AI-Based Social Media Harvesting for Analysis

By

ASC Habibah Saeed

PC Ansa Farrukh

NC Sajjad Ali

PC Usman Nisar

Supervised by:

Assoc Prof Dr Mian Muhammad Waseem Iqbal

Submitted to the faculty of the Department of Electrical Engineering,
Military College of Signals, National University of Sciences and Technology, Islamabad,
in partial fulfillment of the requirements of a B.E Degree in Electrical Engineering.

June 2023

In the name of ALLAH, the Most benevolent, the Most Courteous

CERTIFICATE OF CORRECTNESS AND APPROVAL

This is to officially state that the thesis work contained in this report

"AI-Based Social Media Harvesting for Analysis"

is carried out by

ASC Habibah Saeed, PC Ansa Farrukh, NC Sajjad Ali, PC Usman Nisar

under my supervision and that in my judgment, it is fully ample, in scope and excellence, for the degree of Bachelor of Electrical Engineering in Military College of Signals National University of Sciences and Technology (NUST), Islamabad.

Approved by

Supervisor
Assoc Prof Dr Mian Muhammad Waseem Iqbal
Department of IS, MCS

Date: 26-April-2023

DECLARATION OF ORIGINALITY

We hereby declare that no portion of the work presented in this thesis has been submitted in support of another award or qualification in either this institute or anywhere else.

ACKNOWLEDGEMENTS

We would like to sincerely thank Allah (SWT) for providing us with the courage, knowledge, and endurance necessary to finish this thesis. We appreciate that this accomplishment would not have been achieved without His direction and blessings. We also want to express our gratitude to our parents for their continuous love, support, and inspiration during our academic journey. Our success has been entirely due to their sacrifices and faith in us.

We also appreciate Dr. Mian Muhammad Waseem Iqbal, our supervisor, for his expert advice, patience, helpful feedback, and constant support. His expertise, perspective, and mentoring have been essential to this project's success.

We would like to express our gratitude to all the professors, colleagues, and friends who have given us the materials, suggestions, and help we need to finish this thesis. Our motivation and focus have remained strong thanks to their encouragement and support during these challenging moments. Finally, we would want to express our gratitude to our families for always being there for us, through good times and bad. Their support, encouragement, and compassion have been essential to our achievement. We appreciate your support in joining us on this incredible journey and in helping us reach this significant life milestone.

الحمٍ ل رِ العالمين، وال١ لُ والال ُ على أشرِ النبياِ والمرسلين؛ نبينا محمٍ صلى ١ عليِ وآلِ وصحبِ أجمعين

Dedicated to our parents, siblings, and teachers who have supported us throughout our academic journey. Their unwavering encouragement and guidance have been instrumental in shaping our intellectual and personal growth. We are forever grateful for their love, patience, and sacrifice, without which this achievement would not have been possible.

Plagiarism Certificate (Turnitin Report)

This thesis has a **10 %** similarity index. The Turnitin report endorsed by Supervisor is attached.

ASC Habibah Saeed
325360

PC Ansa Farrukh
325775

NC Sajjad Ali
284912

PC Usman Nisar
325765

Signature of Supervisor

ABSTRACT

Data on social media has always been a source of interest for businesses, government organizations, and other agencies for various reasons. However, dealing with such a plethora of data and its analysis for a particular purpose is not always easy. Therefore, we have customized an Open-source intelligence (OSINT) tool (web application) to gather and analyze data from publicly available sources (currently we are working only on Twitter) which can then be used for the following two purposes:

Business Perspective: - In this era of digitalization, businesses are required to be advertised on Social Media platforms. From the collected data, our tool can help businesses keep track of market intelligence, customer intelligence, competitive analysis, and digital marketing and management. This helps in identifying weaknesses, and planning strategically for the future and thus ensures business growth.

Law Enforcement Agencies (LEA's) Perspective: - If an anti-state trend gets viral on social media, it becomes very difficult for LEAs to trace back to the root originator of this trend. Our customized tool will narrow down the possible suspects. It also provides deeper insights and various visualizations of viral hashtags and trends.

Table of Contents

List of Figures

Chapter 1: Introduction

Engineering is a field of study that involves the operation of scientific, mathematical, and specialized knowledge to design, develop, and upgrade systems, structures, machines, and processes. It plays a pivotal part in working problems and meeting the requirements of people, businesses, and industries. Engineers create new products and technologies, enhance existing ones, and make sure they're durable, safe, and efficient.

Engineering plays a significant part in social media harvesting by furnishing the tools and technologies necessary to collect, store, and explore the vast quantities of data generated by social media platforms. Engineers develop algorithms and software programs that can crawl through social media sites and release applicable information, similar to user demographics, interests, and actions while ensuring that user privacy and security are defended.

