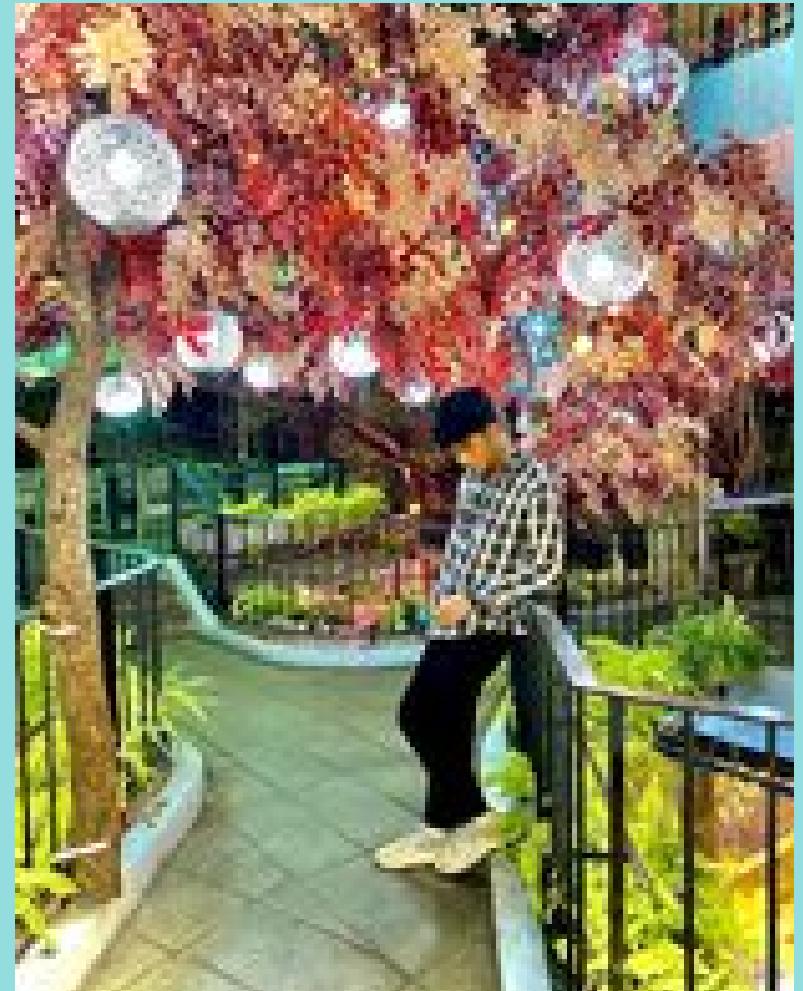




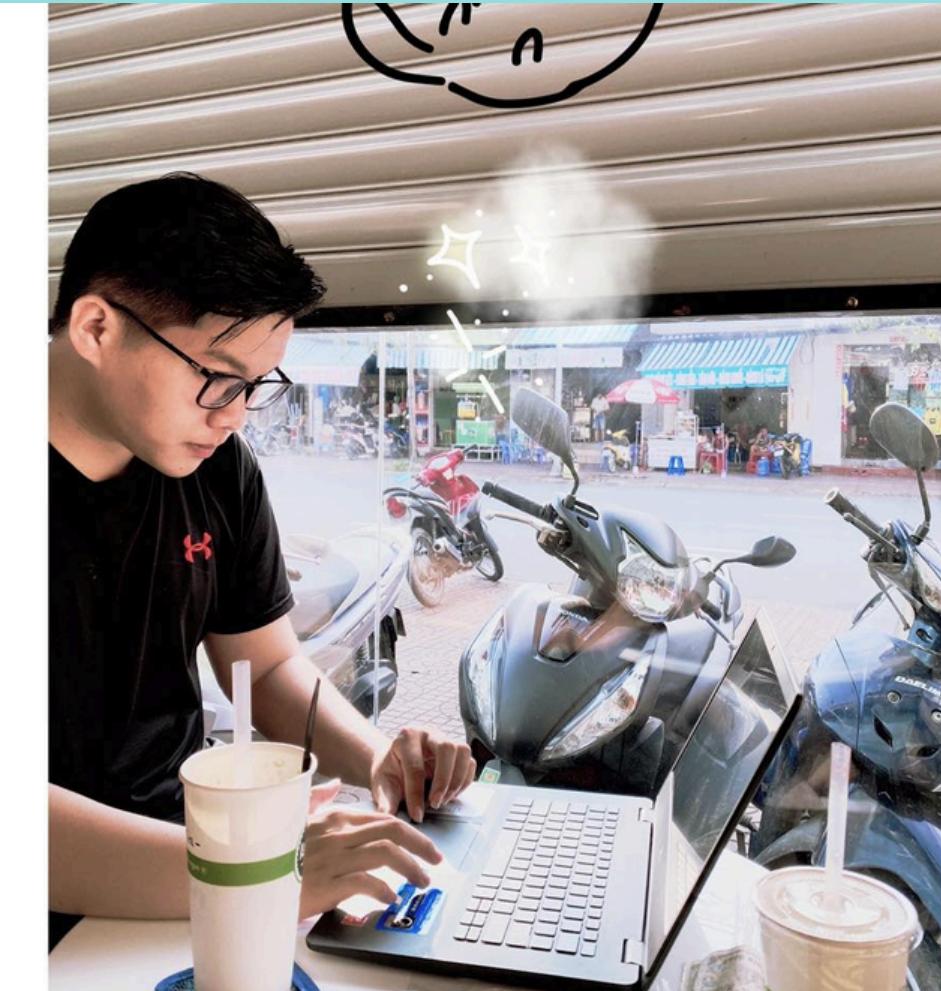
Phân tích về giả mạo trên mạng xã hội

Điểm qua các đặc điểm và cách nhận
biết thông tin sai lệch trên mạng xã hội

Thành viên nhóm 9



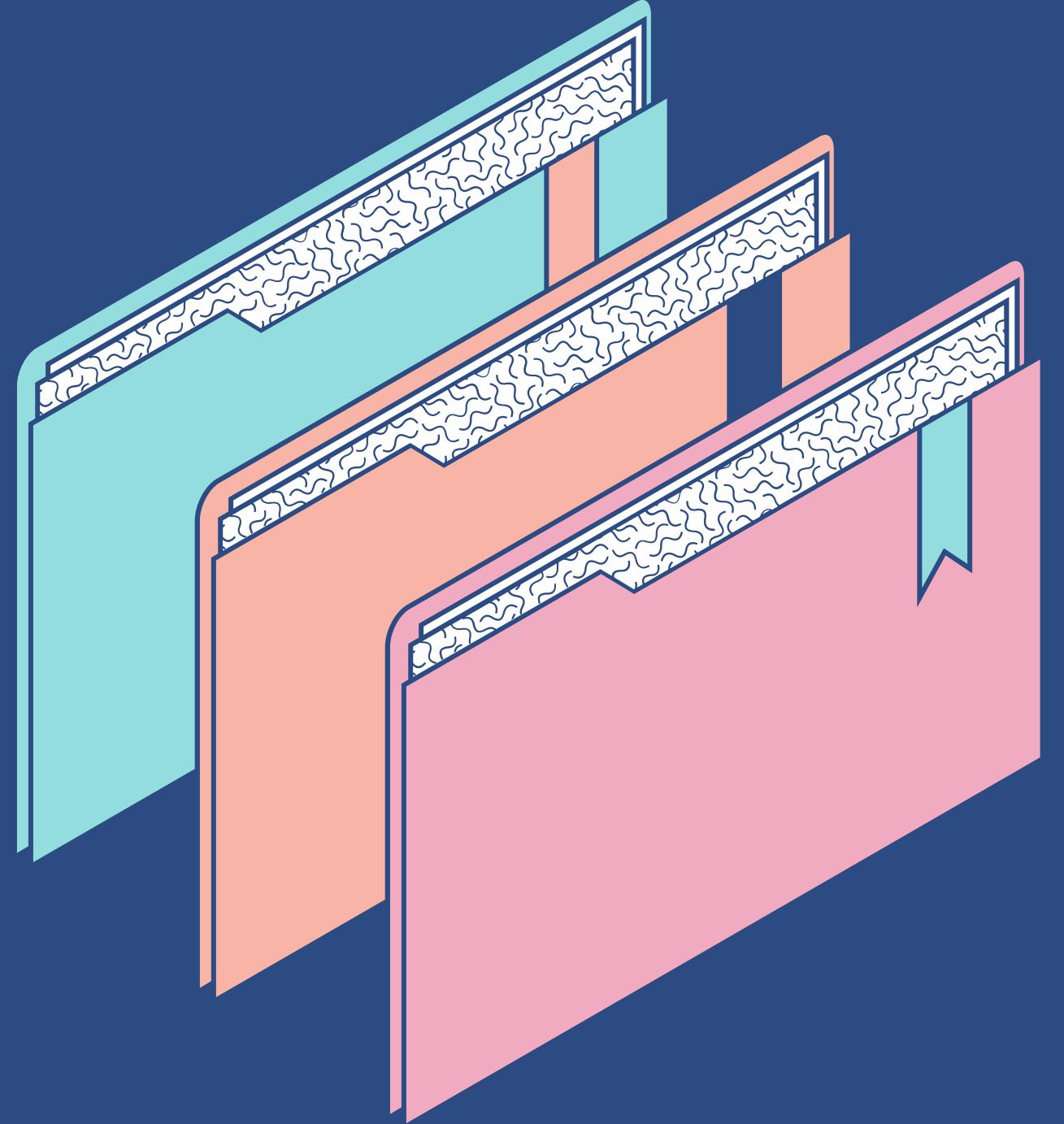
2151010354
Lại Thành Thi



2151010306
Tô Thái Việt Quang



2151010318
Phan Quang Sang



Nội dung

- Sơ lược về sự giả mạo trên mạng xã hội
- Giới thiệu bộ dữ liệu
- Tiền xử lý dữ liệu
- Số liệu cho thấy sự giả mạo rộng rãi trên mạng xã hội
- Cách phòng tránh và nhận biết đâu là tin giả, tin thật
- Kết luận
- Q & A

Sơ lược về giả mạo trên mạng xã hội

- Các thông tin sai sự thật
- Nhiều trang web giả mạo trang chính với mưu đồ bất chính
- Các trang mạng đăng thông tin tuyển dụng sai lệch nhằm thu hút người dùng
- Nhằm vào các đối tượng già, trẻ để làm sai lệch suy nghĩ



