

MỤC TIÊU BÀI TOÁN

1

Phân loại người dùng theo
khu vực địa lý

2

Nhận dạng người
dùng liên quan tới
web3

3

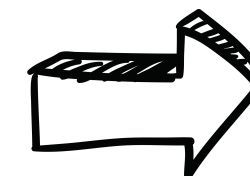
Phân loại người dùng theo
dự án quan tâm

THỰC HIỆN

Data Crawling

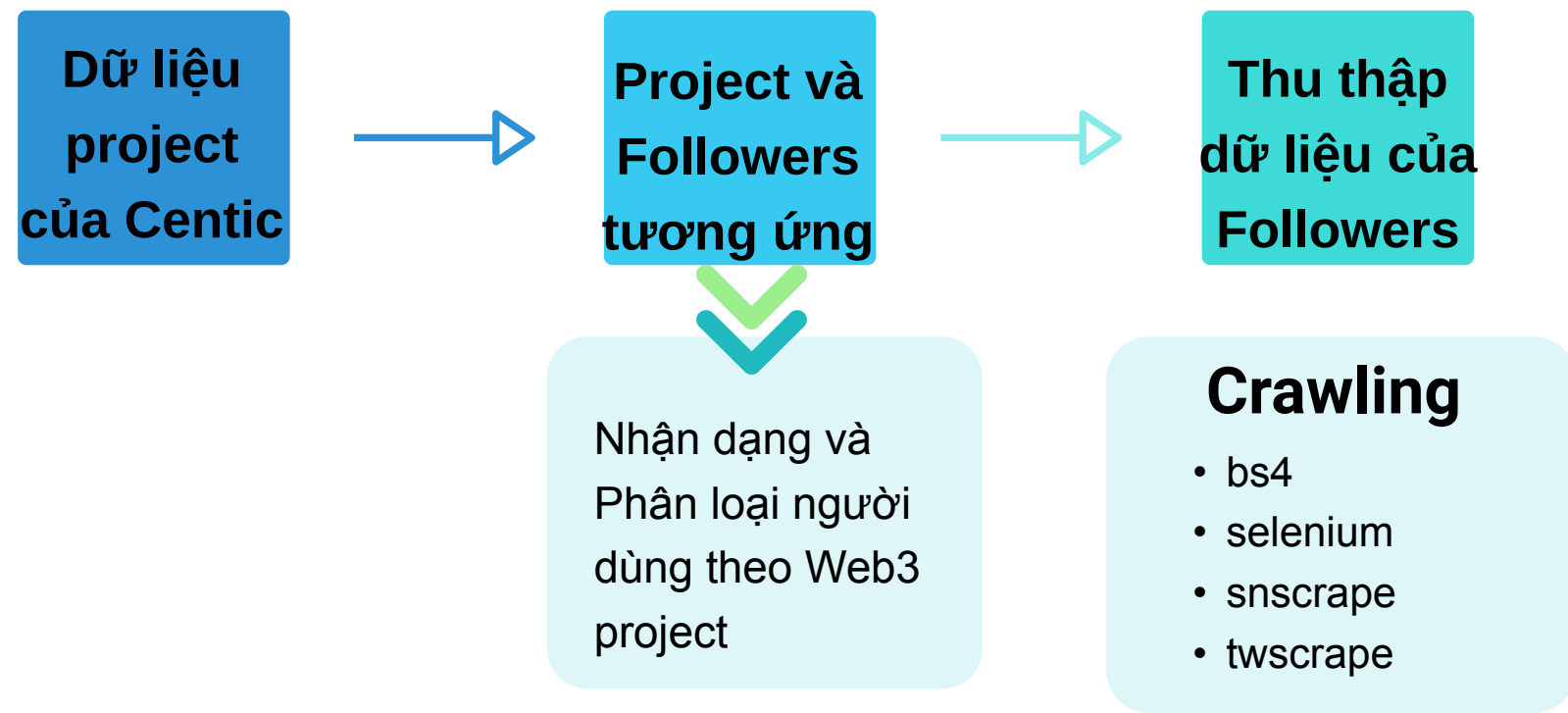


Data Processing



Data Analyzing

I, Thu thập Dữ liệu



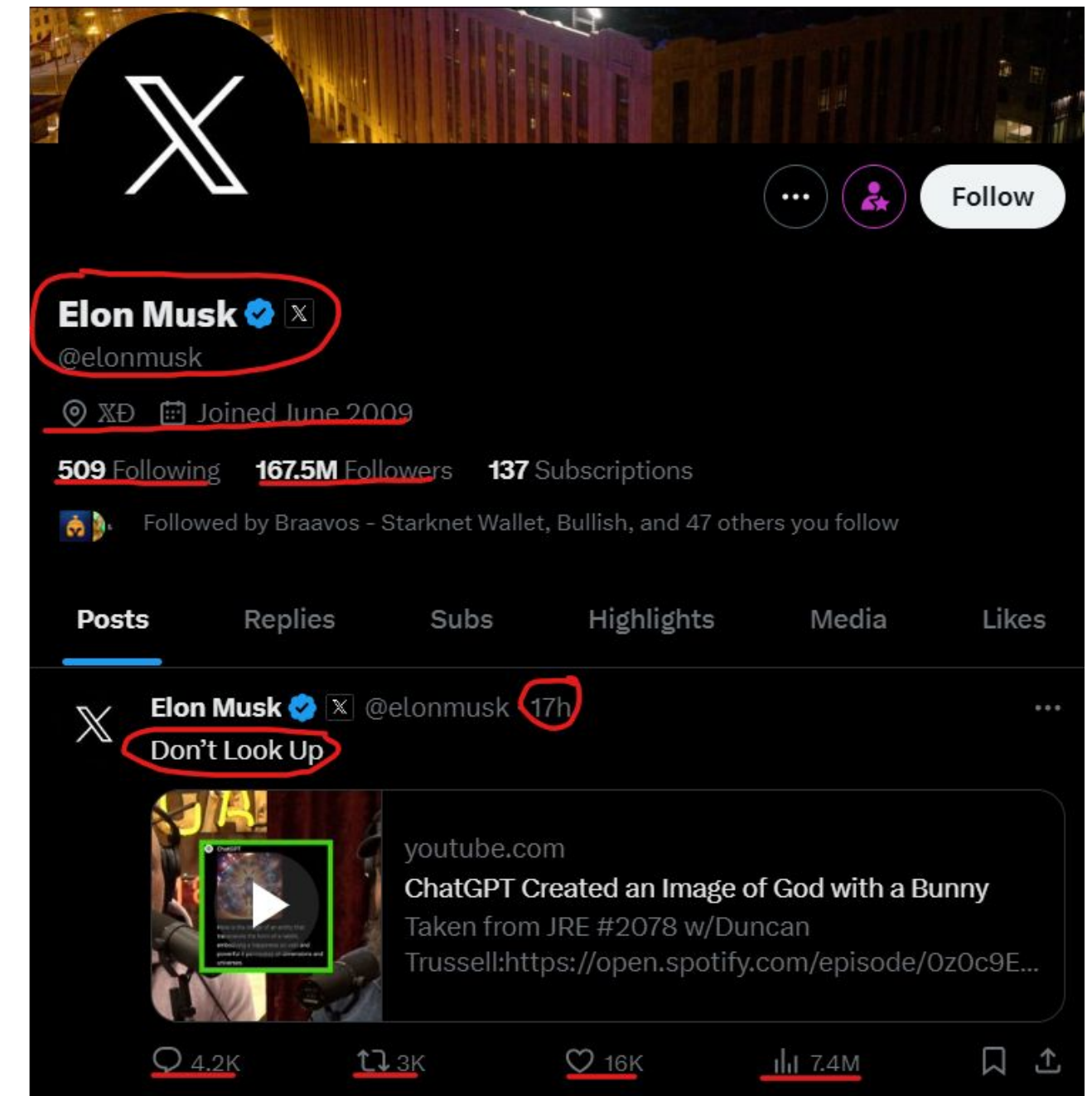
Dữ liệu thông tin người dùng từ X Profile gồm có:

```
{
  "id": 1478269476076351488,
  "username": "VioletAbtahi",
  "displayName": "Violet Abtahi",
  "url": "https://twitter.com/VioletAbtahi",
  "intro": "COO @enyaai | Co-founder @bobanetwork",
  "join_date": "2022-01-04T07:38:16+00:00",
  "location": "San Francisco, CA",
  "friends_count": 1529,
  "favorite_count": 277,
  "followers_count": 1531,
  "media_count": 9,
  "status_count": 190,
  "listed_count": 29,
  "is_protected": null,
  "is_verified": false,
}
```

