AI Academy Vietnam

Final Project Python

TEAM = 【"VAN NAM DANG", "TOMMY NGUYEN", "LILLY VU"】

IDEAS

Topic：Multi-Dimensional Analysis: From DataFrame to the basic Machine Learning basic Models

Sourcecode: Github, class exercises, stack overlow, Thầy_Nam.

#SKlearn #Pandas #Keras #OpenAPI #NLTK #Matplotlib #NLP #Matplotly

NỘI DUNG CHÍNH

01 Ý TƯỞNG

02 THU THẬP DỮ LIỆU

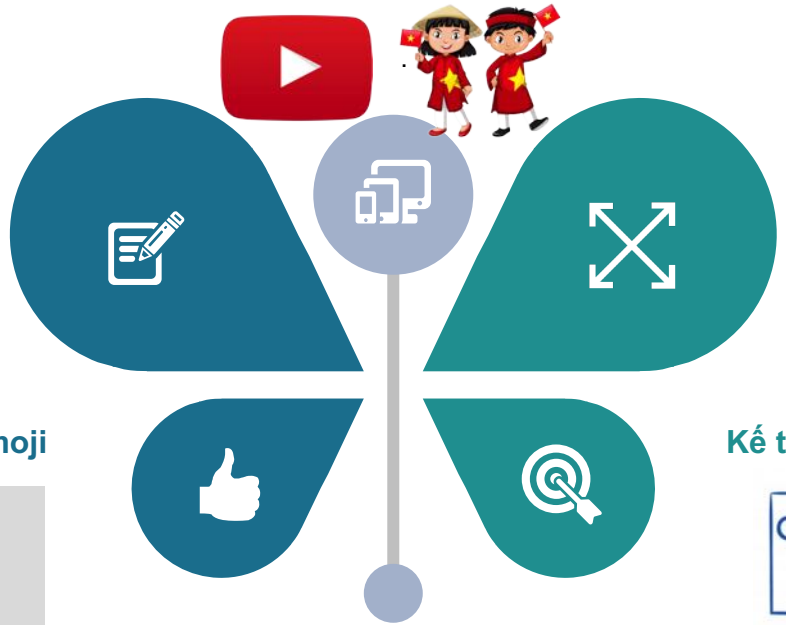
03 PHÂN TÍCH HỌC MÁY

04 DEMO

Ý TƯỞNG:

Người Việt thích clip âm nhạc Youtube?

Dữ liệu Tiếng Việt

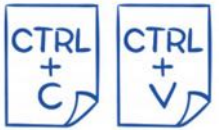


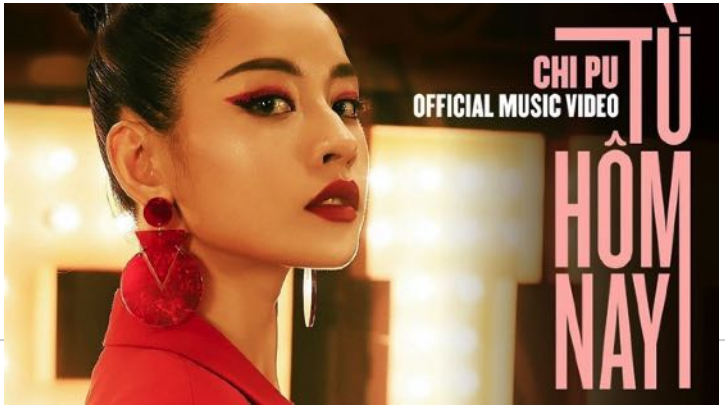
Phân tích kết hợp

Cảm xúc đa chiều

Phân tích cả Emoji

Kế thừa và cải tiến





TA ĐANG LẠC NƠI NÀO ...

