly, she went from idealist to cynic, as a comparison of her /
   two Wellesley commencement addresses show. Way back when, she lamented that â€œfor too long our leaders have viewed politics as the art of the possible, /
   and the challenge now is to practice politics as the art of making what appears to be impossible possible. â€ Now, as the big woman on campus but /
   the odd woman out of the White House, she wonders how her current station is even possible. â€œWhy arenâ€™t I 50 points ahead?â€ she asked in September. /
   In May she asks why she isnâ€™t president. The woman famously dubbed a â€œcongenital liarâ€ by Bill Safire concludes that lies did her in â€”  theirs, /
   mind you, not hers. Getting stood up on Election Day, like finding yourself the jilted bride on your wedding day, inspires dangerous delusions."""]
```

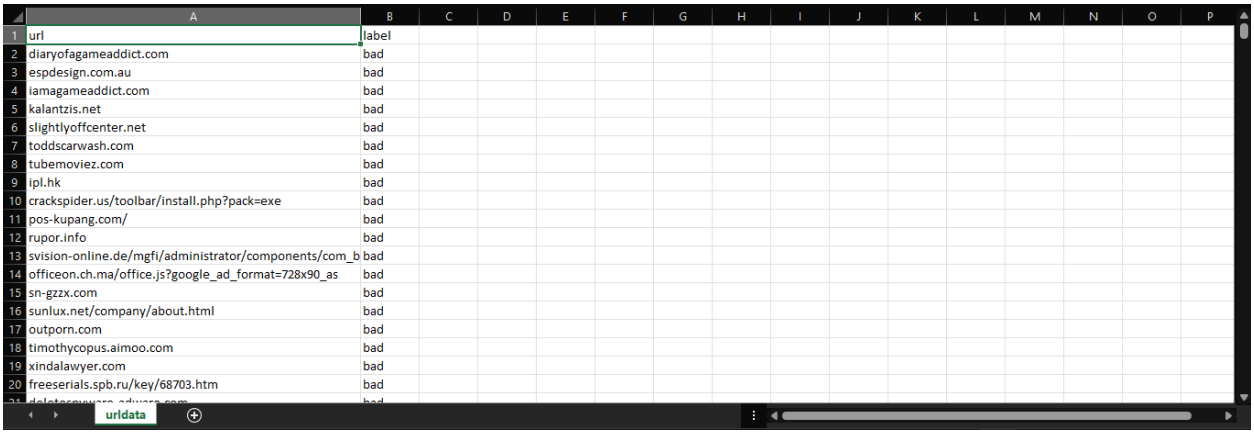
```
X_predict = vectorizer.transform(X_predict)
New_predict = logit.predict(X_predict)
if (New_predict==0):
    print('The news is Real')
else:
    print('The news is Fake')
```

✓ 0.4s

```
The news is Real
```

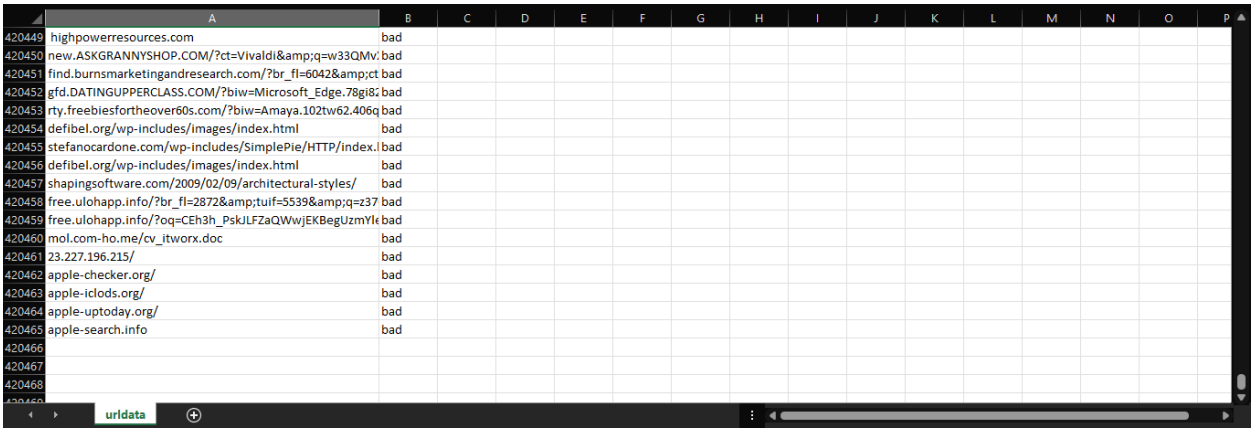
Phát hiện tin giả dựa trên ngữ cảnh xã hội

- Bộ dữ liệu:



	A	B
1	url	label
2	diaryofagameaddict.com	bad
3	espdesign.com.au	bad
4	iamagameaddict.com	bad
5	kalantzis.net	bad
6	slightlyoffcenter.net	bad
7	toddscarwash.com	bad
8	tubemoviez.com	bad
9	ipl.hk	bad
10	crackspider.us/toolbar/install.php?pack=exe	bad
11	pos-kupang.com/	bad
12	rupor.info	bad
13	svision-online.de/mgfi/administrator/components/com_b	bad
14	officeon.ch.ma/office.js?google_ad_format=728x90_as	bad
15	sn-gzzx.com	bad
16	sunlux.net/company/about.html	bad
17	outporn.com	bad
18	timothycopus.aimoo.com	bad
19	xindalawyer.com	bad
20	freeserials.spb.ru/key/68703.htm	bad

urldata



	A	B
420449	highpowerresources.com	bad
420450	new.ASKGRANNYSHOP.COM/?ct=Vivaldi&q=w33QMv	bad
420451	find.burnsmarketingandresearch.com/?br_fl=6042&ct	bad
420452	gfd.DATINGUPPERCLASS.COM/?biw=Microsoft_Edge.78gi8	bad
420453	rty.freebiesfortheover60s.com/?biw=Amaya.102tw62.406q	bad
420454	defibel.org/wp-includes/images/index.html	bad
420455	stefanocardone.com/wp-includes/SimplePie/HTTP/index.l	bad
420456	defibel.org/wp-includes/images/index.html	bad
420457	shapingsoftware.com/2009/02/09/architectural-styles/	bad
420458	free.ulohapp.info/?br_fl=2872&tuif=5539&q=z37	bad
420459	free.ulohapp.info/?oq=CEh3h_PskJLFZaQWwjEKBegUzmYI	bad
420460	mol.com-ho.me/cv_itworx.doc	bad
420461	23.227.196.215/	bad
420462	apple-checker.org/	bad
420463	apple-iclods.org/	bad
420464	apple-uptoday.org/	bad
420465	apple-search.info	bad
420466		
420467		
420468		

urldata

- Nhập các Packages