Profile

```
{
  "user": "VioletAbtahi",
  "tweet_id": 158596754853564160,
  "tweet_content": "love japan web 30 community",
  "tweet_created_at": "2022-10-28T12:11:44+00:00",
  "like_count": 8,
  "retweet_count": 3,
  "comment_count": 2,
  "view_count": null,
  "quote_count": 0,
  "hashtags": [],
  "place": {
    "id": "a56612250c754f23",
    "fullName": "Tokyo-to, Japan",
    "name": "Tokyo-to",
    "type": "admin",
    "country": "Japan",
    "countryCode": "JP"
  },
  "mentioned_users": [],
  "long-lat": [
    -83.54424285888672,
    26.01139259338379
  ]
}
```

Tweets



II, Xử lý dữ liệu

Bot check

- Classification
- Kaggle Labeled Data

Text processing

- Tokenize, lemmatizer NLTK
- Remove emoji, stopword, special character,...

Name	Description	type
id	Twitter ID for user account	integer
account_type	indicator of the account type (bot or human)	string

Model Card for Model Geo-BERT-multilingual

This model predicts the geolocation of short texts (less than 500 words)

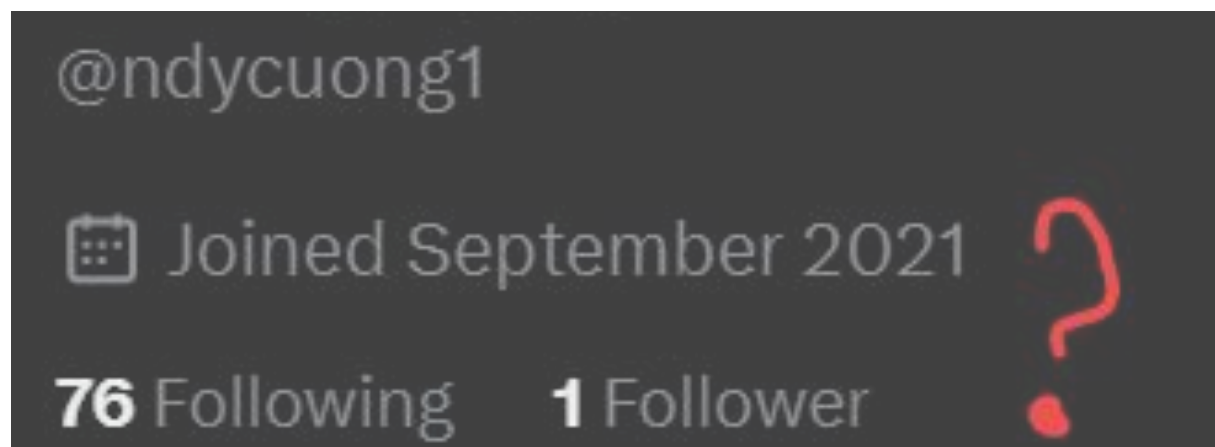
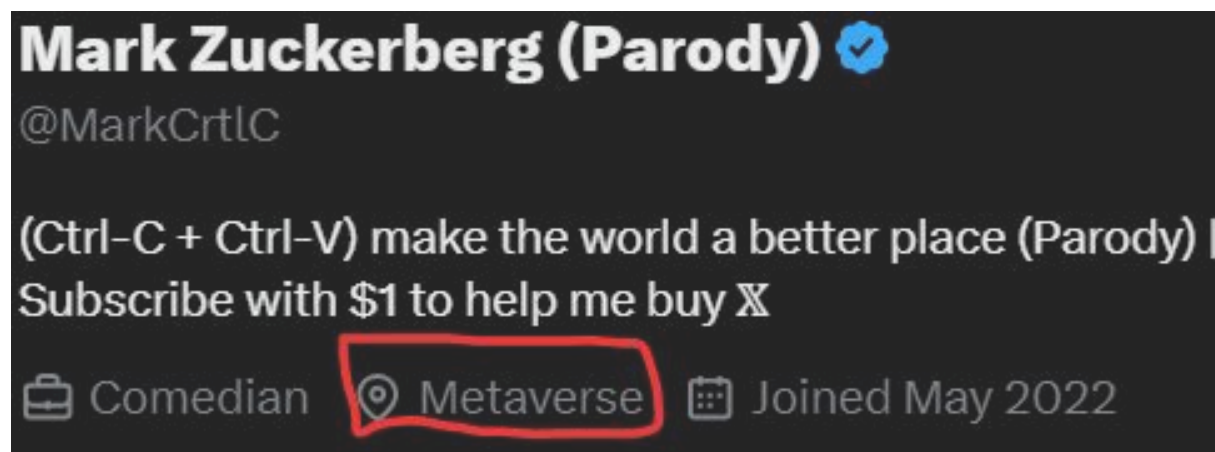
Crawling và
Sử dụng
Model