Twitter is a free social networking platform where users broadcast short posts known as tweets. Tweets sent by users display both on their profiles and in the news feeds of their followers. Users can use hashtags in their postings to weave tweets into a discussion thread or relate them to general information. This makes the tweet hunt under that keyword.

Numerous businesses face challenges when it comes to digital and social media marketing moreover, they don’t know how their product is doing in the market among customers or how to launch a successful campaign strategy grounded on the analysis. So, to give new confines for the business’s growth, Twitter helps them to reach the targeted audience, identify negative trends, analyze contender trends, to help them to identify propaganda networks, and fake biographies, coordinate campaigns, and discover the crucial contributors of that propaganda.

1.1 Overview

Open-Source Intelligence (OSINT) involves the collection or processes of gathering data and profiling intimately available private and public sector information sources about individualities and business intelligence purposes. These sources include the internet and other social media platforms similar to Facebook, email, Twitter, and WhatsApp. Its main idea is to collect, clean, and aggregate data to feed the functional module with composed pointers of concession and other trouble-related data for further data processing and analysis.

Social media harvesting can be conducted through manual methods or by utilizing tools, but with the advancements in technology, AI-based social media harvesting has emerged as a more efficient and precise approach. The rapid growth of social media platforms has established them as essential marketing channels for brands, considering that 75% of the global population eligible to use social media are active users. Furthermore, statistics reveal that a significant portion of Twitter users falls within the age group of 25 to 34 years.

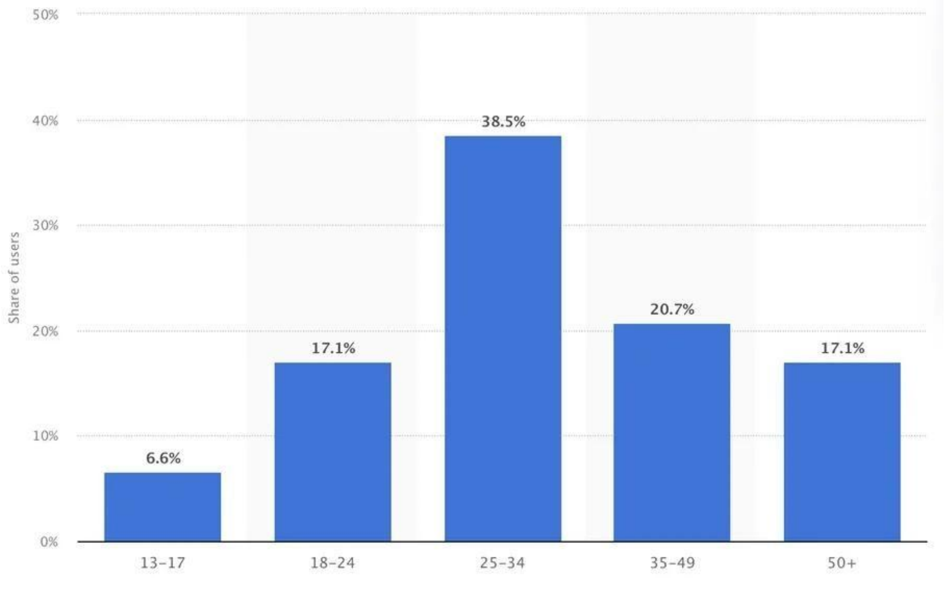


Fig 1: Twitter Users by Age

Protecting cyberspace and the interests of the state and its people while upholding individual rights to express themselves on social media sites is a complex challenge. Monitoring tools can be developed to track and analyze social media activity for harmful speech while adhering to legal and ethical standards. Education and awareness programs can also promote responsible social media use. Striking the right balance is crucial to safeguarding both free expression and the safety of individuals and society.

1.2 Problem Statement

If an anti-state tweet/post/trend gets viral on any social media site, it is very difficult for Law Enforcement Agencies (LEAs) to trace back to the root originator of this trend i.e., the person/s who started it. Also, businesses face challenges when it comes to digital and social media marketing moreover, they don’t know how their product is doing in the market among customers or how to launch a successful campaign strategy grounded on the analysis.

1.3 Proposed Solution

- Developing an Open-source intelligence (OSINT) tool to gather information from publicly available sources (such as Facebook, Instagram, and Twitter) to meet specific intelligence requirements.
- Extraction and Analysis of data from social media posts from target users (the users who have started trends against the state and public).
- From the collected data, our tool can help businesses keep track of market intelligence, customer intelligence, competitive analysis, and digital marketing and management. This helps in identifying weaknesses, and planning strategically for the future and thus ensures business growth.

