Report : Automated Social Media OSINT Aggregation Pipeline

Name: Anjali Mishra
RollNo: 10471
Class: TE Comps B

1. Introduction

What is OSINT?

OSINT (Open Source Intelligence) is the practice of gathering information from publicly available sources such as social media platforms, blogs, forums, news websites, and other online platforms. Unlike traditional intelligence methods, OSINT relies solely on open and legally accessible information. This type of intelligence is widely used by researchers, security analysts, journalists, marketers, and law enforcement agencies to monitor trends, study public opinions, track emerging issues, and make informed decisions.

Lab Objective:

The objective of this project is to develop a fully automated OSINT pipeline that collects posts from multiple social media platforms, cleans and normalizes the text, performs sentiment analysis, and stores the structured data in a database for easy access and analysis. By doing so, researchers and analysts can quickly understand trends, measure public sentiment, and extract actionable insights without manually browsing through hundreds or thousands of posts.

This pipeline also serves as a learning exercise in integrating APIs, handling unstructured social media data, managing databases, and visualizing results. Overall, it demonstrates the potential of OSINT tools to make large-scale social media monitoring efficient, structured, and insightful.

2. Methodology

Platforms Integrated:

- Twitter
- Reddit
- Instagram
- Telegram
- Discord
- GitHub

Tools & Libraries Used:

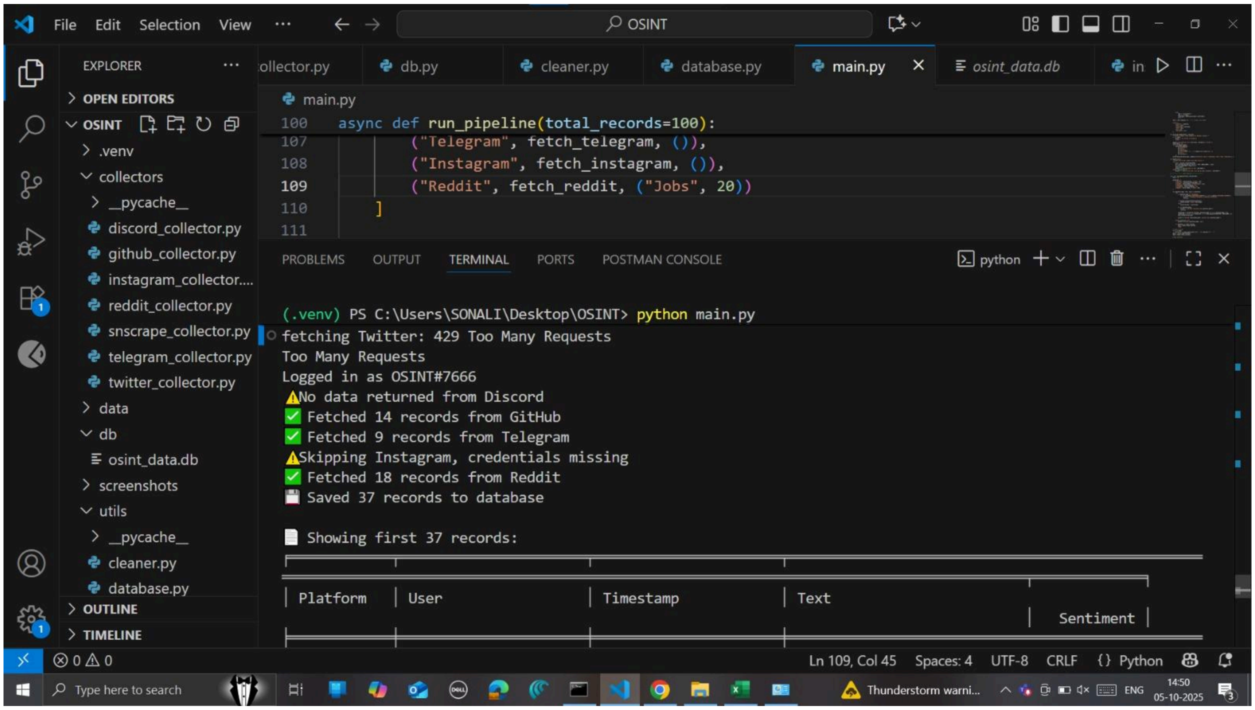
- Python – Main programming language.
- SQLite – To store collected data.
- TextBlob – To perform sentiment analysis.
- snscrape / PRAW / Telethon / API calls – For fetching posts from respective platforms.
- Matplotlib / Pandas – For creating charts and data analysis.
- dotenv – For storing API keys securely in a `.env` file.

Workflow:

1. Fetch posts from social media using APIs or scraping libraries.
2. Normalize and clean the data (remove URLs, emojis, and unwanted symbols).
3. Filter posts by language (English).
4. Analyze sentiment (positive, negative, neutral).
5. Save all structured data to a SQLite database.
6. Generate charts to visualize sentiment distribution.