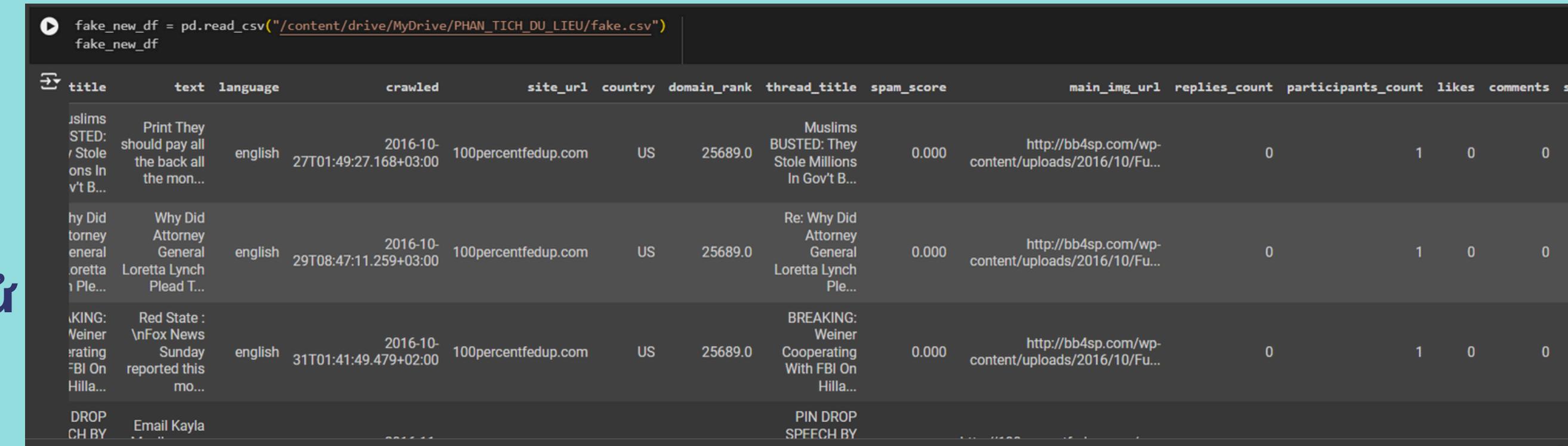
Sơ lược về giả mạo trên mạng xã hội



Giới thiệu về bộ dữ liệu

Có các cột chủ yếu như sau:

- **title:** Tiêu đề
- **language:** Ngôn ngữ sử dụng trong bài báo
- **crawled:** Ngày bài báo được lưu trữ
- **site_url:** URL của các trang báo
- **country:** thành phố nơi các bài báo lưu hành
- **like:** lượt like trên các bài báo
- **comment:** Bình luận trên các bài báo



title	text	language	crawled	site_url	country	domain_rank	thread_title	spam_score	main_img_url	replies_count	participants_count	likes	comments
Muslims STED: / Stole ons In v't B...	Print They should pay all the back all the mon...	english	2016-10- 27T01:49:27.168+03:00	100percentfedup.com	US	25689.0	Muslims BUSTED: They Stole Millions In Gov't B...	0.000	http://bb4sp.com/wp- content/uploads/2016/10/Fu...	0	1	0	0
hy Did torney eneral orett Loretta Lynch Plead T...	Why Did Attorney General Loretta Lynch Plead T...	english	2016-10- 29T08:47:11.259+03:00	100percentfedup.com	US	25689.0	Re: Why Did Attorney General Loretta Lynch Ple...	0.000	http://bb4sp.com/wp- content/uploads/2016/10/Fu...	0	1	0	0
KING: Veiner erating FBI On Hilla...	Red State : \nFox News Sunday reported this mo...	english	2016-10- 31T01:41:49.479+02:00	100percentfedup.com	US	25689.0	BREAKING: Weiner Cooperating With FBI On Hilla...	0.000	http://bb4sp.com/wp- content/uploads/2016/10/Fu...	0	1	0	0
DROP CH RY	Email Kayla						PIN DROP SPEECH RY						

Tiền xử lý dữ liệu

```
[ ] # Hiển thị thông tin của bộ dữ liệu  
fake_new_df.info()  
  
→ <class 'pandas.core.frame.DataFrame'>  
RangeIndex: 12999 entries, 0 to 12998  
Data columns (total 20 columns):  
 #   Column           Non-Null Count  Dtype     
---  --  
 0   uuid             12999 non-null   object    
 1   ord_in_thread    12999 non-null   int64     
 2   author           10575 non-null   object    
 3   published        12999 non-null   object    
 4   title            12319 non-null   object    
 5   text              12953 non-null   object    
 6   language          12999 non-null   object    
 7   crawled           12999 non-null   object    
 8   site_url          12999 non-null   object    
 9   country           12823 non-null   object    
 10  domain_rank      8776 non-null   float64   
 11  thread_title     12987 non-null   object    
 12  spam_score       12999 non-null   float64   
 13  main_img_url     9356 non-null   object    
 14  replies_count    12999 non-null   int64     
 15  participants_count 12999 non-null   int64     
 16  likes             12999 non-null   int64     
 17  comments          12999 non-null   int64     
 18  shares            12999 non-null   int64     
 19  type              12999 non-null   object    
dtypes: float64(2), int64(6), object(12)  
memory usage: 2.0+ MB
```



Tiền xử lý dữ liệu

```
# Xem các thông số thống kê của bộ dữ liệu.  
fake_new_df.describe()
```

	ord_in_thread	domain_rank	spam_score	replies_count	participants_count	likes	comments	shares
count	12999.000000	8776.000000	12999.000000	12999.000000	12999.000000	12999.000000	12999.000000	12999.000000
mean	0.891530	38092.996582	0.026122	1.383183	1.727518	10.831833	0.038311	10.831833
std	6.486822	26825.487454	0.122889	9.656838	6.884239	79.798949	0.827335	79.798949
min	0.000000	486.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	17423.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
50%	0.000000	34478.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
75%	0.000000	60570.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
max	100.000000	98679.000000	1.000000	309.000000	240.000000	988.000000	65.000000	988.000000

```
[ ] # Xem số lượng dòng, cột của bộ dữ liệu.  
fake_new_df.shape
```

```
(12999, 20)
```



Tiến xử lý dữ liệu

```
[ ] # Xóa các cột không cần thiết  
fake_new_df.drop(['domain_rank', 'thread_title', 'main_img_url'], axis = 1, inplace = True)  
# Chuẩn hóa dữ liệu của DataFrame df  
fake_new_df.rename(columns={'uuid':'id'},inplace=True)
```

```
[ ] # Chuyển đổi cột published và crawled từ object sang ngày giờ  
fake_new_df['published']=pd.to_datetime(fake_new_df['published'],utc=True)  
fake_new_df['crawled']=pd.to_datetime(fake_new_df['crawled'],utc=True)
```

```
[ ] # Thêm vào cột 'crawled_by_day' dựa vào cột 'crawled'  
fake_new_df.insert(8,'crawled_by_day',pd.to_datetime(fake_new_df['crawled'],utc=True).dt.strftime('%Y-%m-%d'))  
# Thêm vào cột 'published_by_day' dựa vào cột "published"  
fake_new_df.insert(4,'published_by_day',pd.to_datetime(fake_new_df['published'],utc=True).dt.strftime('%Y-%m-%d'))
```

```
[ ] # Chuyển đổi chữ cái đầu tiên của mỗi từ trong chuỗi thành chữ cái in hoa trong danh sách cột  
fake_new_df.columns = fake_new_df.columns.str.capitalize()
```

```
[ ] # Kiểm tra giá trị trùng lặp  
fake_new_df.duplicated().sum()
```



Tiến x^u_u lý d^ữ liệu

▼ Thông tin số lần xuất hiện của mỗi giá trị sau khi xử lý dữ liệu

```
▶ fake_new_df.Ord_in_thread.value_counts()
```

```
→ Ord_in_thread  
0    9706  
Name: count, dtype: int64
```

```
[ ] fake_new_df.Author.value_counts()
```

```
→ Author  
admin                246  
Editor               100  
Gillian              100  
Starkman              100  
Alex Ansary           100  
...  
11 Things To Let Go Of Before The New Year - Motivate3.com      1  
Jing Jin                1  
USA Today               1  
Millie Weaver             1  
George Washington          1  
Name: count, Length: 1826, dtype: int64
```



Tiền xử lý dữ liệu



```
fake_new_df.Published_by_day.value_counts()
```

```
Published_by_day
```

```
2016-10-27    1378
2016-10-28    1152
2016-10-26     693
2016-11-01     589
2016-11-02     577
2016-10-31     560
2016-10-29     550
2016-11-03     488
2016-11-04     399
2016-10-30     396
2016-11-07     309
2016-11-08     274
2016-11-09     236
2016-11-05     208
2016-11-06     200
2016-11-10     189
```

```
2016-11-11    174
2016-11-16    132
2016-11-14    129
2016-11-15    117
2016-11-12    109
2016-11-23    108
2016-11-17    106
2016-11-18     99
2016-11-21     99
2016-11-13     98
2016-11-22     95
2016-11-19     72
2016-11-24     66
2016-11-20     53
2016-11-25     50
2016-10-25      1
```

```
Name: count, dtype: int64
```



Tiền xử lý dữ liệu

```
[ ] fake_new_df.Title.value_counts()
```

Title	count
Get Ready For Civil Unrest: Survey Finds That Most Americans Are Concerned About Election Violence	6
Will Barack Obama Delay Or Suspend The Election If Hillary Is Forced Out By The New FBI Email Investigation?	5
Let's Be Clear – A Vote For Warmonger Hillary Clinton Is A Vote For World War 3	5
"If Trump Loses, I'm Grabbing My Musket": Former Congressman Ready to Go Full Revolution	5
If Hillary Clinton Is Charged With Obstruction Of Justice She Could Go To Prison For 20 Years	5
..	..
Trump's Amazing Victory Against a Stacked Deck	1
Pope Francis: "...it is the communists who think like Christians"	1
WATCH! Six Communists Arrested After Attacking Trump Supporters, Media Yawns (Video)	1
No, Hate Crimes Have NOT 'Intensified' Since Trump's Election	1
ObamaCare Architect Admits "The Law Is Working As Designed" As Premiums Spike	1
Name: count, Length: 9216, dtype: int64	