Train & Test

Phân loại
account

III, Phân tích dữ liệu

Dự đoán vị trí người dùng



- Mô hình dự đoán vị trí dựa trên text và metadata
- Sử dụng trên nội dung tweets với user chưa có thông tin sau stage 1

User home geolocation prediction task

- TEXT-ONLY: mean 892 km and median 31 km, 74% of Acc@161
- NON-GEO: mean 567 km and median 26 km, 82% of Acc@161

Stage 2

Geo-BERT

Stage 1

API Decoding

Nominatim

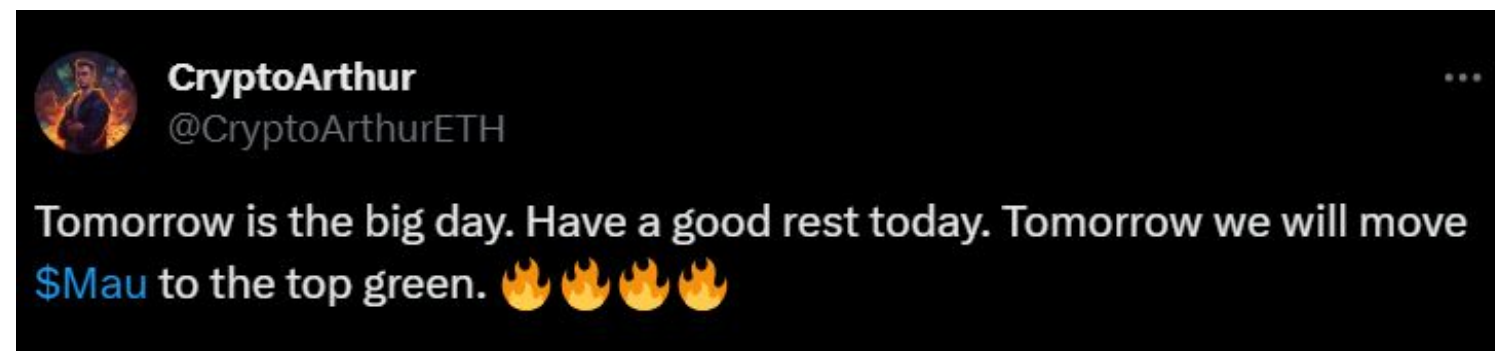
- Lấy thông tin khu vực địa lý từ thông tin trong trường location
- Trả về NULL với không phải dữ liệu vị trí, hoặc trường trống,

III, Phân tích dữ liệu

➔ Tweet Sentiment Analysis

twitter-XLM-roBERTa-base for Sentiment Analysis

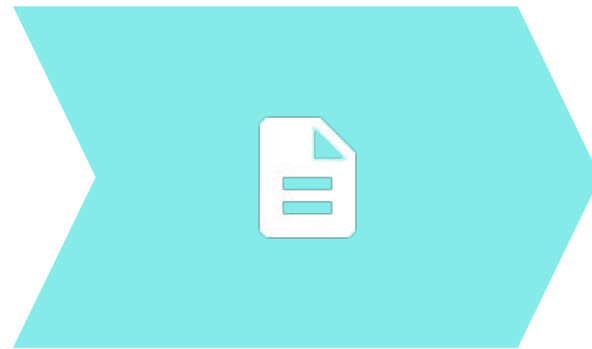
This is a multilingual XLM-roBERTa-base model trained on ~198M tweets and finetuned for sentiment analysis. The sentiment fine-tuning was done on 8 languages (Ar, En, Fr, De, Hi, It, Sp, Pt) but it can be used for more languages (see paper for details).