3. Result:

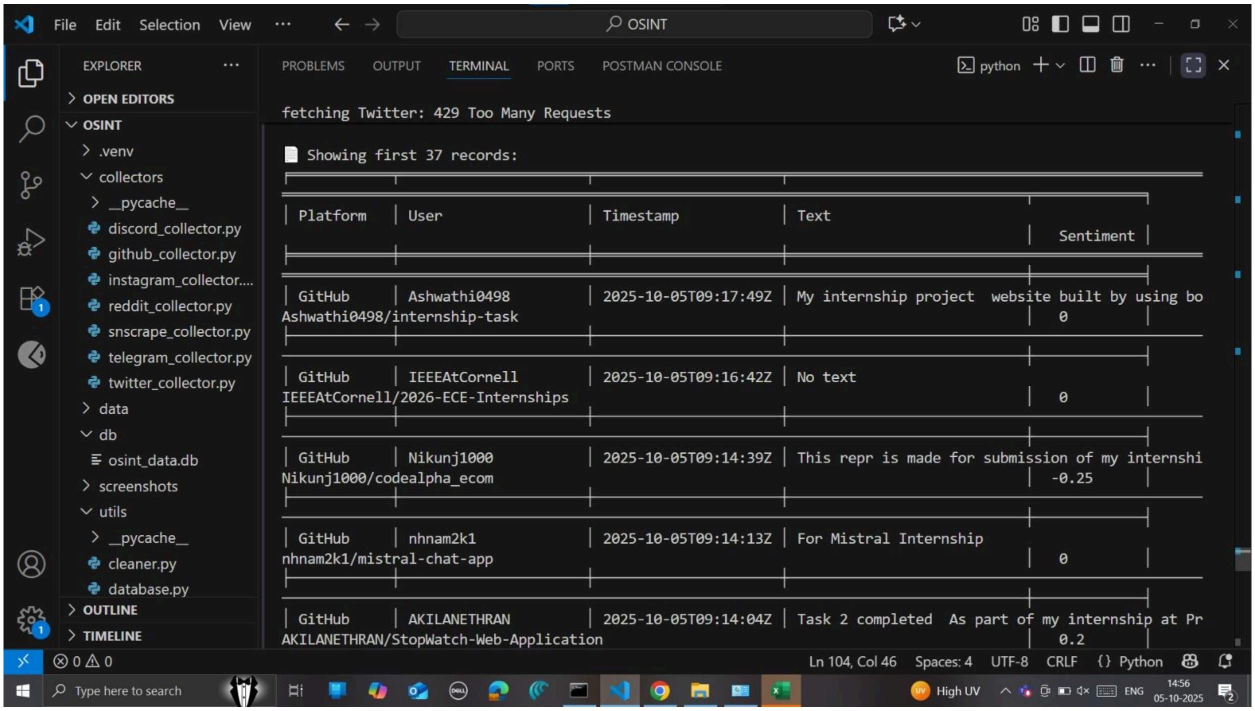
Data Collection:



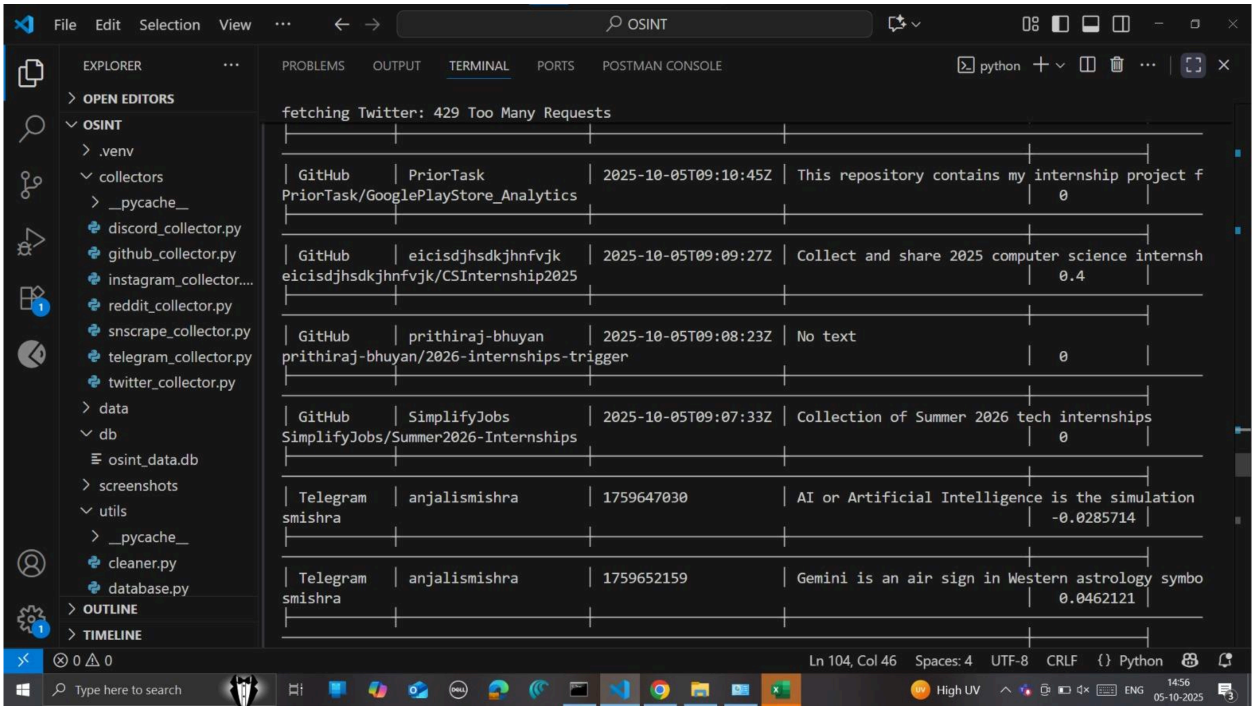
main.py

```python
async def run_pipeline(total_records=100):
        ("Telegram", fetch_telegram, ()),
        ("Instagram", fetch_instagram, ()),
        ("Reddit", fetch_reddit, ("Jobs", 20))
    ]