1.4 Working Principle

The project mainly performs on the principles of web scraping and Twitter API integrated with machine learning algorithms. The work of our project has been done in the following steps:

- Identifying the social media platform to be harvested
- Data Collection
- Data Analysis
- Reporting and Visualization
- Graphical User Interface (GUI) building

1.5 Objectives

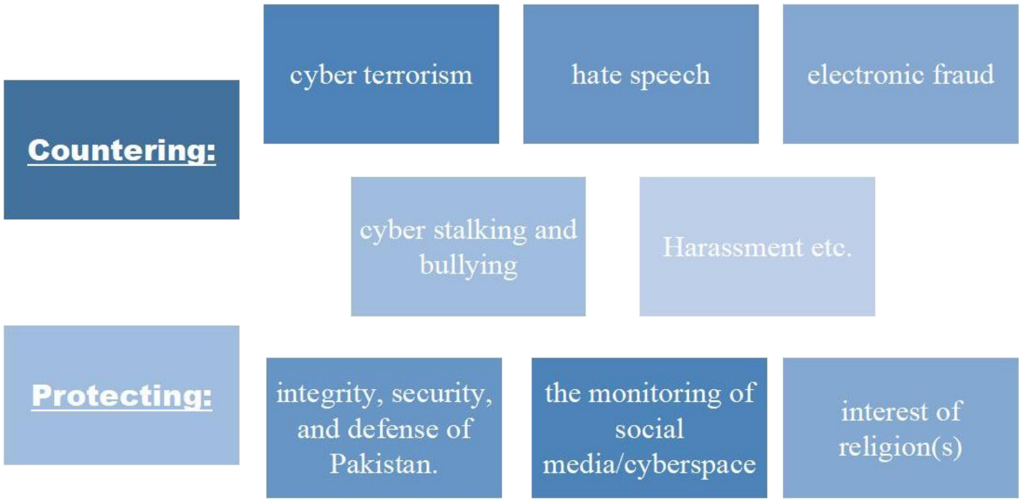
1.5.1 General Objectives:

Our project aims to design a software or tool to identify the originator from big data using the techniques of machine learning (ML) and natural language processing (NLP) to extract insights from the data.

1.5.2 Academic Objectives:

- To monitor and gather data sets of targeted individuals available on social media networks to trace back to the originator of the post/tweet or trend.
- To extract features and useful information from obtained data sets using predefined algorithms and design a tool to identify the root originator from big data.
- To gather information and insights from social media that can assist in decision-making and improve company strategies.

1.6 Scope


Countering:
cyber terrorism
hate speech
electronic fraud
cyber stalking and bullying
Harassment etc.
Protecting:
integrity, security, and defense of Pakistan.
the monitoring of social media/cyberspace
interest of religion(s)

1.7 Potential Impact

Spammers and bots on social media sites like Instagram and Twitter can be found more easily due to social media data mining. Government agencies are increasingly focusing their actions on welfare. It will be possible to access Twitter's raw data through its operating program interface (API), which will be mined for data. The information provided used algorithms created by the company and included measurements of influence as well as predictions of the likely gender of the Twitter stoner. Before analysis, tweets were manually reused on a tweet-by-tweet basis. All tweets, even irrelevant tweets, that had nothing to do with the miracle under discussion were deleted.

Every time a user views a tweet, a Twitter print is produced. The effectiveness of a hashtag or a tweet can be evaluated using Twitter printouts. These prints are accessible to the account owner only and can be found in the Twitter Analytics section of our Twitter profile. The prints of other individuals are invisible to us. Using the users' followers who mention the hashtag, we calculate the hashtag's popularity. Because we are unable to determine the precise number of people who viewed the tweet, these prints are known as "implicit prints." We will say that someone earned 1,000 impressions if they used the campaign's hashtag and had 1,000 followers.

Big data enables marketers to recognize social media trends and gather perceptions, which can be used to form ideas on involvement, such as which individuals to contact or which demographic should accept marketing emails, etc. It is useful to keep tabs on customer preferences and trends, particularly what visitors enjoy, what they do, and how it evolves. The tweets are gathered via the Twitter API in the proposed work, and counting styles and other machine-learning algorithms are used to discover trending patterns on Twitter. An assortment of styles that can be utilized to connect with the operation is provided by this API.

1.8 Deliverables

1.8.1 User Interface:

An easy-to-use interface that enables the contributors of the project to enter their needs and track the development of the data collecting and analysis.

1.8.2 Documentation:

Detailed explanations of the project's methodology, algorithms, and software are provided in the documentation.