Cào Dữ Liệu

HDYLT Github

Youtube3 API

Cào 5-40mins

Lạc Trôi – 45K
Sơn Tùng MTP

Từ Hôm Qua – 47K
ChiPu

HãyTraoChoAnh – 780K
Sơn Tùng MTP

Lạc Trôi


```
(45100, 6)
Done process at :  2020-11-09 10:13:31
```

	Comment	Comment ID	Reply Count	Like Count	Updated At	Viewer Rating
0	LẠC TRÔI (TRIPLE D REMIX) I 360 DEGREE MV I S...	UgxkMePlfTxtvesUSJl4AaABAg	446	27410	none	2017-09-08T13:16:21Z\r
1	Drift away\n\nDrifting.....\nIs that the world...	UgxBx7XzRki4L7GmssB4AaABAg	0	2	none	2020-11-06T15:57:15Z\r
2	231M ai còn nghe ko	UgzEofSxjREzcicg7fl4AaABAg	0	1	none	2020-11-06T15:50:36Z\r
3	Bn tôi nay có tâm sự à	Ugzfup8Gq6Y_Z9XVKxB4AaABAg	1	1	none	2020-11-06T15:40:24Z\r
4	Vấn vuong	Ugz8FnabTqhshShoBT94AaABAg	0	1	none	2020-11-06T15:37:33Z\r
...	...	...	...	...	...	...
95	em tên gia bảo	UgzCjCZ5LmqnZPnhgKJ4AaABAg	0	3	none	2020-11-02T13:41:34Z\r
96	em thích anh	Ugxp0KDJXr0jhm7u7RZ4AaABAg	0	2	none	2020-11-02T13:41:21Z\r
97	Sau 3 năm thì mn thấy bài này còn hay ko \nTôi...	Ugz6f5NPvs-eS8M7Kyx4AaABAg	0	3	none	2020-11-02T12:19:19Z\r
98	Hay 😍😍😍😍	UgwhtcGDCww8DVyi7Np4AaABAg	0	3	none	2020-11-02T12:15:30Z\r
99	Ông hoàng v-pop là đây <3	UgxovpqJW4bM_2Ueo9x4AaABAg	0	3	none	2020-11-02T11:17:13Z\r

100 rows × 6 columns

```
time: 147 ms
```

Lạc Trôi

predicted_lang	Reply Count	Like Count	text_length	predicted_conf
Vietnamese	3.958110	202.119407	5.599915	0.175774
Malay (macrolanguage)	3.175426	142.723801	3.018880	0.079394
Thai	5.439510	83.014107	3.262488	0.311761
Swedish	2.269029	69.303782	1.996007	0.141440
Lao	3.255124	64.222011	1.024695	0.307640
En	2.299175	60.701571	8.362197	0.256718
Azerbaijani	2.121320	52.325902	15.556349	0.294864
Central Khmer	2.211540	51.961699	1.489356	0.210717
Indonesian	1.248529	49.711093	3.381998	0.295821
Occitan (post 1500)	2.274388	39.933484	1.961326	0.157698
Chinese	3.052569	36.414902	7.874286	0.300635
Portuguese	1.308824	31.150676	3.552344	0.265567
Danish	0.733493	29.857948	6.768952	0.337665
Telugu	1.154701	29.737743	8.962886	0.067506
Romansh	2.121320	26.870058	4.242641	0.023335
Western Frisian	1.113704	23.772627	2.489305	0.252562
Modern Greek (1453-)	1.567021	23.649289	1.229273	0.166086
Afrikaans	1.000000	22.186708	3.366502	0.201571
Turkish	0.550033	21.751395	2.431675	0.276987
Russian	0.784777	19.671101	2.316480	0.202308

time: 79.7 ms

Từ Hôm Nay

predicted_lang	Reply Count	Like Count	text_length	predicted_conf
Finnish	4.478343	59.319672	2.255712	0.120878
Vietnamese	2.567057	29.818166	31.842629	0.145133
Piemontese	4.734250	24.615359	4.248033	0.171061
Serbo-Croatian	2.556707	24.442236	1.911993	0.221637
German	0.981069	21.611511	12.148200	0.242105
En	1.462840	14.119374	23.471414	0.286627
Sinhala	0.288675	12.901433	3.785939	0.173556
Scottish Gaelic	0.332106	11.003008	3.733788	0.139063
Lombard	3.027840	9.593820	4.072476	0.118925
Romanian	2.711478	9.424179	17.250139	0.168980
Panjabi	0.000000	9.324400	0.843274	0.069778
Italian	0.144139	9.310704	10.362946	0.240906
Spanish	0.434274	8.910883	3.210909	0.285034
Catalan	0.728740	8.399254	2.290067	0.203309
Croatian	1.239691	8.307466	3.155809	0.186955
Welsh	1.178314	7.895198	15.284616	0.154976
Manx	1.100016	7.744916	2.144797	0.178420
Egyptian Arabic	0.000000	7.609205	1.788854	0.130408
Slovak	0.928477	6.397929	1.975839	0.111137
Gujarati	3.889087	4.949747	0.755929	0.070567