Tiền xử lý dữ liệu

```
fake_new_df.Language.value_counts()
```

Language

english	9325
spanish	146
german	108
russian	58
french	27
turkish	9
ignore	7
italian	7
arabic	5
portuguese	5
norwegian	3
greek	2
finnish	1
chinese	1
polish	1
dutch	1
Name: count, dtype: int64	

Crawled_by_day

2016-10-27	1332
2016-10-28	1198
2016-10-26	634
2016-11-01	584
2016-10-29	574
2016-11-02	568
2016-10-31	561
2016-11-03	489
2016-11-04	380
2016-10-30	377
2016-11-07	311
2016-11-09	274
2016-11-05	249
2016-11-08	241
2016-11-06	197

2016-11-10	193
------------	-----

2016-11-11	184
---	---
2016-11-14	129
---	---
2016-11-16	127
---	---
2016-11-17	115
---	---
2016-11-23	111
---	---
2016-11-12	110
---	---
2016-11-15	109
---	---
2016-11-18	101
---	---
2016-11-13	101
---	---
2016-11-21	95
---	---
2016-11-22	90
---	---
2016-11-25	74
---	---
2016-11-19	70
---	---
2016-11-20	66
---	---
2016-11-24	62
---	---
Name: count, dtype: int64	



Tiền xử lý dữ liệu

```
[ ] fake_new_df.Site_url.value_counts()
```

Site_url	count
thedailysheep.com	100
conservativetribune.com	100
thecommonsenseshow.com	100
naturalblaze.com	100
nakedcapitalism.com	100
...	
therundownlive.com	1
educateinspirechange.org	1
blacklistednews.com	1
opednews.com	1
reductress.com	1
Name: count, Length: 181, dtype: int64	

```
[ ] fake_new_df.Country.value_counts()
```

Country	count
US	8433
GB	542
RU	124
EU	111
TV	101
ES	100
IS	93
DE	62
FR	36
NL	34
ME	34
IN	23
BG	6
CA	3
ZA	2
CO	1
SE	1
Name: count, dtype: int64	



Tiền xử lý dữ liệu

```
fake_new_df.Likes.value_counts()
```

```
↳ Likes
0      9376
3       9
1       8
2       5
12      5
...
328     1
696     1
87      1
397     1
714     1
Name: count, Length: 231, dtype: int64
```

```
fake_new_df.Comments.value_counts()
```

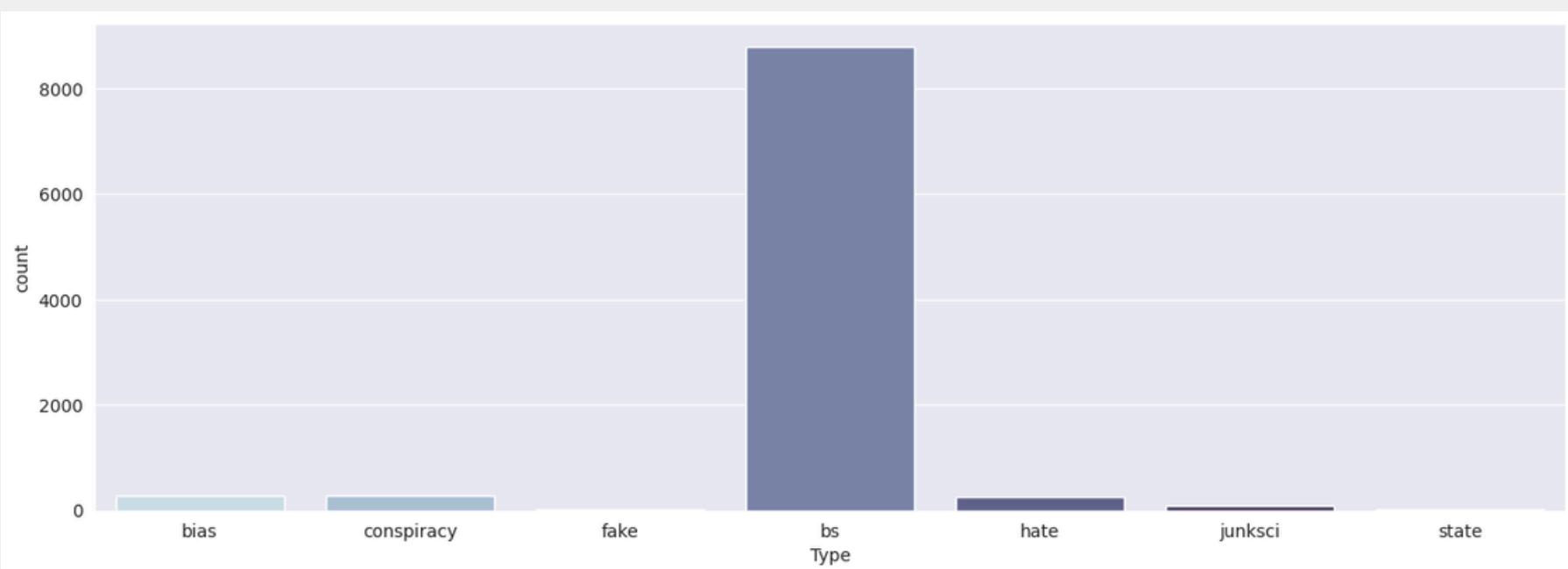
```
↳ Comments
0      9668
3       8
1       6
2       5
4       4
9       3
8       2
17      2
7       2
5       2
6       1
26      1
30      1
15      1
Name: count, dtype: int64
```

```
fake_new_df.Shares.value_counts()
```

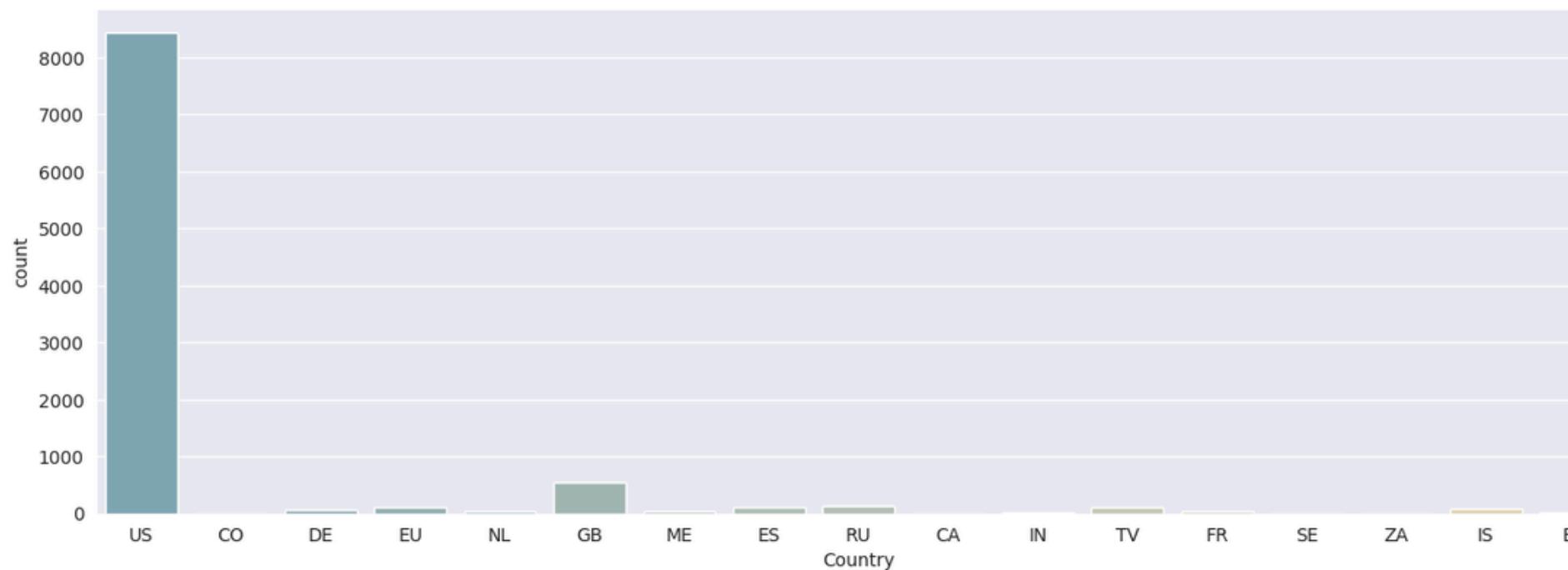
```
↳ Shares
0      9376
3       9
1       8
2       5
12      5
...
328     1
696     1
87      1
397     1
714     1
Name: count, Length: 231, dtype:
```



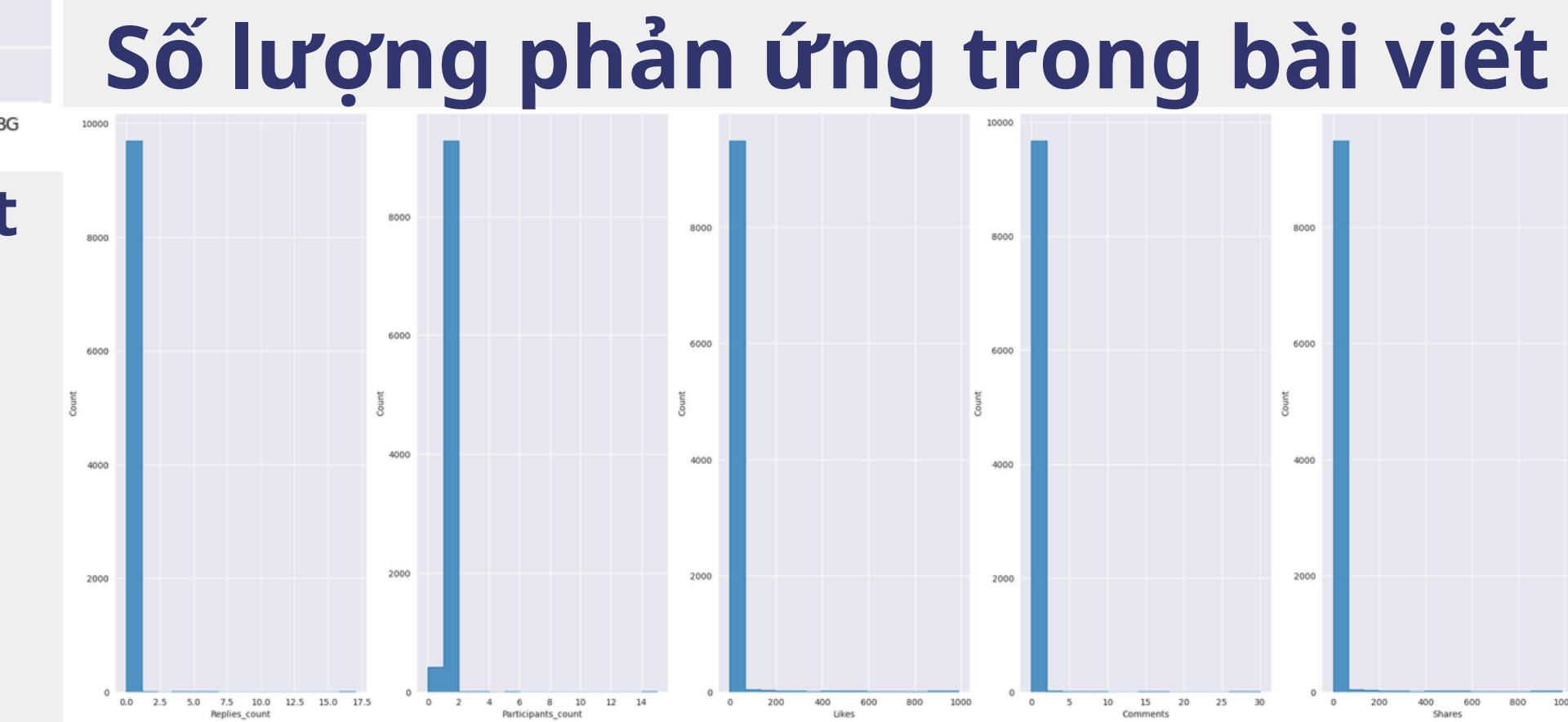
Số liệu về giả mạo, gian lận trên mạng xã hội