```
import pandas as pd
import numpy as np
import random
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```
✓ 2.9s

- Đọc bộ dữ liệu

```
urls_data = pd.read_csv("urldata.csv")
```
✓ 0.9s

- Tách các kí tự

```
def makeTokens(f):
    tkns_BySlash = str(f.encode('utf-8')).split('/')
    total_Tokens = []
    for i in tkns_BySlash:
        tokens = str(i).split('-')
        tkns_ByDot = []
        for j in range(0,len(tokens)):
            temp_Tokens = str(tokens[j]).split('.')
            tkns_ByDot = tkns_ByDot + temp_Tokens
        total_Tokens = total_Tokens + tokens + tkns_ByDot
    total_Tokens = list(set(total_Tokens))
    if 'com' in total_Tokens:
        total_Tokens.remove('com')
    return total_Tokens
```
✓ 0.2s

- Tách nhãn ra khỏi dữ liệu

```
y = urls_data["label"]
```
✓ 0.3s

```
url_list = urls_data["url"]
```
✓ 0.3s

```
vectorizer = TfidfVectorizer(tokenizer=makeTokens)
✓ 0.3s
```

```
X = vectorizer.fit_transform(url_list)
✓ 9.6s
```

- Tách tập dữ liệu thành dữ liệu đào tạo và kiểm tra

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
✓ 0.1s
```

- Huấn luyện mô hình hồi quy Logistic

```
logit = LogisticRegression()
logit.fit(X_train, y_train)
✓ 26.4s
```

- Độ chính xác trên tập dữ liệu huấn luyện

```
print("Accuracy ",logit.score(X_train, y_train))
✓ 1.1s
Accuracy  0.9724322251323687
```

- Độ chính xác trên tập dữ liệu kiểm tra

```
print("Accuracy ",logit.score(X_test, y_test))
✓ 0.3s
Accuracy  0.9618041929768233
```

- Kiểm tra hệ thống dự đoán

```
X_predict = ["google.com/search=jcharistech",
"google.com/search=faizanahmad",
"pakistanifacebookforever.com/getpassword.php/",
"www.radsport-voggel.de/wp-admin/includes/log.exe",
"ahrenhei.without-transfer.ru/nethost.exe ",
"www.itidea.it/centroesteticosothys/img/_notes/gum.exe"]
```
✓ 0.3s

```
X_predict = vectorizer.transform(X_predict)
New_predict = logit.predict(X_predict)
```
✓ 0.4s

```
print(New_predict)
```
✓ 0.3s

```
['good' 'good' 'good' 'bad' 'bad' 'bad']
```

7. Kết luận

Nhiệm vụ phân loại tin tức theo cách thủ công đòi hỏi kiến thức chuyên sâu về lĩnh vực và kiến thức chuyên môn để xác định các điểm bất thường trong văn bản. Phương pháp được sử dụng trong bài báo cáo này cụ thể là hồi quy Logistic.

Trong quá trình nghiên cứu, tìm hiểu và hoàn thành bài báo cáo, em đã thu nhận được thêm những kiến thức và em cũng nhận thấy xử lý ngôn ngữ tự nhiên là một lĩnh vực nghiên cứu rộng lớn, còn nhiều điều cần phải khám phá. Trong đề tài "Phát hiện tin tức giả mạo" em đã cố gắng tập trung tìm hiểu về xử lý ngôn ngữ tự nhiên, một số thuật toán phân lớp. Từ đó em đã xây dựng được chương trình mô phỏng "Phát hiện tin tức giả mạo". Do thời gian thực hiện bài báo cáo còn hạn chế nên em mới chỉ tìm hiểu được một số bước trong quá trình xử lý ngôn ngữ tự nhiên và chương trình mô phỏng còn chưa được hoàn thiện như mong muốn. Trong thời gian tới em sẽ cố gắng tiếp tục nghiên cứu và hoàn thiện việc tìm hiểu xử lý ngôn ngữ tự nhiên và chương trình mô phỏng "Phát hiện tin tức giả mạo".

8. Tài liệu tham khảo

[1] https://www.kaggle.com/teseract/urldataset

[2] https://www.kaggle.com/c/fake-news/data

[3] https://www.kaggle.com/jihenbelhoudi/fake-news-classification-using-logistic-regression

[4] https://www.hindawi.com/journals/complexity/2020/8885861/

[5] NICT24_paper_66.pdf

[6]https://www.researchgate.net/publication/348580170_Identifying_Fake_News_on_Social_Networks_Based_on_Natural_Language_Processing_Trends_and_Challenges