```

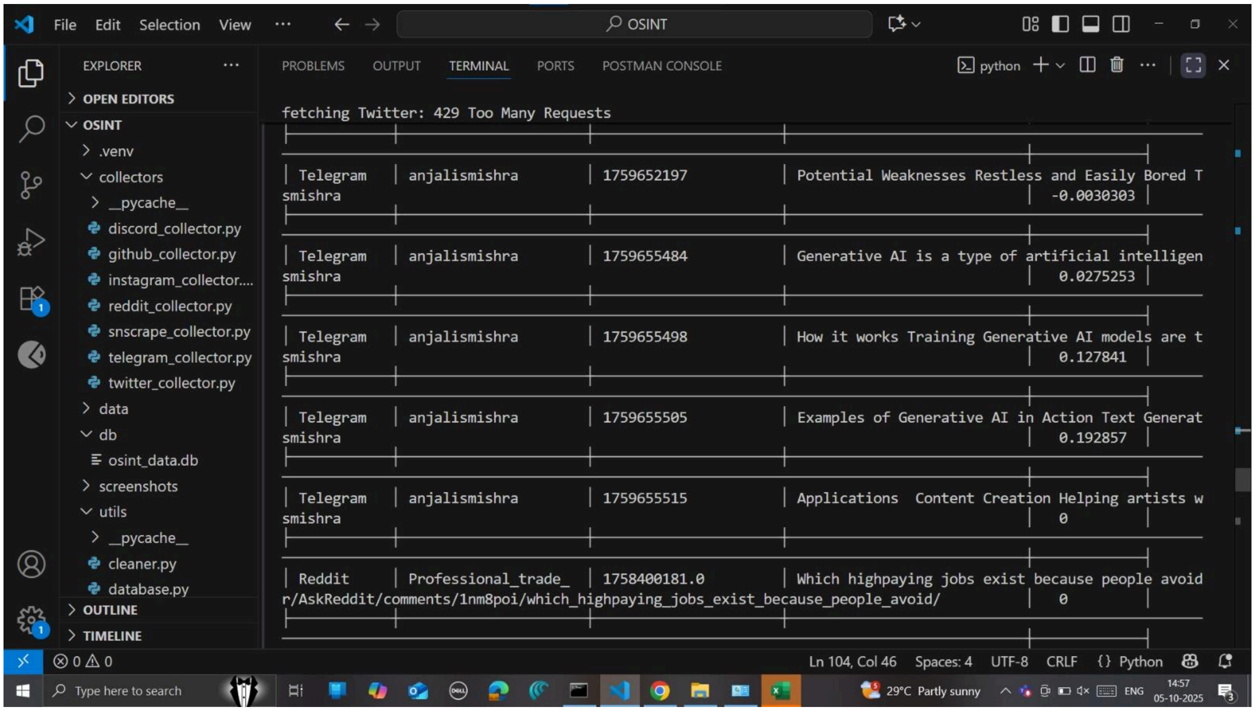
```
(.venv) PS C:\Users\SONALI\Desktop\OSINT> python main.py
fetching Twitter: 429 Too Many Requests
Too Many Requests
Logged in as OSINT#7666
⚠No data returned from Discord
✅ Fetched 14 records from GitHub
✅ Fetched 9 records from Telegram
⚠Skipping Instagram, credentials missing
✅ Fetched 18 records from Reddit
💾 Saved 37 records to database

📄 Showing first 37 records:
╒══════════╤═════════════════╤═════════════════╤══════════════════════════════════
═══════════════════════════════════════════════════════════════════════╤═══════════╕
│ Platform │ User            │ Timestamp       │ Text
                                                                       │ Sentiment │
╞══════════╪═════════════════╪═════════════════╪══════════════════════════════════
```