time: 76.7 ms

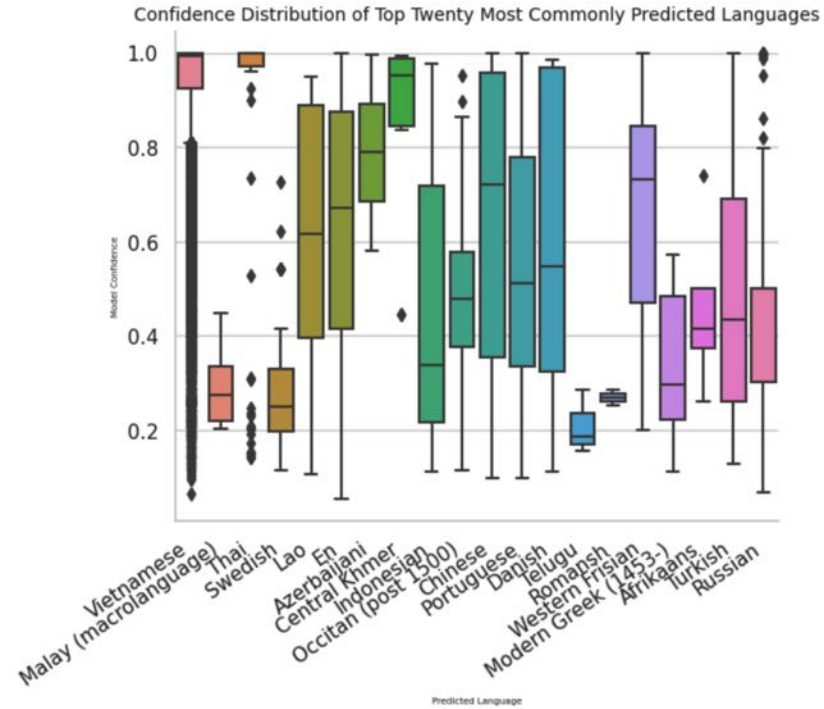
Lạc Trôi

592	Italian	phiuu
593	Vietnamese	vẫn thích hợp
595	Catalan	rấtlàhay
597	Norwegian	lạc
599	Vietnamese	giờ có còn ai nghe nữa k
600	Vietnamese	lâu lâu nghe có hứng
603	Vietnamese	lạc trôi
604	Portuguese	tem nghe 16/10/2020

Từ Hôm Nay

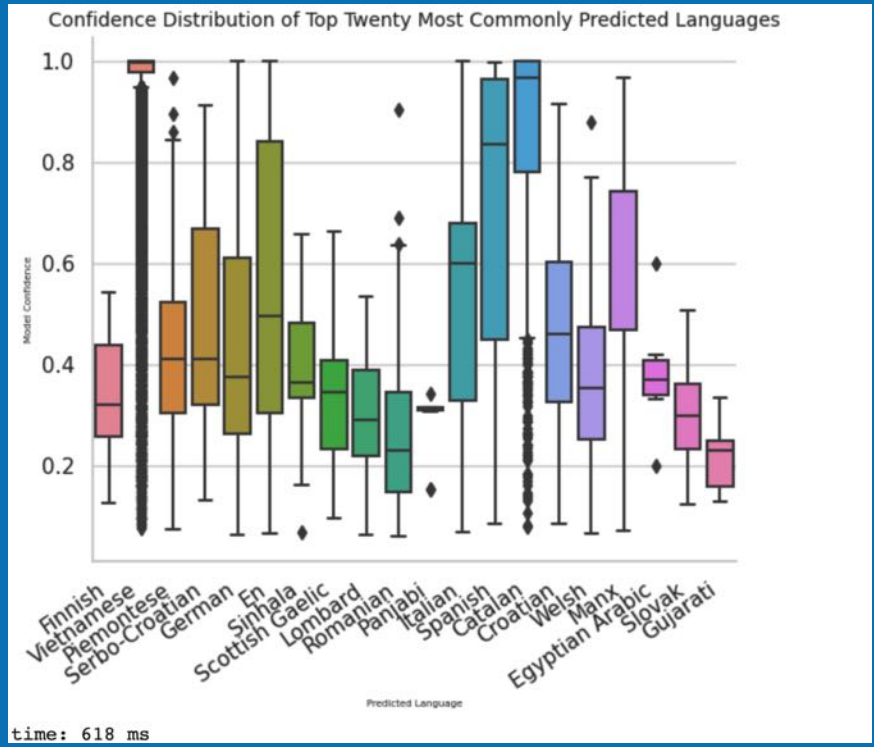
69	Vietnamese	2020 ai còn xem thì cho like nha
71	Vietnamese	tôi vẫn ở đây vẫn yêu quý cô ấy <3 nghe bản re...
72	Vietnamese	ghét vl. nghe k có 1 tý cảm tình. nhảm nhí
73	Vietnamese	tôi từ binz hát từ hôm nay quay lại, chắc ko a...
74	Vietnamese	chi pu làm bài này dở
75	Spanish	hay vcl
76	German	chipu
77	Vietnamese	hay quá