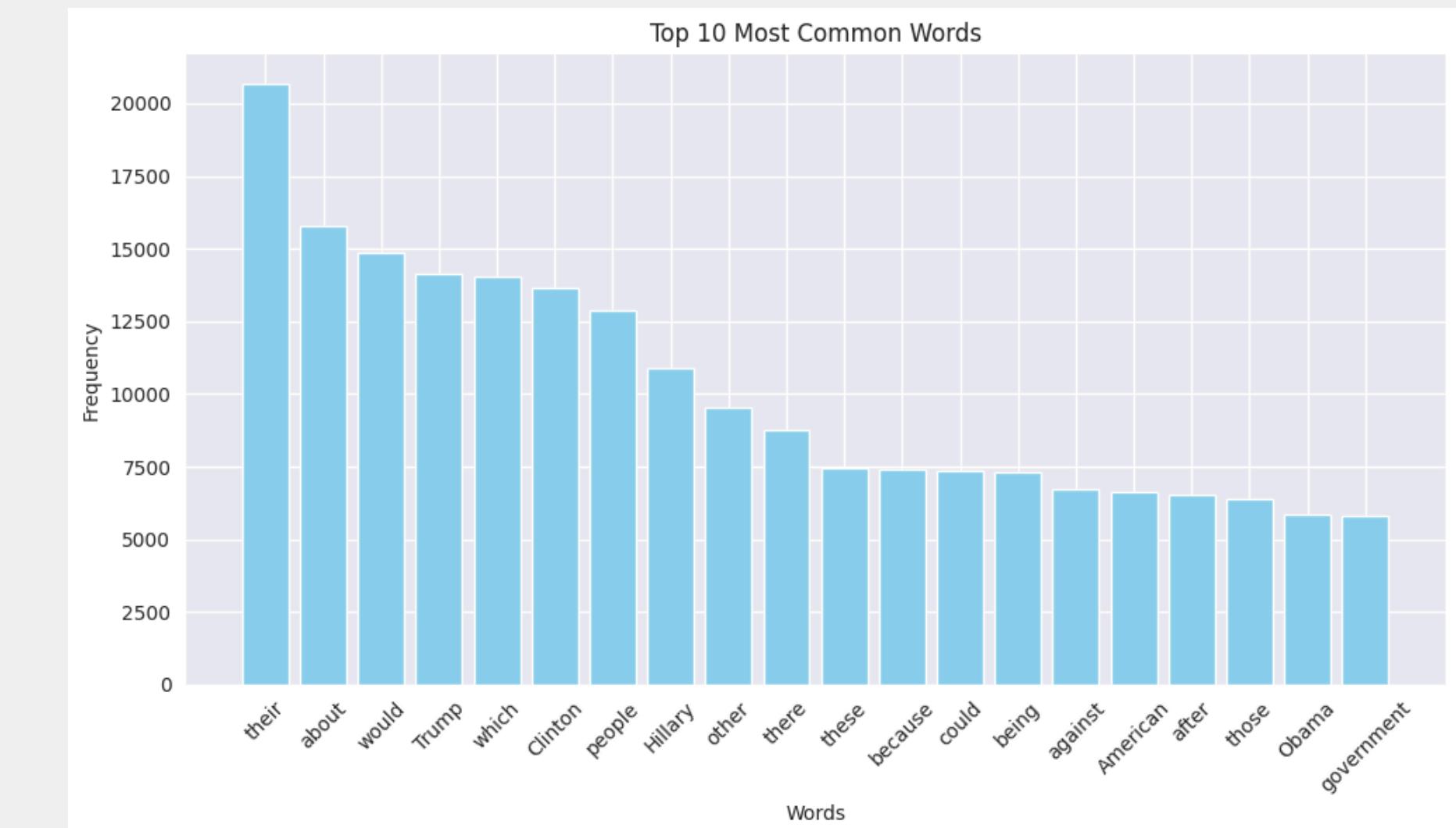
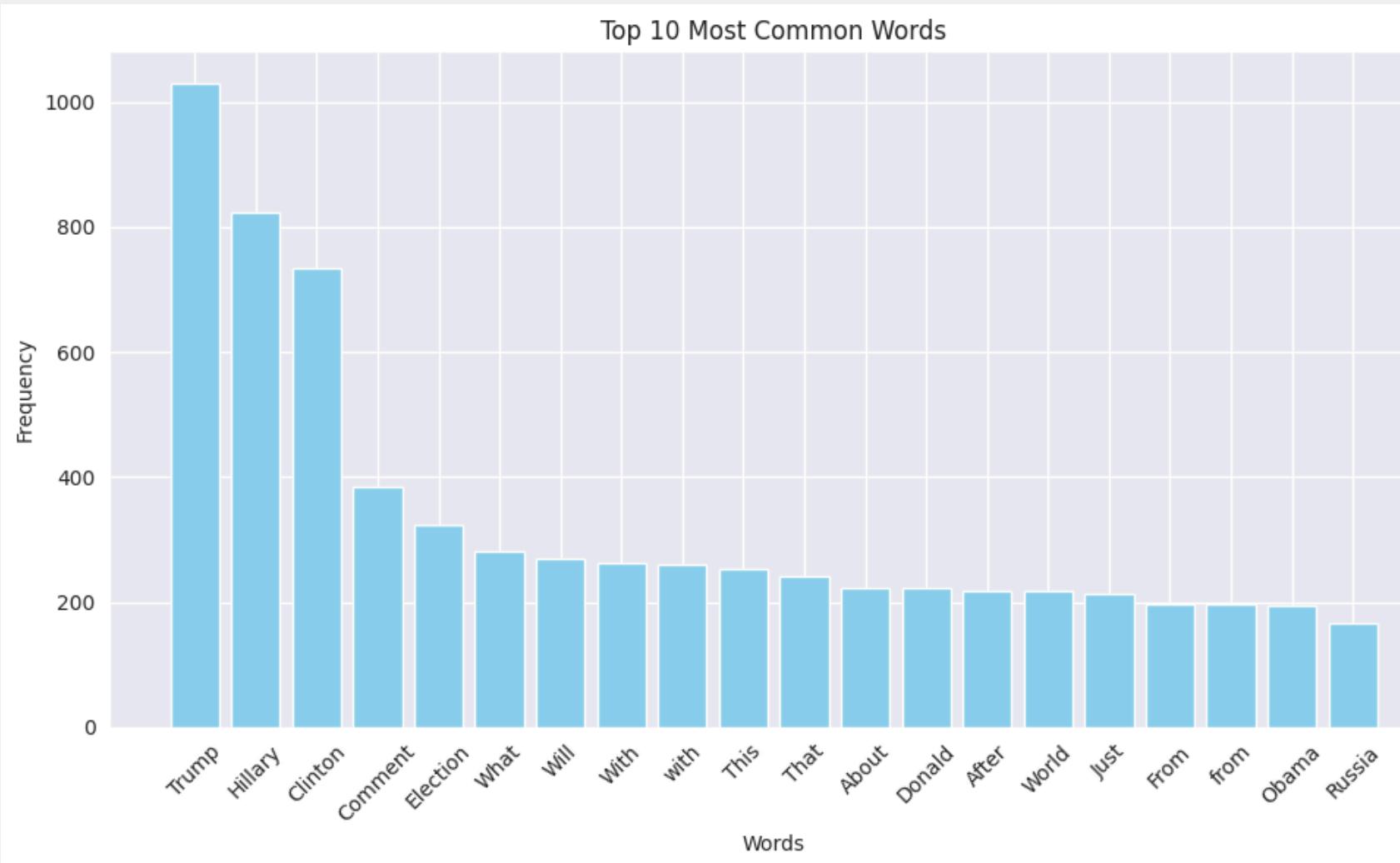
Số liệu về giả mạo, gian lận trên mạng xã hội