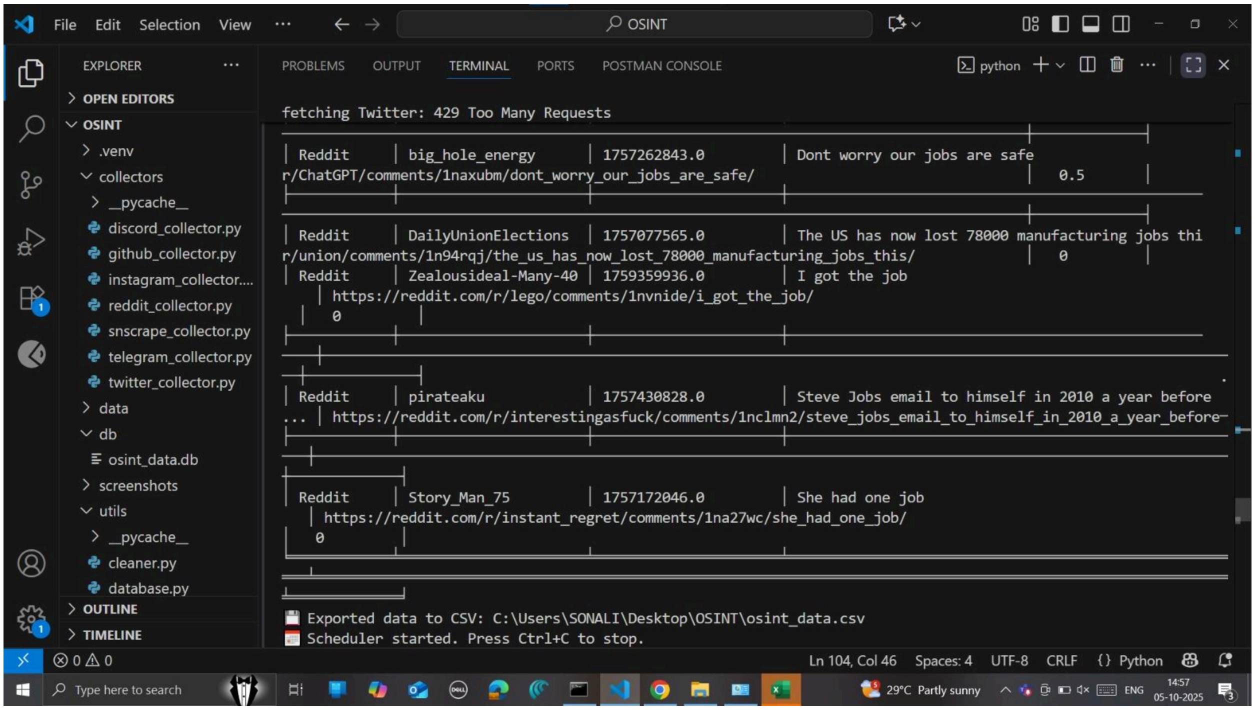
```
fetching Twitter: 429 Too Many Requests

📄 Showing first 37 records:
╒══════════╤═════════════════╤═════════════════╤══════════════════════════════════
═══════════════════════════════════════════════════════════════════════╤═══════════╕
│ Platform │ User            │ Timestamp       │ Text
                                                                       │ Sentiment │
╞══════════╪═════════════════╪═════════════════╪══════════════════════════════════
═══════════════════════════════════════════════════════════════════════╪═══════════╡
│ GitHub   │ Ashwathi0498    │ 2025-10-05T09:17:49Z │ My internship project  website built by using bo
Ashwathi0498/internship-task                                           │     0     │
├──────────┼─────────────────┼─────────────────┼──────────────────────────────────
───────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ IEEEAtCornell   │ 2025-10-05T09:16:42Z │ No text
IEEEAtCornell/2026-ECE-Internships                                     │     0     │
├──────────┼─────────────────┼─────────────────┼──────────────────────────────────
───────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ Nikunj1000      │ 2025-10-05T09:14:39Z │ This repr is made for submission of my internshi
Nikunj1000/codealpha_ecom                                              │   -0.25   │
├──────────┼─────────────────┼─────────────────┼──────────────────────────────────
───────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ nhnam2k1        │ 2025-10-05T09:14:13Z │ For Mistral Internship
nhnam2k1/mistral-chat-app                                              │     0     │
├──────────┼─────────────────┼─────────────────┼──────────────────────────────────
───────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ AKILANETHRAN    │ 2025-10-05T09:14:04Z │ Task 2 completed  As part of my internship at Pr
AKILANETHRAN/StopWatch-Web-Application                                 │    0.2    │
```