Lạc Trôi



time: 768 ms

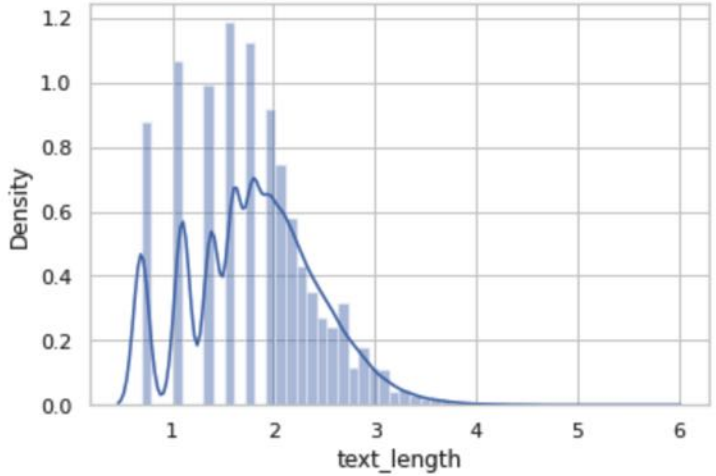
Từ Hôm Nay



time: 618 ms

Lạc Trôi

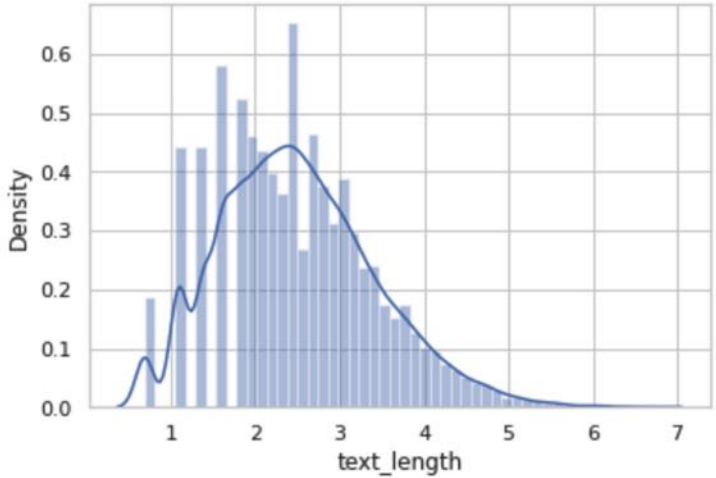
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7fa9aea860>
```




```
time: 657 ms
```

Từ Hôm Nay

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f91b66b1748>
```




```
time: 619 ms
```

Lạc Trôi

	emoji_count
count	584.000000

	emoji	emoji_count	sentiment	text
3	□	2219	0.0	
18	👍	1452	0.0	:thumbs_up:
432	🌑	630	0.0	:new_moon:
433	🌕	576	0.0	:full_moon:
161	💪	567	0.0	:flexed_biceps:
17	👇	505	0.0	:backhand_index_pointing_down:
109	🙌	424	0.0	:raising_hands:
146	👏	388	0.0	:clapping_hands:
170	🎶	246	0.0	:musical_notes:
158	🌹	233	0.0	:rose:
252	😋	227	0.0	:face_savoring_food:
40	🌻	209	0.0	:sunflower:
141	🤾	209	0.0	:woman_dancing:
31	🇳	176	0.0	:regional_indicator_symbol_letter_n:
167	🤩	172	0.0	:star-struck:

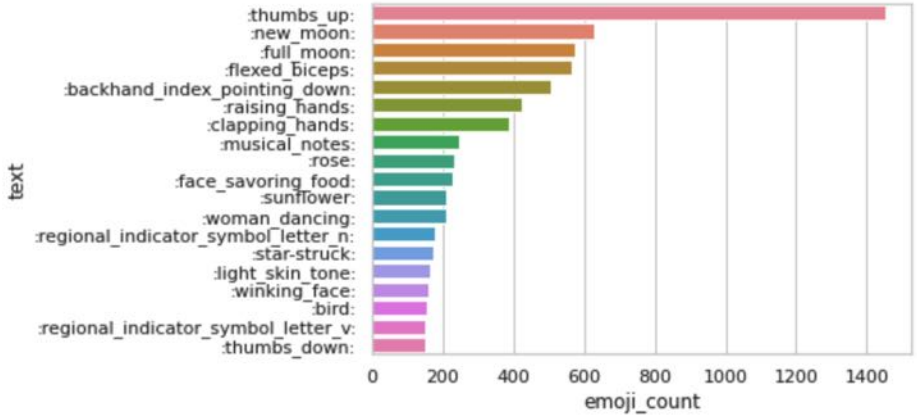
Từ Hôm Nay

	emoji_count
count	390.000000

	emoji	emoji_count	sentiment	text
55	👎	1017	0.0	:thumbs_down:
15	👍	1006	0.0	:thumbs_up:
2	□	875	0.0	
93	😏	387	0.0	:smirking_face:
66	👏	297	0.0	:clapping_hands:
53	😑	268	0.0	:expressionless_face:
51	🏻	224	0.0	:light_skin_tone:
226	💪	207	0.0	:flexed_biceps:
58	😉	173	0.0	:winking_face:
85	😒	156	0.0	:unamused_face:
74	🤔	153	0.0	:thinking_face:
79	😡	130	0.0	:pouting_face:
183	☁	118	0.0	:cloud:

Lạc Trôi

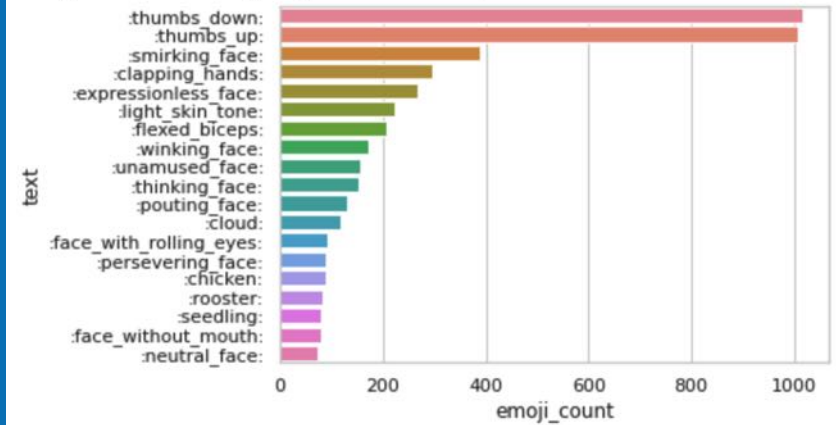
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7f95e74ac8>
```




```
time: 381 ms
```

Từ Hôm Nay

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f91b31e0588>
```




```
time: 314 ms
```

Phân Bổ của Từ Ngữ #LạcTrôi

time: 54.7 ms

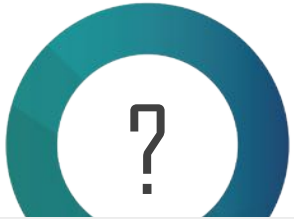
Phân Bổ của Từ Ngữ #Từ Hôm Nay

time: 73.5 ms

WordCloud of #LạcTrôi

WordCloud of #TừHômNay

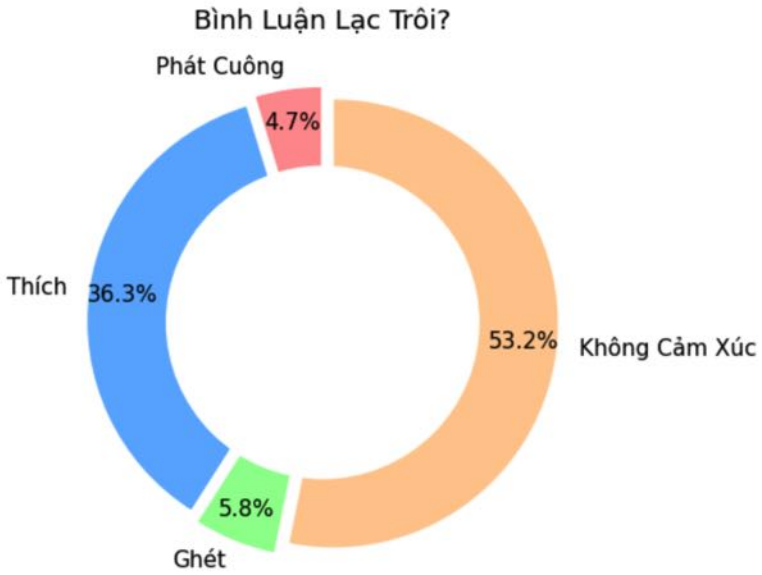
Từ Điển




```python
#Xây dựng hàm chuẩn hóa dữ liệu tiếng việt
def normalize_text(text):

    #Remove các ký tự kéo dài: vd: đẹpppppppp
    text = re.sub(r'([A-Z])\1+', lambda m: m.group(1).upper(), text, flags=re.IGNORECASE)

    # Chuyển thành chữ thường
    text = text.lower()

    #Chuẩn hóa tiếng Việt, xử lý emoj, chuẩn hóa tiếng Anh, thuật ngữ
    replace_list = {
        'òa': 'oà', 'óa': 'oá', 'ỏa': 'oả', 'õa': 'oã', 'ọa': 'oạ', 'òe': 'oè', 'óe': 'oé','ỏe': 'oẻ',
        'õe': 'oẽ', 'ọe': 'oẹ', 'ùy': 'uỳ', 'úy': 'uý', 'ủy': 'uỷ', 'ũy': 'uỹ','ụy': 'uỵ', 'uả': 'ủa',
        'ả': 'ả', 'ố': 'ố', 'u´': 'ố','ỗ': 'ỗ', 'ồ': 'ồ', 'ổ': 'ổ', 'á': 'á', 'ã': 'ã', 'à': 'à',
        'ầ': 'ầ', 'ỏ': 'ỏ', 'ề': 'ề','ễ': 'ễ', 'ắ': 'ắ', 'ủ': 'ủ', 'ế': 'ế', 'ở': 'ở', 'ỉ': 'ỉ',
        'ẻ': 'ẻ', 'àk': u' à ','a`': 'à', 'i`': 'ì', 'ă´': 'ắ','ử': 'ử', 'e˜': 'ẽ', 'y˜': 'ỹ', 'a´': 'á',
        #Quy các icon về 2 loại emoj: Tích cực hoặc tiêu cực
        "🔞": "negative", "☠️": "positive", "💃": "positive",'👍': ' positive ', '👍': ' positive ',
        "💄": "positive", "💎": "positive", "💩": "positive","🙂": "negative", "😱": "negative", "🐱": "positive",
        "😿": "negative", "🚫": "negative",  "😡": "negative","🦋": "positive", "❤️": "positive",'🐶':' positive ',
        '👎': ' negative ', '😒': ' negative ','✨': ' positive ', '❣': ' positive ','☀': ' positive ',
        '♥': ' positive ', '🤩': ' positive ', 'like': ' positive ', '💌': ' positive ',
        '🤣': ' positive ', '🖤': ' positive ', '☺️': ' positive ', ':(': ' negative ', '😢': ' negative ',
        '❤': ' positive ', '😍': ' positive ', '😘': ' positive ', '😪': ' negative ', '😊': ' positive ',
        '?': ' ? ', '😁': ' positive ', '💖': ' positive ', '😉': ' negative ', '😭': ' negative ',
        '💯': ' positive ', '💗': ' positive ', '♡': ' positive ', '💜': ' positive ', '🤗': ' positive ',
        '^^': ' positive ', '😨': ' negative ', '☺': ' positive ', '💋': ' positive ', '👌': ' positive ',
        '😖': ' negative ', '😀': ' positive ', ':((': ' negative ', '😠': ' negative ', '😠': ' negative ',
        '😏': ' negative ', '🙂': ' positive ', '😔': ' negative ', '😜': ' positive ', '😄': ' positive ',
        '😶': ' positive ', '😤': ' negative ', '😎': ' positive ', '😆': ' positive ', '💚': ' positive ',
```