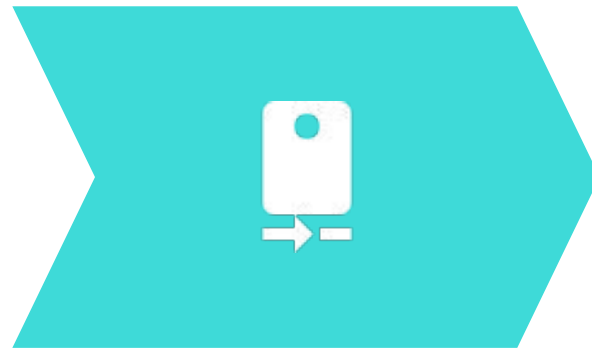
```
[{'label': 'positive', 'score': 0.7345331311225891}]
```

PROCESS

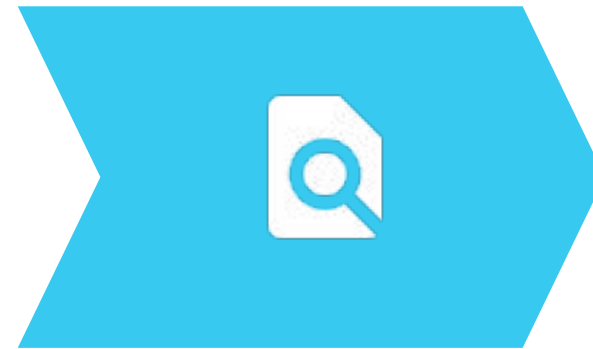