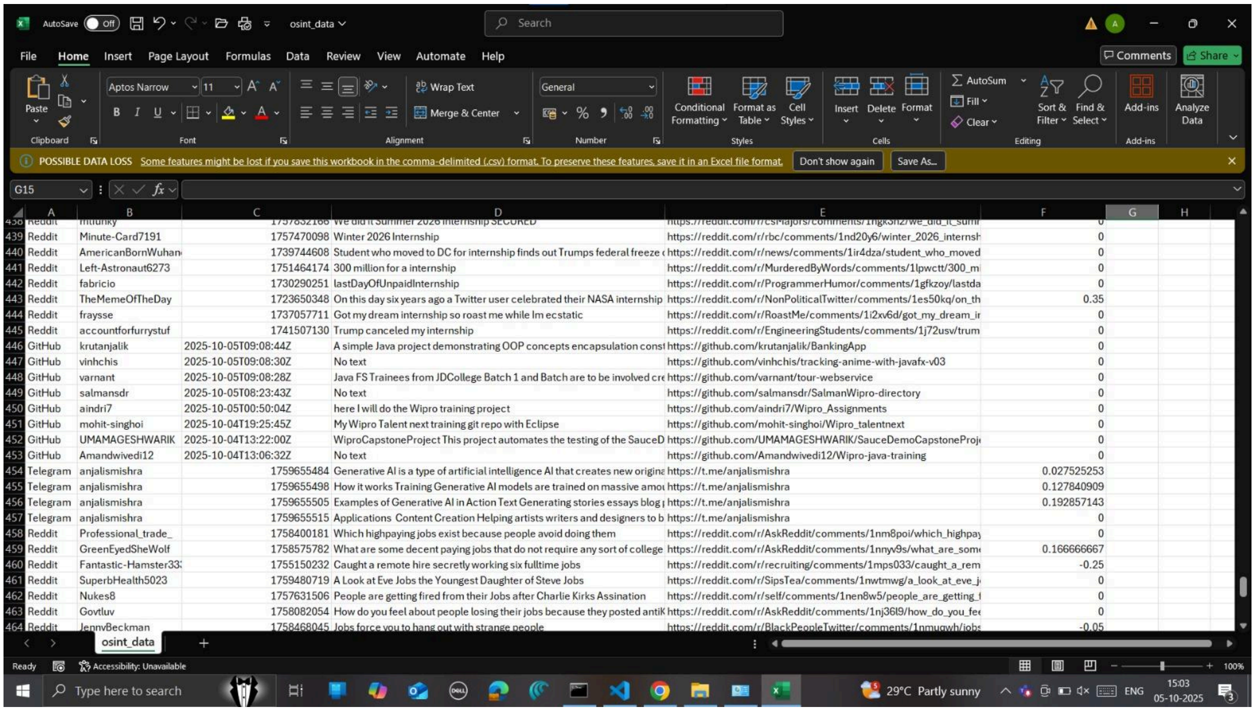
```
fetching Twitter: 429 Too Many Requests
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ PriorTask            │ 2025-10-05T09:10:45Z │ This repository contains my internship project f
PriorTask/GooglePlayStore_Analytics                                   │   0       │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ eicisdjhsdkjhnfvjk   │ 2025-10-05T09:09:27Z │ Collect and share 2025 computer science internsh
eicisdjhsdkjhnfvjk/CSInternship2025                                   │   0.4     │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ prithiraj-bhuyan     │ 2025-10-05T09:08:23Z │ No text
prithiraj-bhuyan/2026-internships-trigger                             │   0       │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ GitHub   │ SimplifyJobs         │ 2025-10-05T09:07:33Z │ Collection of Summer 2026 tech internships
SimplifyJobs/Summer2026-Internships                                   │   0       │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759647030           │ AI or Artificial Intelligence is the simulation
smishra                                                               │ -0.0285714 │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759652159           │ Gemini is an air sign in Western astrology symbo
smishra                                                               │   0.0462121 │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
```



```
fetching Twitter: 429 Too Many Requests
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759652197           │ Potential Weaknesses Restless and Easily Bored T
smishra                                                               │ -0.0030303 │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759655484           │ Generative AI is a type of artificial intelligen
smishra                                                               │   0.0275253 │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759655498           │ How it works Training Generative AI models are t
smishra                                                               │   0.127841  │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759655505           │ Examples of Generative AI in Action Text Generat
smishra                                                               │   0.192857  │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Telegram │ anjalismishra        │ 1759655515           │ Applications  Content Creation Helping artists w
smishra                                                               │   0       │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
│ Reddit   │ Professional_trade_  │ 1758400181.0         │ Which highpaying jobs exist because people avoid
r/AskReddit/comments/1nm8poi/which_highpaying_jobs_exist_because_people_avoid/ │   0       │
├─────────┼──────────────────────┼──────────────────────┼──────────────────────────────────────
──────────────────────────────────────────────────────────────────────┼───────────┤
```