Lạc Trôi

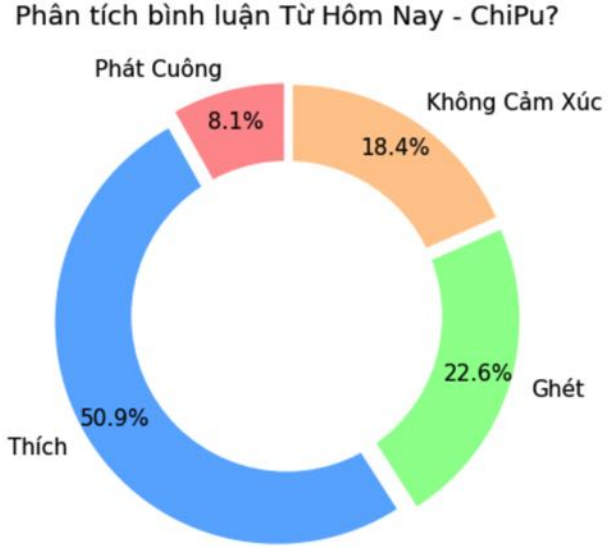
```
---------------- TỔNG HỢP CÁC CON SỐ -------------
Bình luận tích cực:  12705  Vs  Bình luận tiêu cực:  1813
Bình luận không có gì :  16489
Tổng số Bình luận:  31007
```




```
time: 145 ms
```

Từ Hôm Nay

```
---------------- TỔNG HỢP CÁC CON SỐ -------------
Bình luận tích cực:  26682  Vs  Bình luận tiêu cực:  10195
Bình luận không có gì :  8305
Tổng số Bình luận:  45182
```