Google stored Data



Spark clean data



Elasticsearch index



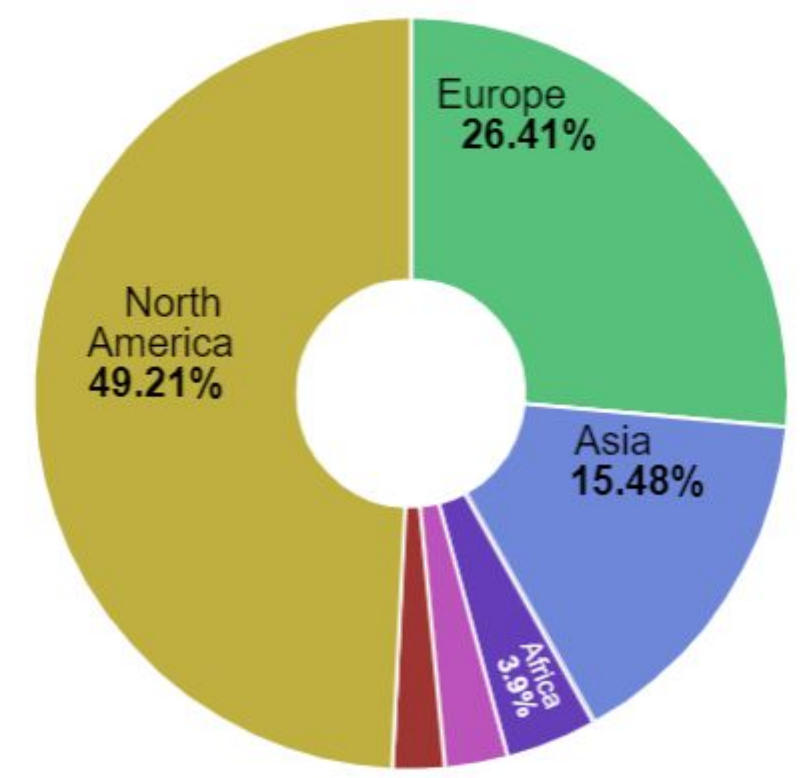
Kibana visualize



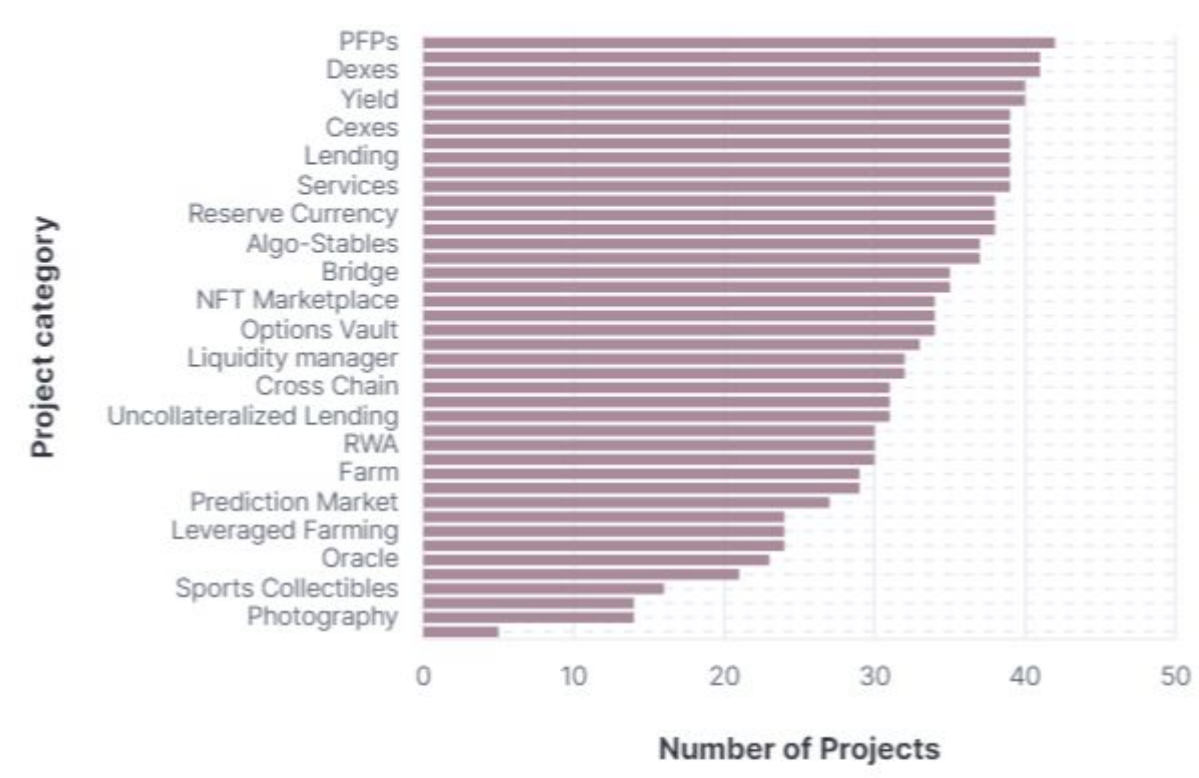