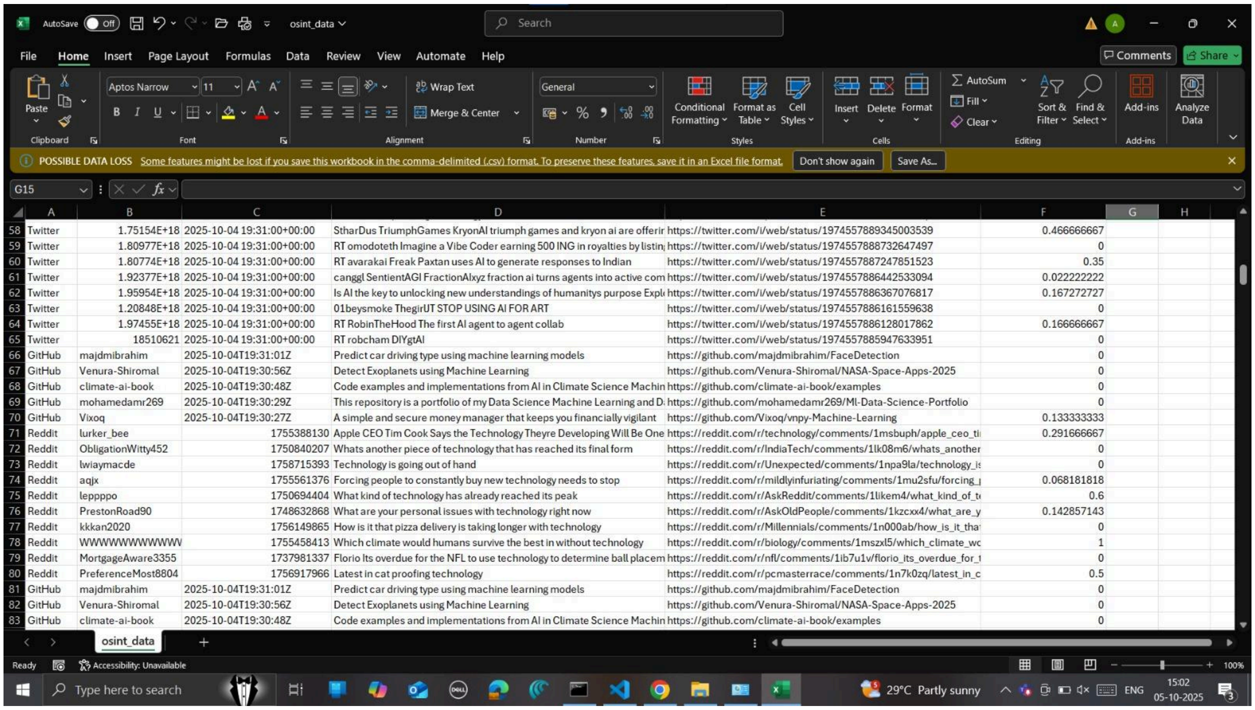


```
fetching Twitter: 429 Too Many Requests
                                                                     ┼               ┤
│ Reddit    │ big_hole_energy        │ 1757262843.0          │ Dont worry our jobs are safe
r/ChatGPT/comments/1naxubm/dont_worry_our_jobs_are_safe/              │   0.5         │
├───────────┼────────────────────────┼───────────────────────┼───────────────────────────────
                                                                     ┼               ┤
│ Reddit    │ DailyUnionElections    │ 1757077565.0          │ The US has now lost 78000 manufacturing jobs thi
r/union/comments/1n94rqj/the_us_has_now_lost_78000_manufacturing_jobs_this/         │   0
│ Reddit    │ Zealousideal-Many-40   │ 1759359936.0          │ I got the job
    │ https://reddit.com/r/lego/comments/1nvnide/i_got_the_job/
  │   0   │
├───────────┼────────────────────────┼───────────────────────┼───────────────────────────────

──┼───────────────┤
│ Reddit    │ pirateaku              │ 1757430828.0          │ Steve Jobs email to himself in 2010 a year before
... │ https://reddit.com/r/interestingasfuck/comments/1nclmn2/steve_jobs_email_to_himself_in_2010_a_year_before─
├───────────┼────────────────────────┼───────────────────────┼───────────────────────────────
──┼
├───────────┤
│ Reddit    │ Story_Man_75           │ 1757172046.0          │ She had one job
   │ https://reddit.com/r/instant_regret/comments/1na27wc/she_had_one_job/
   0   │
└───────────┴────────────────────────┴───────────────────────┴───────────────────────────────
══════════════════════════════════════════════════════════════════════════════════════════════
└───────────┘
💾 Exported data to CSV: C:\Users\SONALI\Desktop\OSINT\osint_data.csv
📅 Scheduler started. Press Ctrl+C to stop.
```



Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B	C	D	E	F	G	H
438	Reddit	mtfunky	1757832166	We did it Summer 2026 Internship SECURED	https://reddit.com/r/csMajors/comments/1ngk3n2/we_did_it_sum	0		
439	Reddit	Minute-Card7191	1757470098	Winter 2026 Internship	https://reddit.com/r/rbc/comments/1nd20y6/winter_2026_internsh	0		
440	Reddit	AmericanBornWuhan	1739744608	Student who moved to DC for internship finds out Trumps federal freeze (	https://reddit.com/r/news/comments/1ir4dza/student_who_moved	0		
441	Reddit	Left-Astronaut6273	1751464174	300 million for a internship	https://reddit.com/r/MurderedByWords/comments/1lpwctt/300_mi	0		
442	Reddit	fabricio	1730290251	lastDayOfUnpaidInternship	https://reddit.com/r/ProgrammerHumor/comments/1gfkzoy/lastda	0		
443	Reddit	TheMemeOfTheDay	1723650348	On this day six years ago a Twitter user celebrated their NASA internship	https://reddit.com/r/NonPoliticalTwitter/comments/1es50kq/on_th	0.35		
444	Reddit	fraysse	1737057711	Got my dream internship so roast me while Im ecstatic	https://reddit.com/r/RoastMe/comments/1i2xv6d/got_my_dream_ir	0		
445	Reddit	accountforfurrystuf	1741507130	Trump canceled my internship	https://reddit.com/r/EngineeringStudents/comments/1j72usv/trum	0		
446	GitHub	krutanjalik	2025-10-05T09:08:44Z	A simple Java project demonstrating OOP concepts encapsulation const	https://github.com/krutanjalik/BankingApp	0		
447	GitHub	vinhchis	2025-10-05T09:08:30Z	No text	https://github.com/vinhchis/tracking-anime-with-javafx-v03	0		
448	GitHub	varnant	2025-10-05T09:08:28Z	Java FS Trainees from JDCollege Batch 1 and Batch are to be involved cre	https://github.com/varnant/tour-webservice	0		
449	GitHub	salmansdr	2025-10-05T08:23:43Z	No text	https://github.com/salmansdr/SalmanWipro-directory	0		
450	GitHub	aindri7	2025-10-05T00:50:04Z	here I will do the Wipro training project	https://github.com/aindri7/Wipro_Assignments	0		
451	GitHub	mohit-singhoi	2025-10-04T19:25:45Z	My Wipro Talent next training git repo with Eclipse	https://github.com/mohit-singhoi/Wipro_talentnext	0		
452	GitHub	UMAMAGESHWARIK	2025-10-04T13:22:00Z	WiproCapstoneProject This project automates the testing of the SauceD	https://github.com/UMAMAGESHWARIK/SauceDemoCapstoneProj	0		
453	GitHub	Amandwivedi12	2025-10-04T13:06:32Z	No text	https://github.com/Amandwivedi12/Wipro-java-training	0		
454	Telegram	anjalismishra	1759655484	Generative AI is a type of artificial intelligence AI that creates new origina	https://t.me/anjalismishra	0.027525253		
455	Telegram	anjalismishra	1759655498	How it works Training Generative AI models are trained on massive amou	https://t.me/anjalismishra	0.127840909		
456	Telegram	anjalismishra	1759655505	Examples of Generative AI in Action Text Generating stories essays blog p	https://t.me/anjalismishra	0.192857143		
457	Telegram	anjalismishra	1759655515	Applications Content Creation Helping artists writers and designers to b	https://t.me/anjalismishra	0		
458	Reddit	Professional_trade_	1758400181	Which highpaying jobs exist because people avoid doing them	https://reddit.com/r/AskReddit/comments/1nm8poi/which_highpay	0		
459	Reddit	GreenEyedSheWolf	1758575782	What are some decent paying jobs that do not require any sort of college	https://reddit.com/r/AskReddit/comments/1nnyv9s/what_are_some	0.166666667		
460	Reddit	Fantastic-Hamster33	1755150232	Caught a remote hire secretly working six fulltime jobs	https://reddit.com/r/recruiting/comments/1mps033/caught_a_rem	-0.25		
461	Reddit	SuperbHealth5023	1759480719	A Look at Eve Jobs the Youngest Daughter of Steve Jobs	https://reddit.com/r/SipsTea/comments/1nwtmwg/a_look_at_eve_j	0		
462	Reddit	Nukes8	1757631506	People are getting fired from their Jobs after Charlie Kirks Assination	https://reddit.com/r/self/comments/1nen8w5/people_are_getting_f	0		
463	Reddit	Govtluv	1758082054	How do you feel about people losing their jobs because they posted antiK	https://reddit.com/r/AskReddit/comments/1nj36l9/how_do_you_fee	0		
464	Reddit	JennyBeckman	1758468045	Jobs force you to hang out with strange people	https://reddit.com/r/BlackPeopleTwitter/comments/1nmuqwh/jobs	-0.05		



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again Save As...