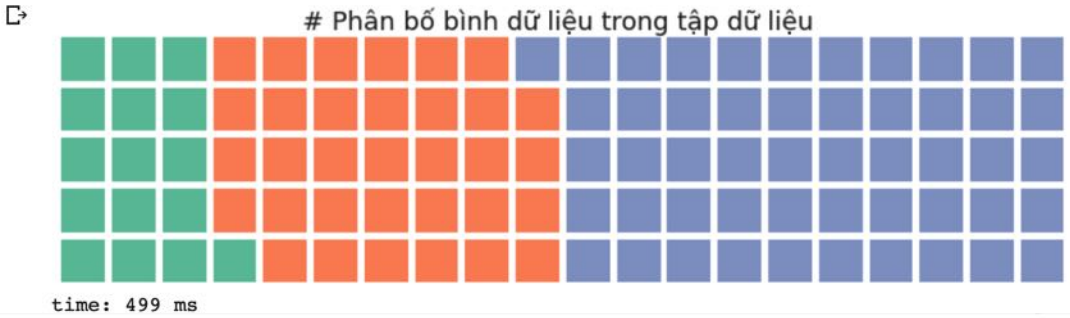



```
time: 149 ms
```

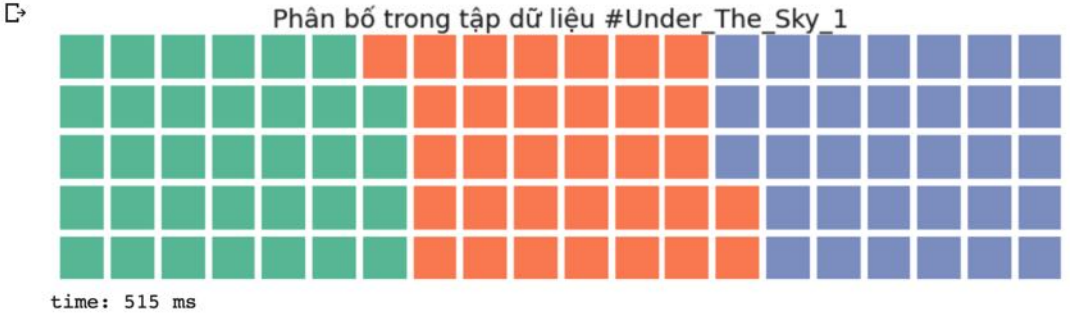
Tạo Tập Dữ Liệu

	clean_comment			
	count	unique	top	freq
label				
-1.0	12000	11497	thảm_hoạ negative	55
0.0	12000	9774	lạc trôi	153
1.0	12000	10643	hay quá positive	180

time: 81.2 ms



time: 499 ms



time: 515 ms

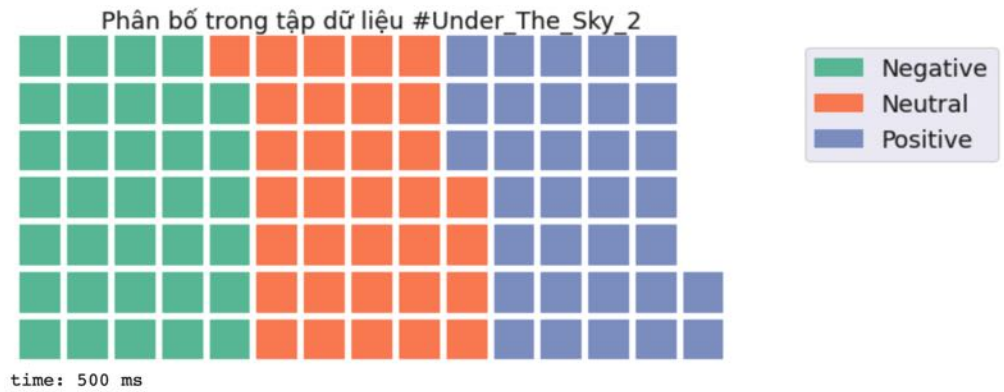
Tạo Tập Dữ Liệu

	clean_comment	label
4	nghiện nghiện nghiện say say say vấn_vương pos...	1.0
5	star bài này chil nhất n ae positive positive ...	1.0
6	tôi thích bài hát này positive positive	1.0
7	được đấy positive	1.0
8	lâu_lâu nghe lại hay ghê ý positive	1.0
...	...	...

	clean_comment	label
4	nghien nghien nghien say say say van_vuong pos...	1.0
5	star bai nay chil nhat n ae positive positive ...	1.0
6	toi thich bai hat nay positive positive	1.0
7	duoc day positive	1.0
8	lau_lau nghe lai hay ghe y positive	1.0
...	...	...

	clean_comment			
	count	unique	top	freq
label				
-1.0	24000	11469	tham_hoa negative	110
0.0	24000	9647	lac troi	342
1.0	24000	10573	hay qua positive	430

time: 79.1 ms



time: 500 ms

Chọn Mô Hình

#Under_The_Sky_1

	model_name	accuracy_score	precision_score	recall_score	f1_score
0	Super Vector Machine	0.973148	0.974213	0.973148	0.973235
1	Random Forest	0.963241	0.964250	0.963241	0.963326
2	Stochastic Gradient Descent	0.958704	0.961573	0.958704	0.958978
3	Decsision Tree	0.931296	0.931457	0.931296	0.931354
4	AdaBoost	0.930370	0.931244	0.930370	0.930518
5	K Nearest Neighbor	0.869722	0.883094	0.869722	0.870432
6	Gaussian Naive Bayes	0.609352	0.674126	0.609352	0.569422
7	Dummy	0.332315	0.332308	0.332315	0.332309