Mỹ là quốc gia được sử dụng dữ liệu nhiều nhất

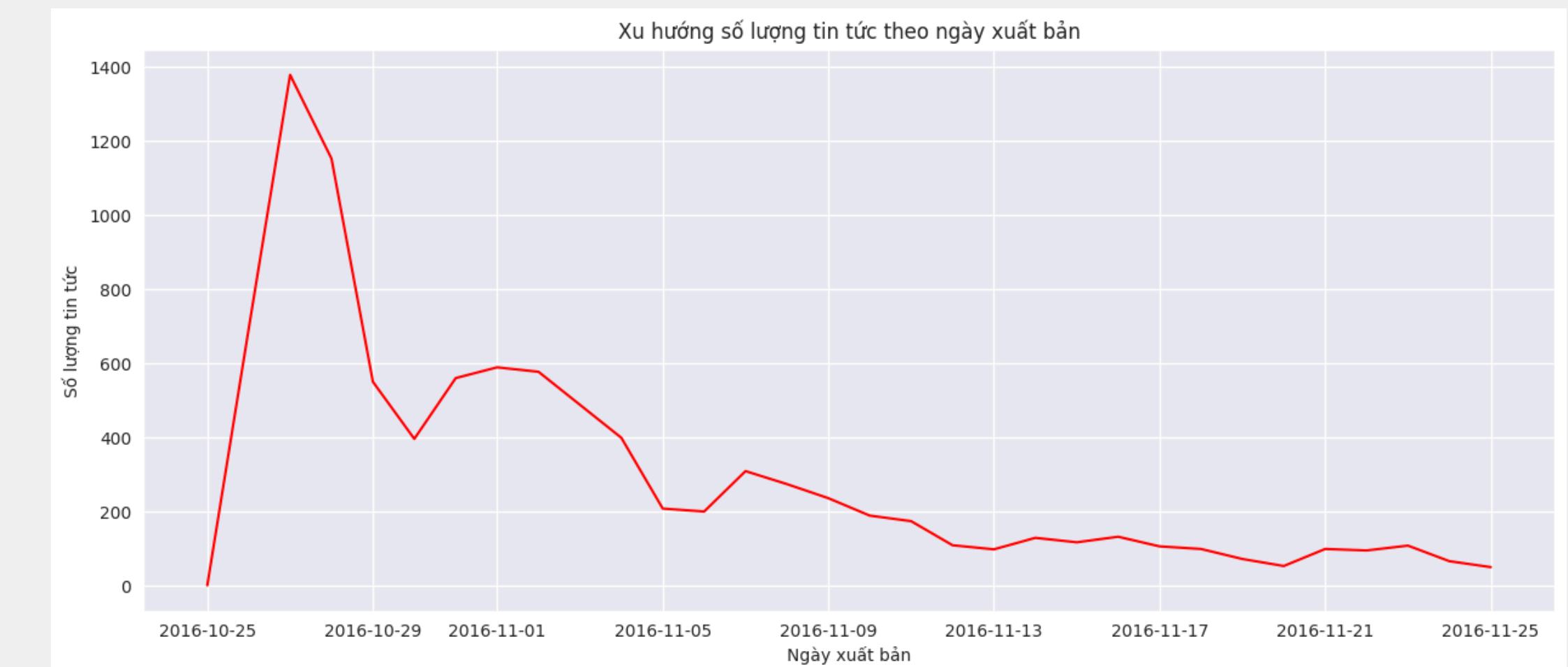
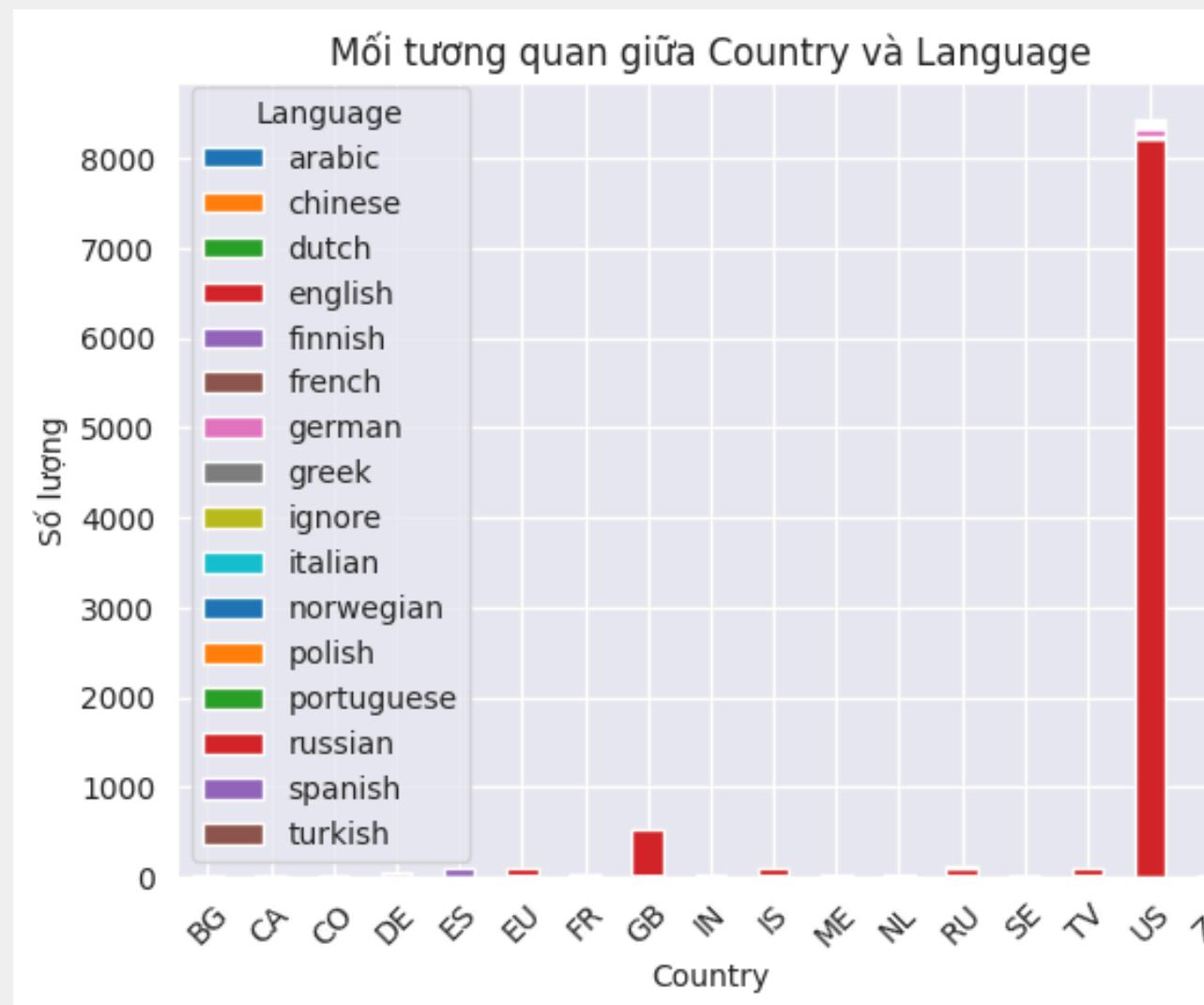


Số liệu về giả mạo, gian lận trên mạng xã hội



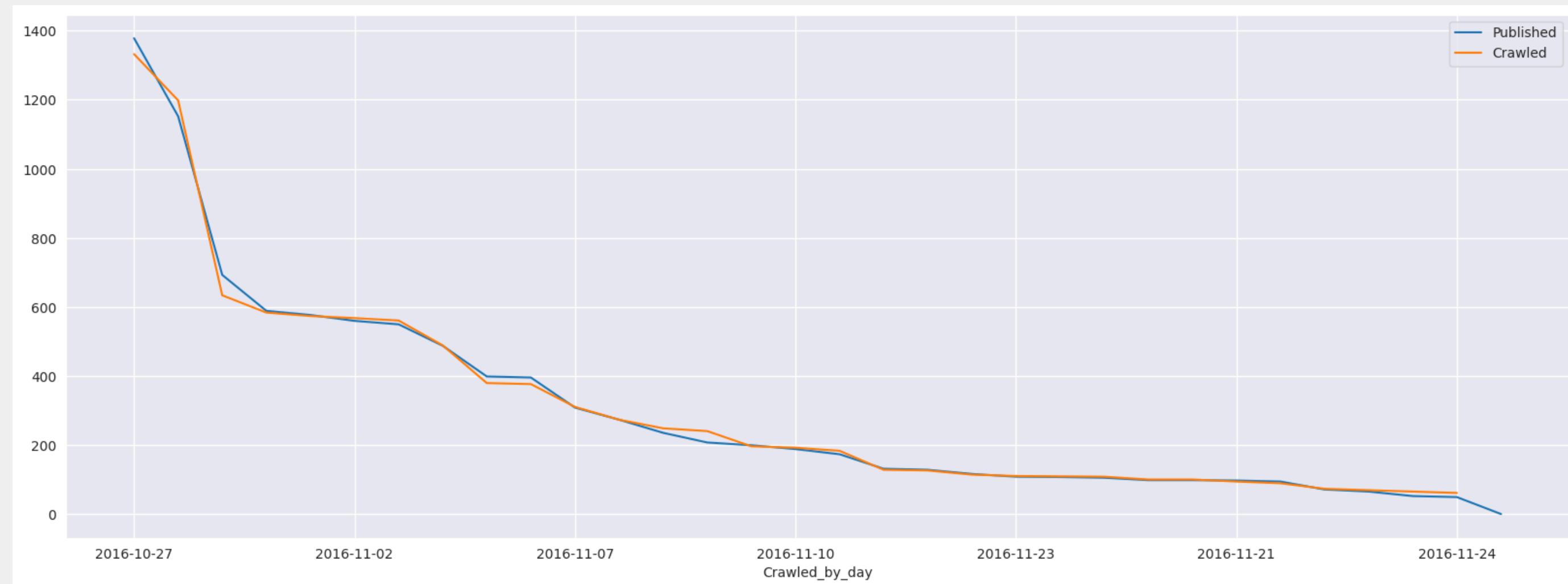
Từ khoá được tìm kiếm nhiều nhất trong dữ liệu