	A	B	C	D	E	F
58	Twitter	1.75154E+18	2025-10-04 19:31:00+00:00	StharDus TriumphGames KryonAI triumph games and kryon ai are offerir	https://twitter.com/i/web/status/1974557889345003539	0.466666667
59	Twitter	1.80977E+18	2025-10-04 19:31:00+00:00	RT omodoteth Imagine a Vibe Coder earning 500 ING in royalties by listin	https://twitter.com/i/web/status/1974557888732647497	0
60	Twitter	1.80774E+18	2025-10-04 19:31:00+00:00	RT avarakai Freak Paxtan uses AI to generate responses to Indian	https://twitter.com/i/web/status/1974557887247851523	0.35
61	Twitter	1.92377E+18	2025-10-04 19:31:00+00:00	canggl SentientAGI FractionAIxyz fraction ai turns agents into active com	https://twitter.com/i/web/status/1974557886442533094	0.022222222
62	Twitter	1.95954E+18	2025-10-04 19:31:00+00:00	Is AI the key to unlocking new understandings of humanitys purpose Expl	https://twitter.com/i/web/status/1974557886367076817	0.167272727
63	Twitter	1.20848E+18	2025-10-04 19:31:00+00:00	01beysmoke ThegirUT STOP USING AI FOR ART	https://twitter.com/i/web/status/1974557886161559638	0
64	Twitter	1.97455E+18	2025-10-04 19:31:00+00:00	RT RobinTheHood The first AI agent to agent collab	https://twitter.com/i/web/status/1974557886128017862	0.166666667
65	Twitter	18510621	2025-10-04 19:31:00+00:00	RT robcham DIYgtAI	https://twitter.com/i/web/status/1974557885947633951	0
66	GitHub	majdmibrahim	2025-10-04T19:31:01Z	Predict car driving type using machine learning models	https://github.com/majdmibrahim/FaceDetection	0
67	GitHub	Venura-Shiromal	2025-10-04T19:30:56Z	Detect Exoplanets using Machine Learning	https://github.com/Venura-Shiromal/NASA-Space-Apps-2025	0
68	GitHub	climate-ai-book	2025-10-04T19:30:48Z	Code examples and implementations from AI in Climate Science Machin	https://github.com/climate-ai-book/examples	0
69	GitHub	mohamedamr269	2025-10-04T19:30:29Z	This repository is a portfolio of my Data Science Machine Learning and D	https://github.com/mohamedamr269/Ml-Data-Science-Portfolio	0
70	GitHub	Vixoq	2025-10-04T19:30:27Z	A simple and secure money manager that keeps you financially vigilant	https://github.com/Vixoq/vnpy-Machine-Learning	0.133333333
71	Reddit	lurker_bee	1755388130	Apple CEO Tim Cook Says the Technology Theyre Developing Will Be One	https://reddit.com/r/technology/comments/1msbuph/apple_ceo_ti	0.291666667
72	Reddit	ObligationWitty452	1750840207	Whats another piece of technology that has reached its final form	https://reddit.com/r/IndiaTech/comments/1lk08m6/whats_another	0
73	Reddit	lwiaymacde	1758715393	Technology is going out of hand	https://reddit.com/r/Unexpected/comments/1npa9la/technology_is	0
74	Reddit	aqjx	1755561376	Forcing people to constantly buy new technology needs to stop	https://reddit.com/r/mildlyinfuriating/comments/1mu2sfu/forcing_	0.068181818
75	Reddit	leppppo	1750694404	What kind of technology has already reached its peak	https://reddit.com/r/AskReddit/comments/1likem4/what_kind_of_t	0.6
76	Reddit	PrestonRoad90	1748632868	What are your personal issues with technology right now	https://reddit.com/r/AskOldPeople/comments/1kzcxx4/what_are_y	0.142857143
77	Reddit	kkkan2020	1756149865	How is it that pizza delivery is taking longer with technology	https://reddit.com/r/Millennials/comments/1n000ab/how_is_it_tha	0
78	Reddit	WWWWWWWWWWW	1755458413	Which climate would humans survive the best in without technology	https://reddit.com/r/biology/comments/1mszxl5/which_climate_wc	1
79	Reddit	MortgageAware3355	1737981337	Florio Its overdue for the NFL to use technology to determine ball placem	https://reddit.com/r/nfl/comments/1ib7u1v/florio_its_overdue_for_t	0
80	Reddit	PreferenceMost8804	1756917966	Latest in cat proofing technology	https://reddit.com/r/pcmasterrace/comments/1n7k0zq/latest_in_c	0.5
81	GitHub	majdmibrahim	2025-10-04T19:31:01Z	Predict car driving type using machine learning models	https://github.com/majdmibrahim/FaceDetection	0
82	GitHub	Venura-Shiromal	2025-10-04T19:30:56Z	Detect Exoplanets using Machine Learning	https://github.com/Venura-Shiromal/NASA-Space-Apps-2025	0
83	GitHub	climate-ai-book	2025-10-04T19:30:48Z	Code examples and implementations from AI in Climate Science Machin	https://github.com/climate-ai-book/examples	0

4. Challenges

1. API Limits: Twitter and Reddit limit the number of requests per hour. Had to reduce fetch count and handle rate-limiting.

2. Authentication: Some platforms require API keys or tokens. Managing them securely was necessary.

3. Data Cleaning: Social media text often has emojis, symbols, and URLs that needed to be removed.

4. Errors & Exceptions: Occasionally some platforms returned empty data or network errors. Had to implement error handling.

5. Conclusion:

- The pipeline successfully collected and analyzed data from multiple platforms.

- Sentiment analysis provides a quick overview of public opinion on topics.

- The database structure allows easy querying and further analysis.

Future Improvements:

- Add more platforms like TikTok, Facebook, Mastodon.
- Enhance sentiment analysis using advanced models (like transformers).
- Add real-time alerts for trending topics.
- Implement better visualization dashboards for easier insights.