time: 38 ms

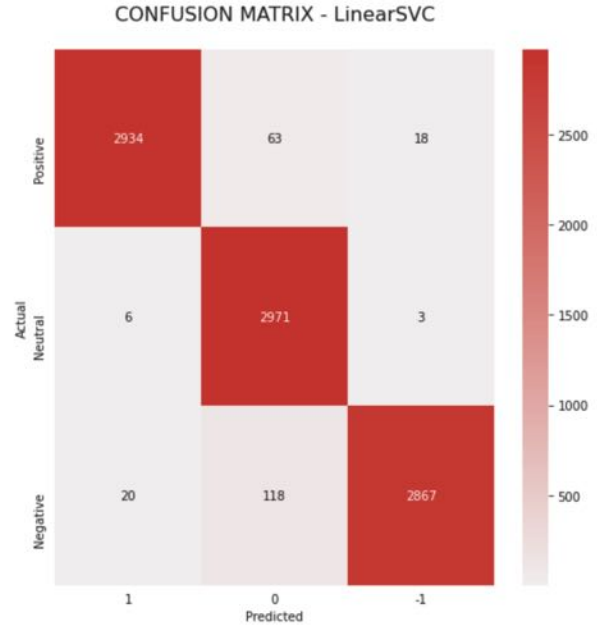
Áp Dụng Mô Hình SVM

#Under_The_Sky_1

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
1.0	0.99	0.97	0.98	3015
0.0	0.94	1.00	0.97	2980
-1.0	0.99	0.95	0.97	3005
accuracy			0.97	9000
macro avg	0.98	0.97	0.97	9000
weighted avg	0.98	0.97	0.97	9000

time: 792 ms



time: 264 ms

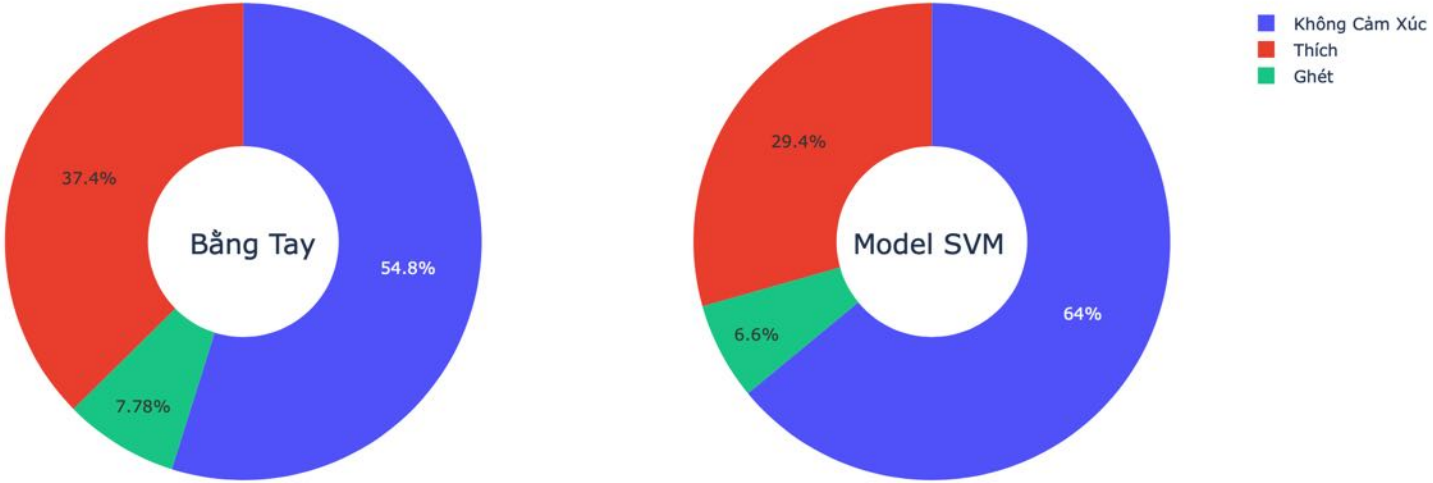
Demo Mẫu

#Under_The_Sky_1

```
 Áp Dụng Thử Nha:
----------------------------------------------------------
 Ví dụ :
Bình luận - Comment :  thks tks ! Thầy Nam hát hayyyyyyyy quá ❤️
----------------------------------------------------------
Text sau khi xử lý  :  cám ơn cám ơn thầy nam hát hay quá positive  positive positive positive
----------------------------------------------------------
Mô hình dự đoán     :  ['1.0']
time: 11.5 ms

----------------------------------------------------------
 Ví dụ :
Bình luận - Comment :   Mình thích hát nhưng giọng mình dở quá 😨
----------------------------------------------------------
Text sau khi xử lý  :  mình thích hát nhưng giọng mình dở quá negative positive negative
----------------------------------------------------------
Mô hình dự đoán     :  ['-1.0']
time: 12.5 ms
```

#HTCA

So sánh mô hình Phân Lớp Học Máy #HãyTraoChoAnh



time: 13.2 ms

#Happy20/11

AI Academy Vietnam