Số liệu về giả mạo, gian lận trên mạng xã hội



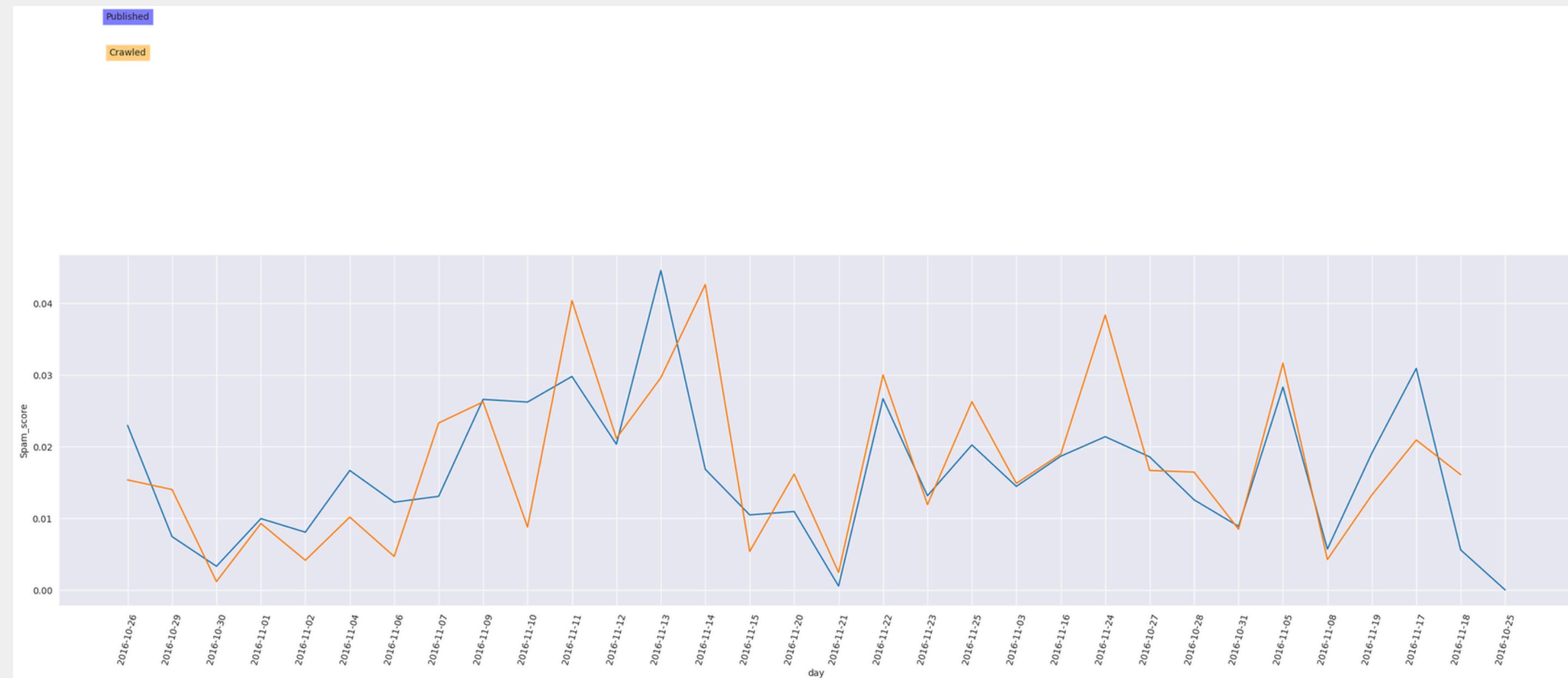
Số lượng xuất bản tin tức theo ngày

Số liệu về giả mạo, gian lận trên mạng xã hội



Số lượng xuất bản tin tức giảm dần theo thời
gian

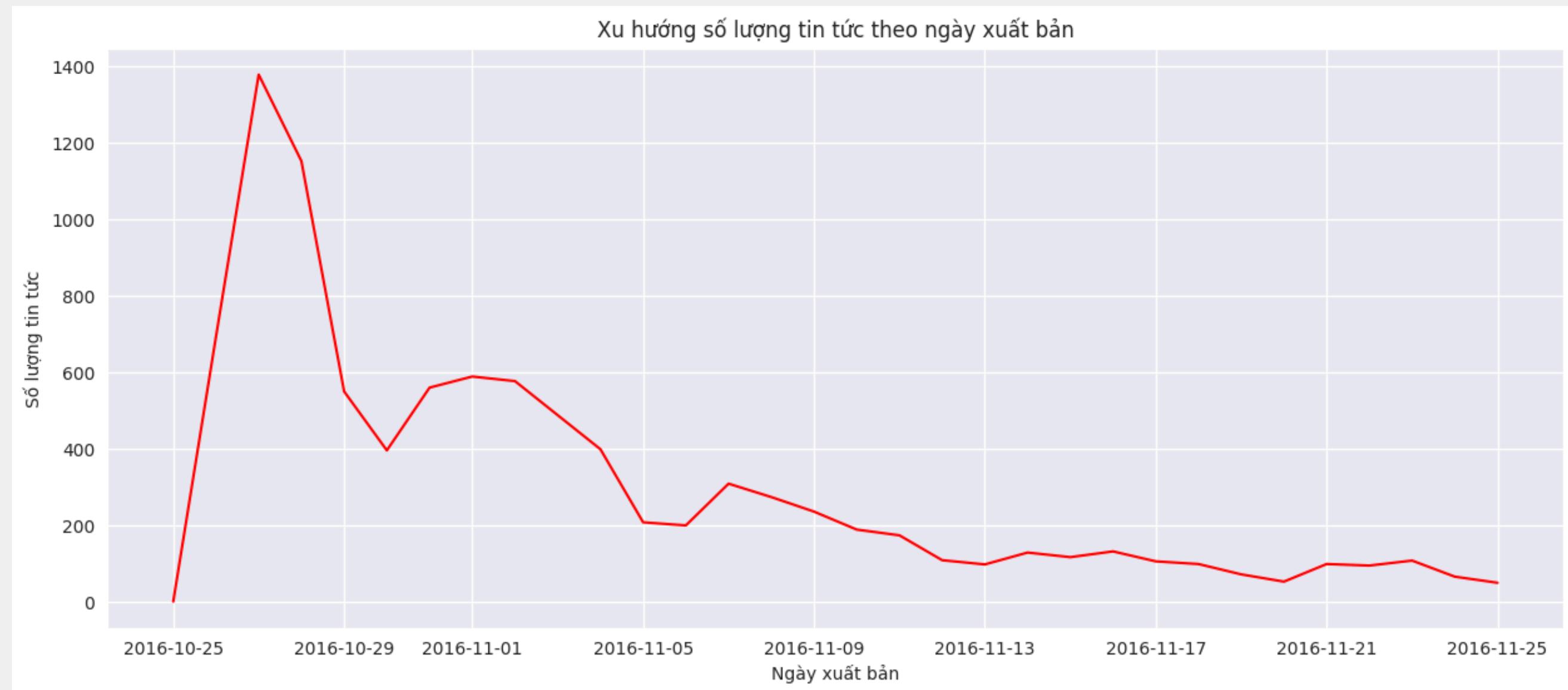
Số liệu về giả mạo, gian lận trên mạng xã hội



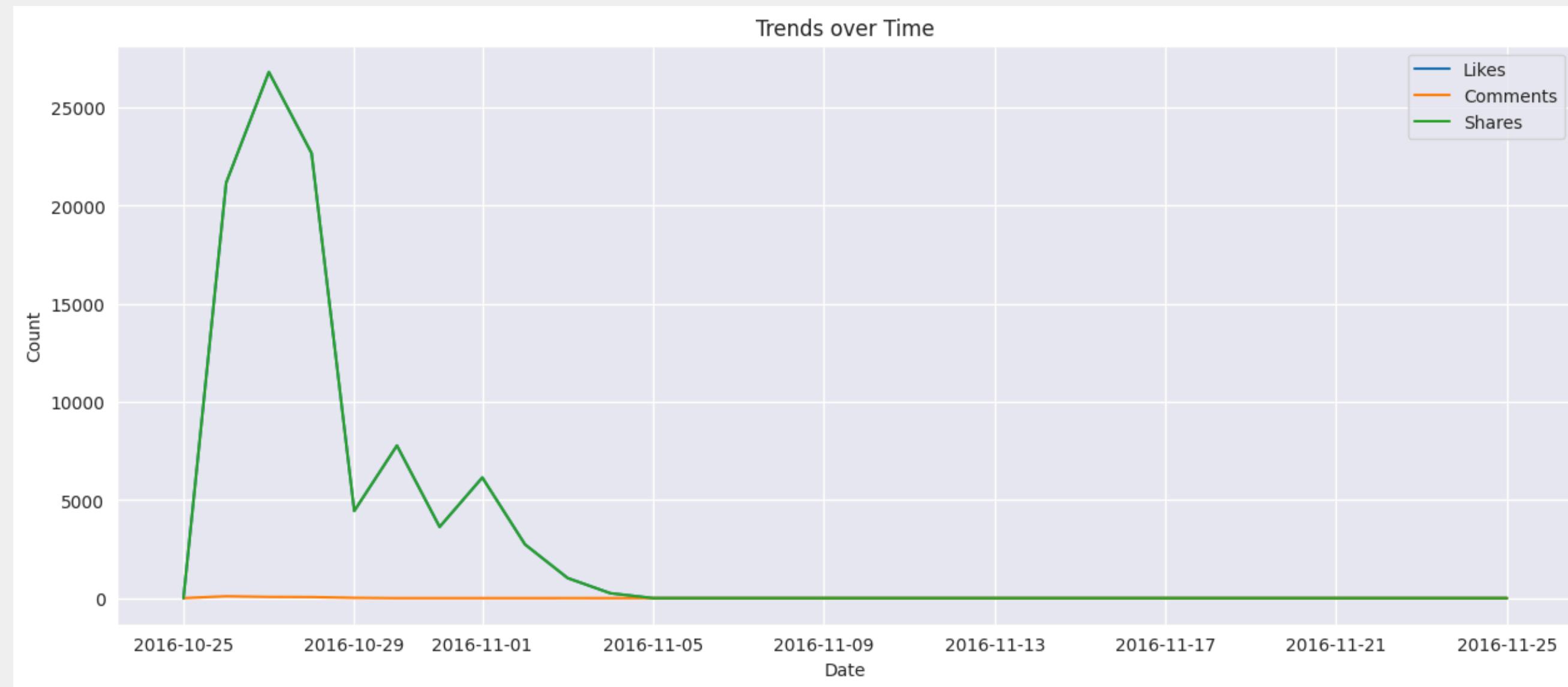
Số lượng ngày xuất bản và thu nhập dữ liệu từ
ngày 7-11 tới 23-11