Demo



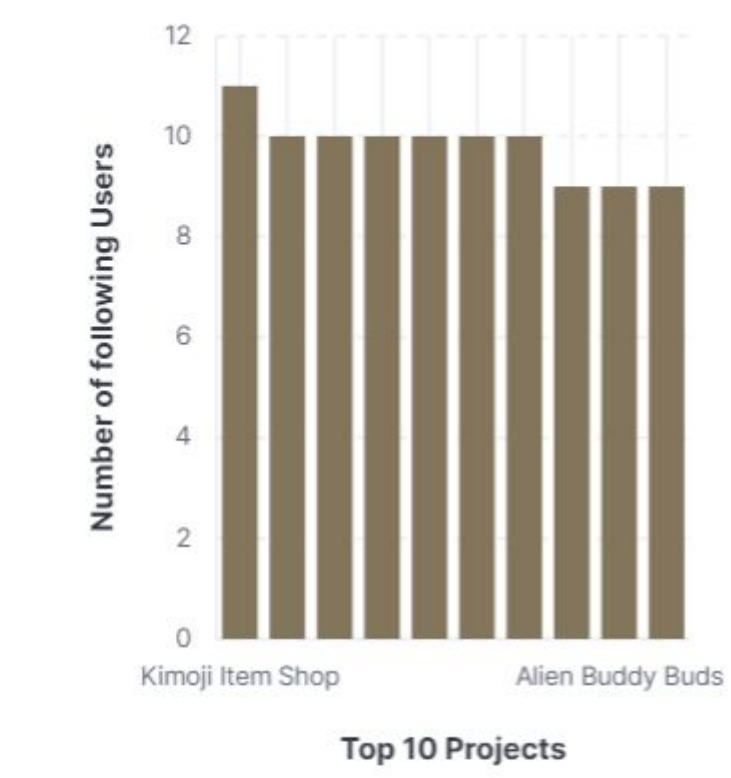
Continent category ⓘ



Project Category



Following Projects



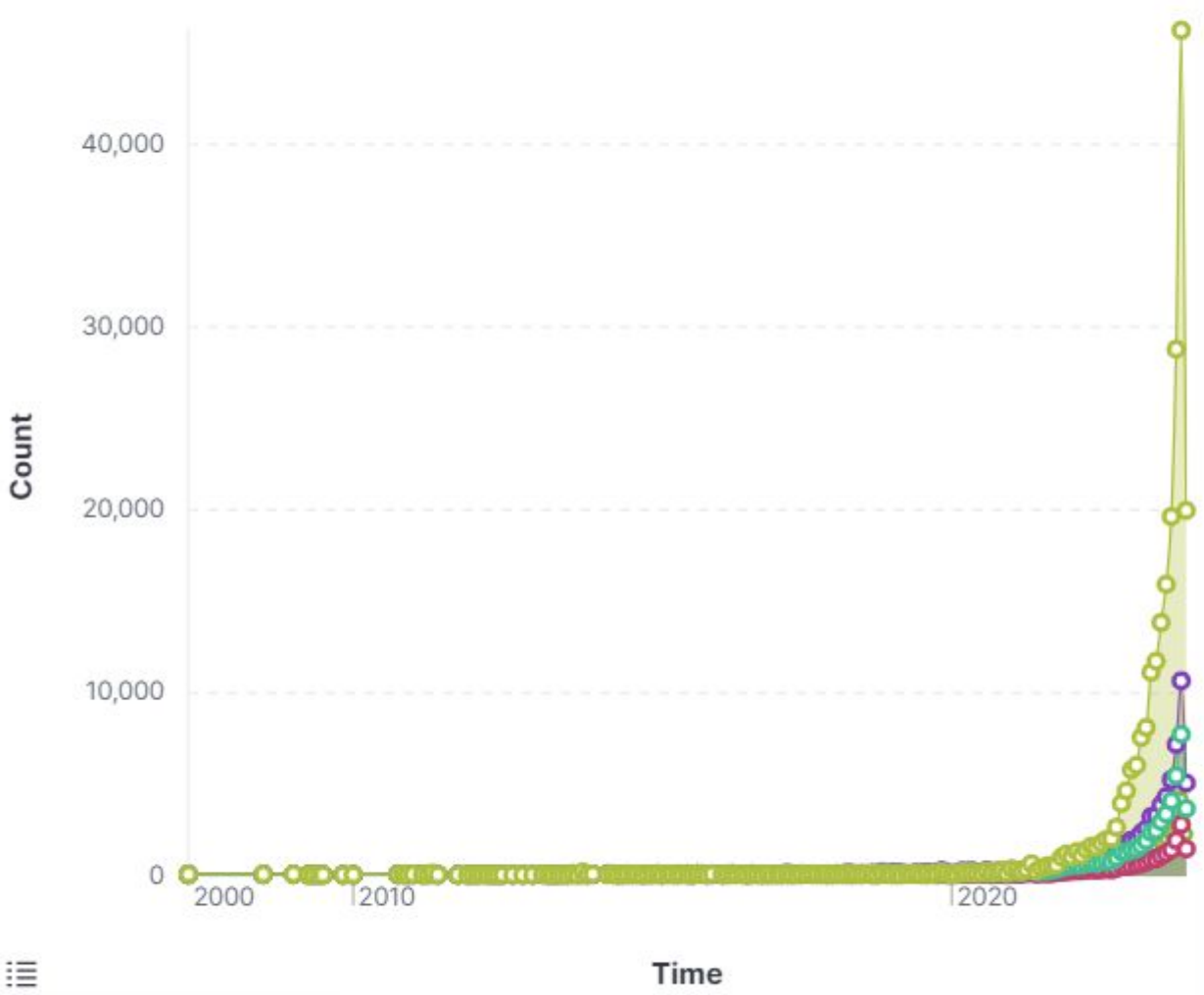
Total Users