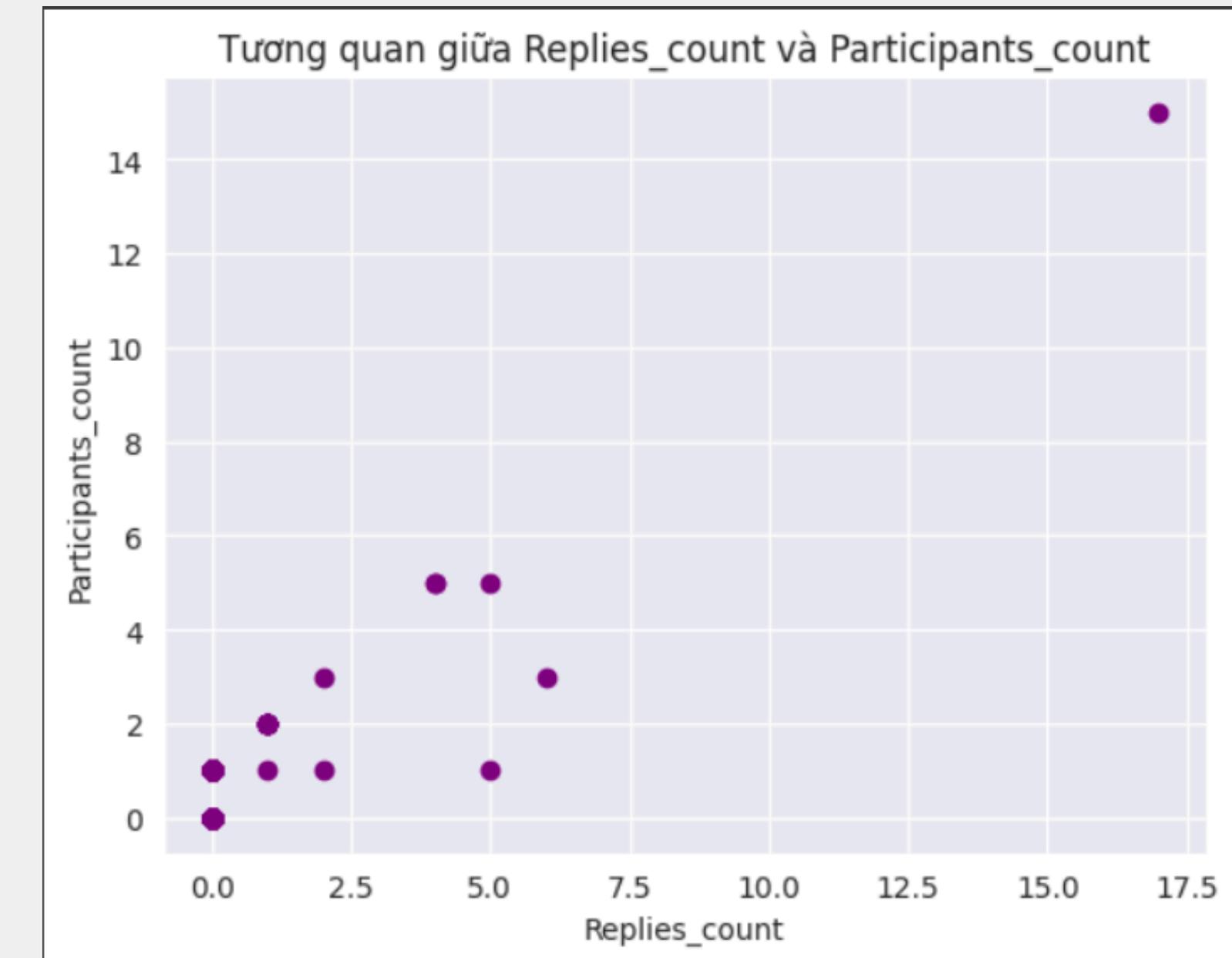
Số liệu về giả mạo, gian lận trên mạng xã hội



Số liệu về giả mạo, gian lận trên mạng xã hội

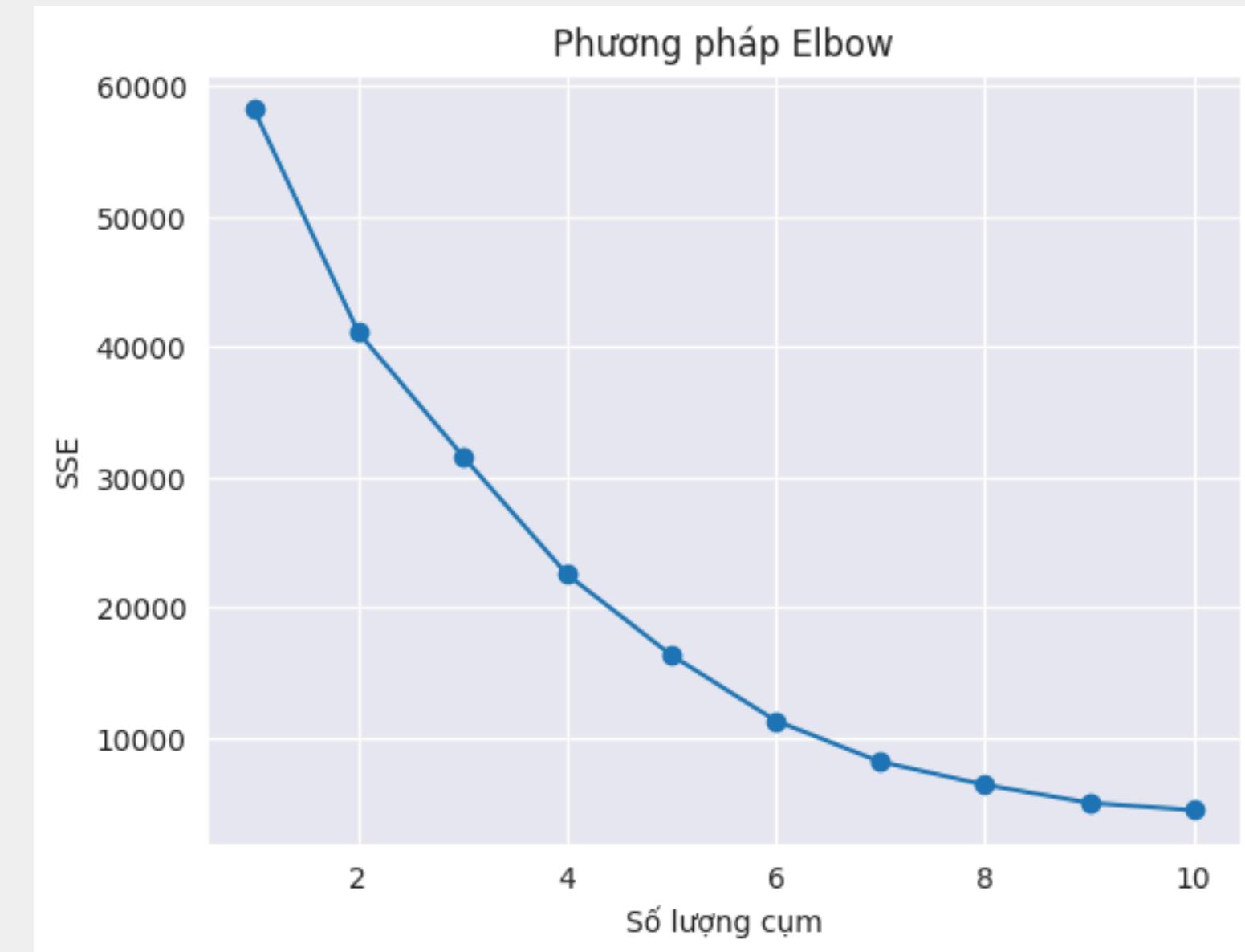


Số liệu về già mạo, gian lận trên mạng xã hội



Phân tích mối tương quan giữa Replies_count và
Participants_count

Số liệu về giả mạo, gian lận trên mạng xã hội



Elbow

Cách phòng tránh và nhận biết đâu là tin giả, tin thật

1

**Tìm hiểu từ
nhiều nguồn tin**

Sàng lọc và coi từ
nhiều nguồn để xác
định đâu là tin giả
đâu là tin thật

2

**Đọc kỹ các
thông tin để
nhìn nhận sự
đúng sai**

Tìm kiếm bài viết
khác từ chủ nhân
đưa tin để xác
nhận

3

**Không đưa ra
quan điểm khi
thấy các bài viết**

Các bài viết sai sự
thật nhầm gây ra
hiềm khích đối với
tất cả mọi người
nếu không tìm hiểu
kỹ thông tin

4

**Báo cáo ngay cho
các cơ quan chức
năng quản lý
thông tin mạng**

Hãy thông báo có các
cấp chính quyền về
những thông tin sai
lệch

5

**Không lan
truyền, đồn xa
khi chưa chắc
 chắn có phải
thông tin chính
xác**

Đọc và chỉ bản thân
bạn biết, không lan
truyền, đồn xa, để
tránh gây ra hậu
hoạ đến bản thân
mình