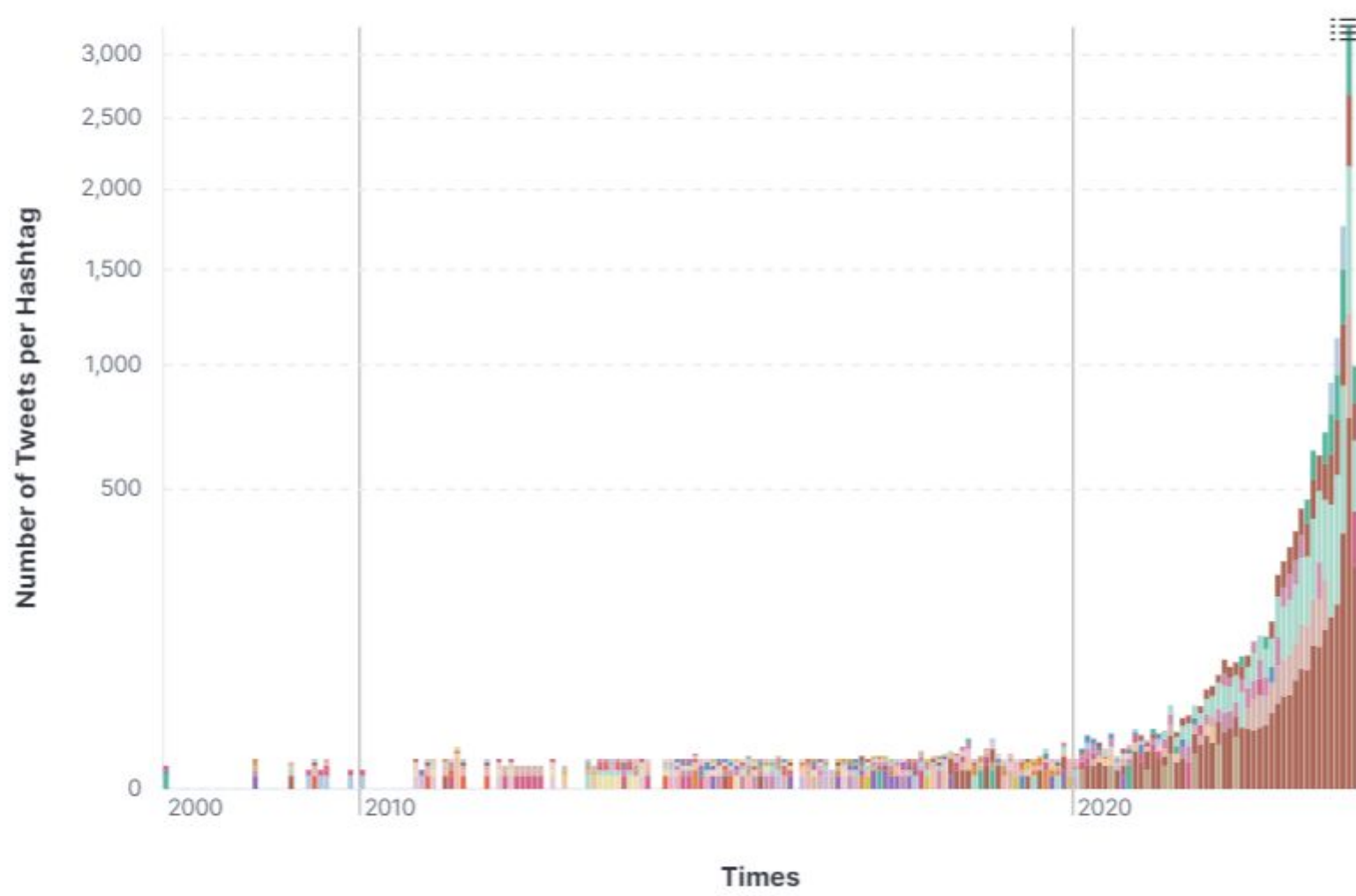
Total Tweets



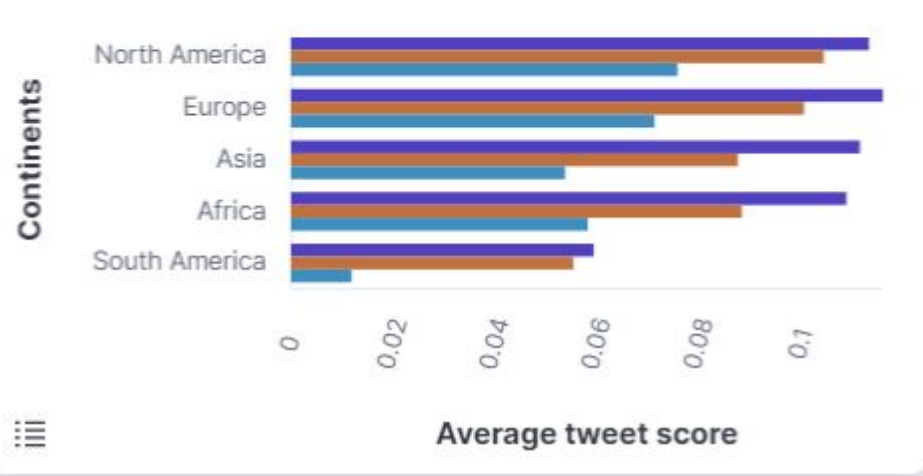
Iterations



Top Hashtags



Tweets Score



User Score