Cách phòng tránh và nhận biết đâu là tin giả, tin thật

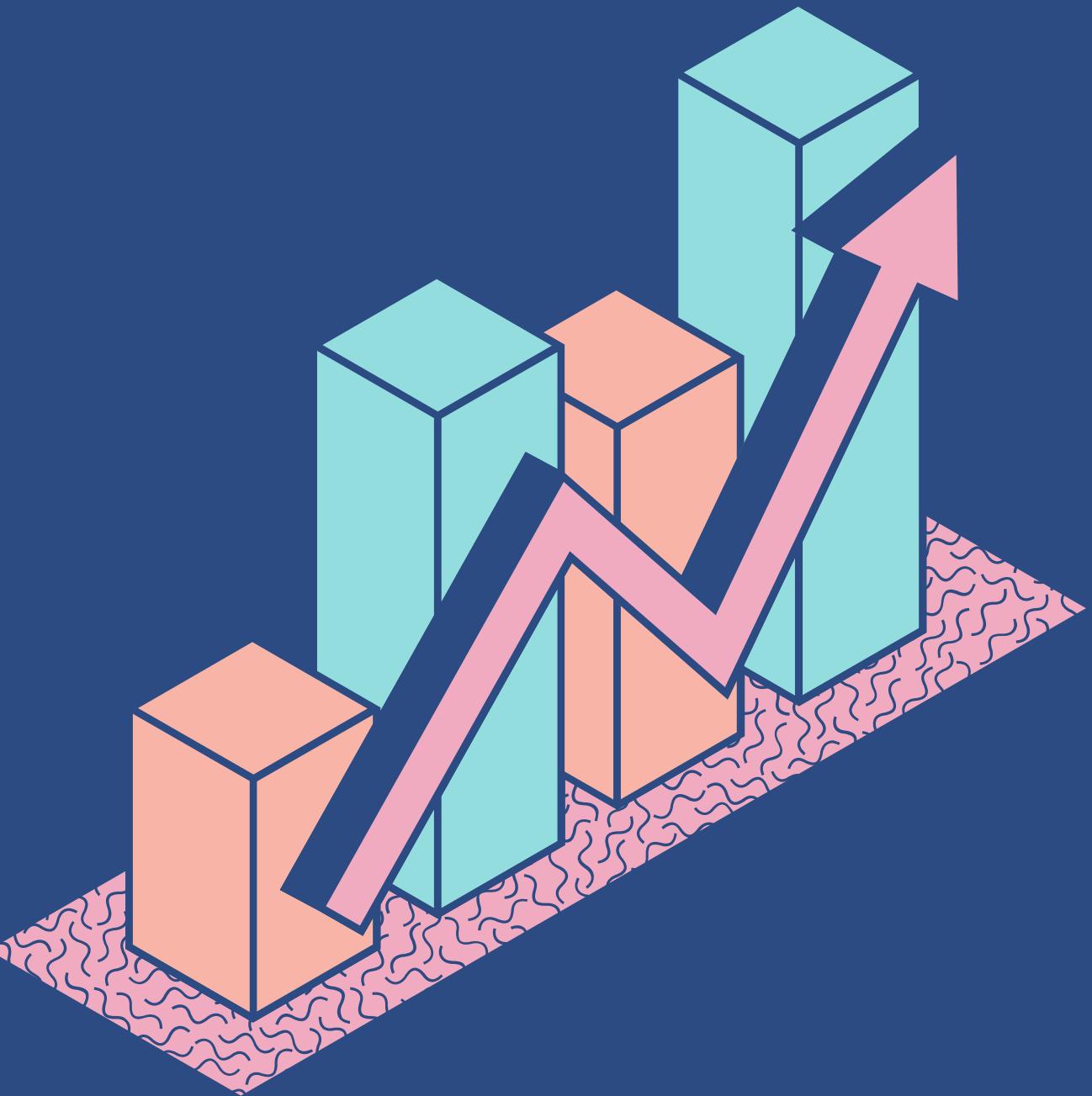


Hệ quả của việc tung tin đồn thất thiệt, giả mạo lên mạng xã hội

**Tung tin giả mạo trên mạng xã hội
bị phạt từ 10-20 triệu đồng**



Kết luận



Phải giữ vững tâm trí, không được lung lay trước các tin báo lá cải không rõ nguồn gốc

**Không lan truyền những
tin tức giả mạo khi chưa
được xác định rõ ràng**

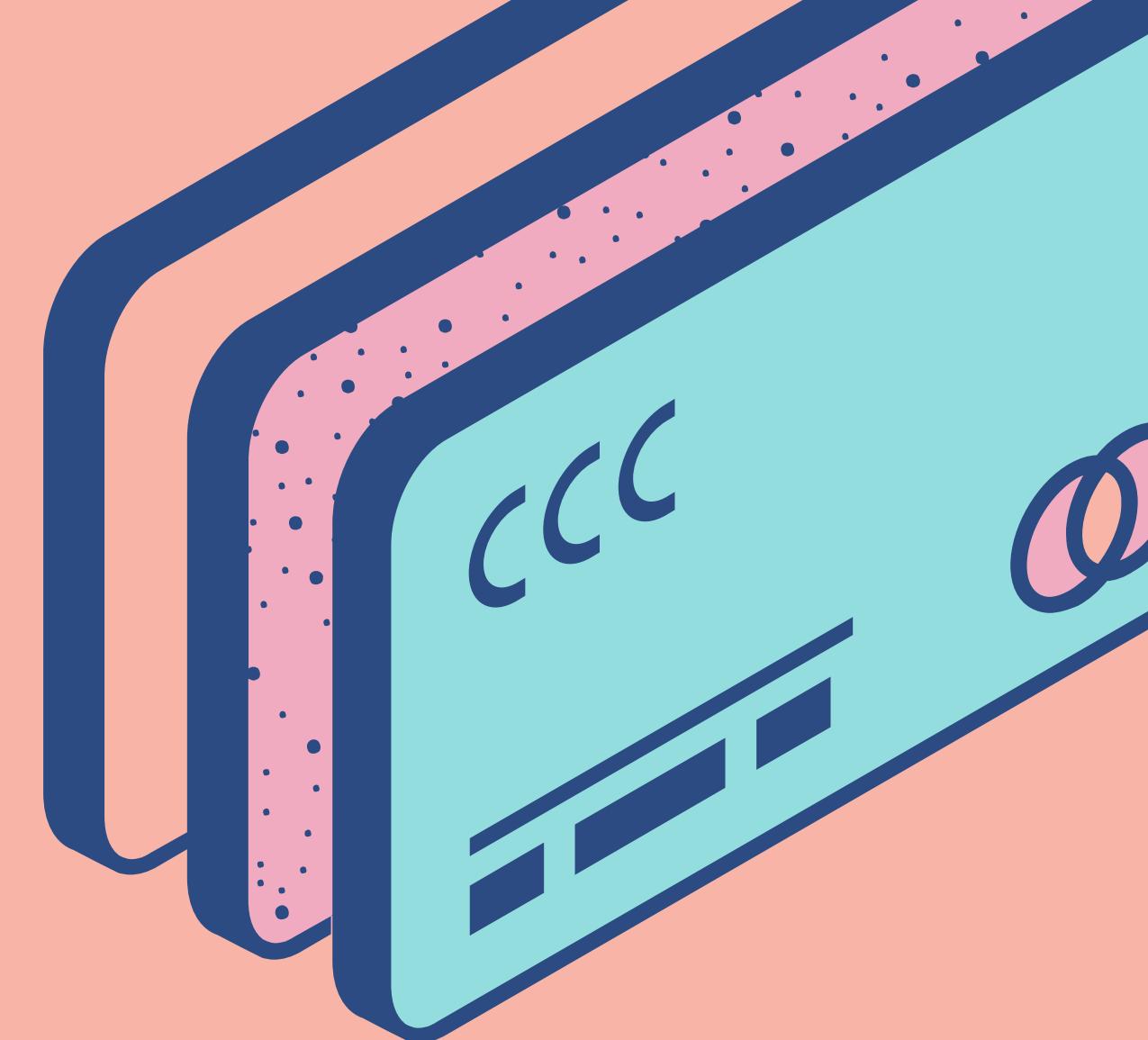
Mọi người phải nhận thức được
chính xác đâu là tin thật đâu
là tin giả khi thấy các bài báo
trên mạng xã hội, phải xác
định rõ nguồn đăng và tính
chính xác của bài viết.

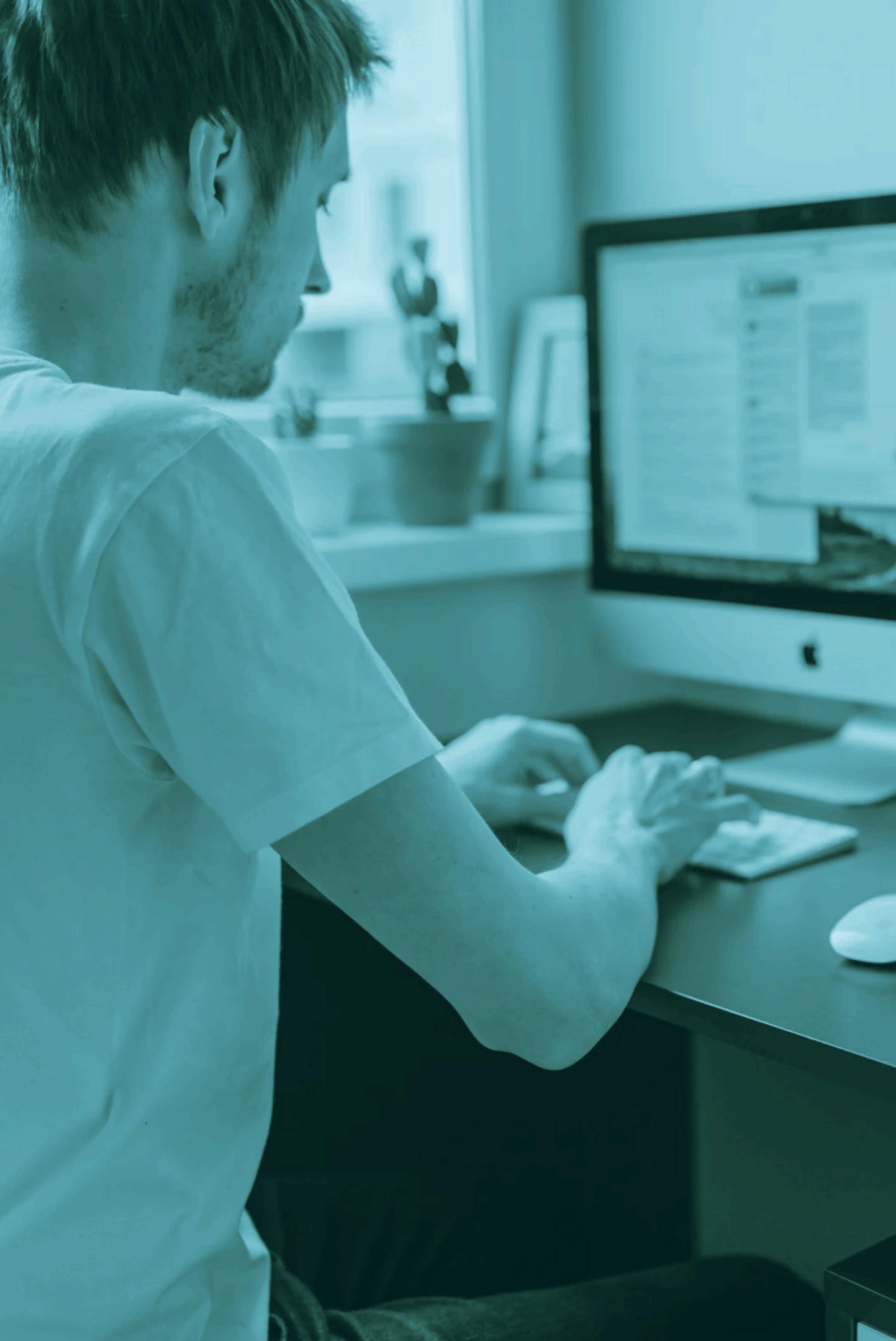
**Báo cáo cho các cấp
chính quyền có thẩm
định về các bài báo giả
mạo gây hiềm khích nội
bộ trong nước**

Việc thấy các bài báo đưa sai
thông tin từ các nguồn không
đáng tin cậy, mọi người có thể
thông báo ngay cho chính
quyền để họ xử lý các trường
hợp đưa thông tin sai sự thật.

**Tìm hiểu rõ ràng nguồn
đăng, không đôn đai để
tránh hậu hoạ sau này**

Nếu mọi người không tìm hiểu
kỹ các bài báo được đăng tại
nguồn không rõ ràng, có thể bị
chính quyền mời lên để quyết,
xử phạt hành chính.





Q&A



THANKS
FOR
LISTENING