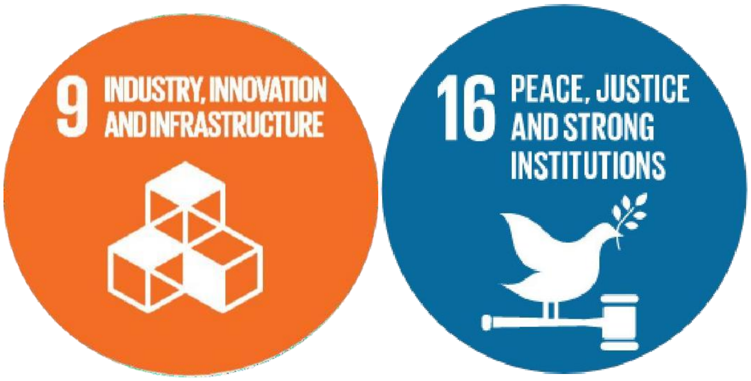
1.8.3 Training Resources:

Educational resources that can aid users in learning how to use the tools and software created for this project.

1.8.4 Maintenance and Assistance:

Ongoing services to ensure that the software and tools continue to operate correctly after the project is completed.

1.9 Relevant Sustainable Development Goals


9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
16 PEACE, JUSTICE AND STRONG INSTITUTIONS

GOAL 09: Figure flexible structure, promote inclusive and sustainable industrialization, and foster innovation. It also highlights the role of media in the fight against corruption.

GOAL 16: Promote peaceful and inclusive societies for sustainable development, give access to justice for all, and make effective, responsible, and inclusive institutions at all levels. It ensures public access to information and covers fundamental freedoms, in agreement with public legislation and transnational agreements.

1.10 Structure of Thesis

Chapter 2 briefly discussed the background terminologies and the previously published work.

Chapter 3 contains the design and methodology of the project.

Chapter 4 is about the experimental setup which is the practical implementation of our project.

Chapter 5 describes the conclusion of this project.

Future work directions are discussed in Chapter 6.

Chapter 2: Literature Review

The features of recently released, comparable items are modified and improved to create a new product. A literature review is a crucial stage in the creation of a new product. Finding out what is previously known about a given issue, as well as any knowledge gaps, is the main goal of a literature review. It entails a thorough study of the research's advantages and disadvantages, including its techniques and conclusions. Our research is divided into the following points.

- Industrial Background
- Existing methods and their shortcomings
- Research Papers

2.1 Industrial background

The way we connect with businesses has been completely transformed by the AI industry. The new industrial era, which places a premium on effective logistics above human error, is characterized by the widespread use of AI in practically all-important technologies.

Machine learning algorithms are used in AI-based social media harvesting to automatically gather and examine significant amounts of social media data, including tweets, posts, comments, and likes, to spot patterns and trends. Making better-educated decisions, such as altering marketing strategy, enhancing customer service, and spotting new business prospects, can be facilitated by this kind of data.

By employing AI to target and respond to particular groups that can affect the outcomes of your marketing efforts, social media bots are AI solutions that can accomplish this alone.

2.2 Existing Methods and their shortcomings

The subject of AI-based social media harvesting is fast developing, and there are numerous solutions already in existence that can be applied in this context. These solutions often gather and analyze social media data using machine learning algorithms to identify trends and insights. Following are some solutions that are formally being set and being applied.

- Scraping for Big Data Without Breaking the Bank
- Using Osint to Collect Information About a User from Various Social Networks
- Web Scraping vs. Twitter API: A Credibility Analysis
- Cyber Security for Social Networking Sites
- Identifying Cyber Security-Related Twitter Accounts and Sub-Groups

2.2.1 Scraping for Big Data Without Breaking the Bank

Big data is abundant, but the coding skills required to access it are not; useful data is frequently accessed by researchers and institutions paying pricey third-party social media "scraping" corporations, which hire programmers to create tools.[1]

This constrained research prospects and outraged the coding community because individual academics were forced to choose between manually scraping Twitter data one tweet at a time or paying pricey third-party companies. Twitter introduced its Application Programming Interface (API) to close this gap, enabling programmers and developers to filter through massive numbers of tweets by specifying particular words, phrases, and hashtags.

Users can apply through the Twitter developer portal using an active Twitter account. After being approved, the user gains access to the Twitter API developer level and can use authorization keys (passwords used in coding) to view the complete database history. One of the recognized coding languages and applications that support them can be used to create rules and functions for the data retrieval procedure. This free solution offers more precise information and more large data sets, potentially eliminating the need for expensive third-party tools.

2.2.2 Using Osint to Gather Information About a User from Multiple Social Networks

OSINT is the practice of reusing public information for intelligence and investigations, including social media content that is not protected by privacy settings. As social media users, we are beginning to notice the growth of specific social networking sites like Twitter and LinkedIn that offer datasets that both overlap and differ from those of more general-purpose social networking sites like Facebook.

Online social networks provide a wealth of information for OSINT, including social relationships, activities, and personal information about a person of interest. Nevertheless, contrary to popular belief, not all information on the internet is readily available. [2] There are many challenges for investigators, including limitations on platform and privacy, as well as the accessibility and lifespan of data.

Numerous social networking sites have included more privacy control features to limit access to personal data in response to growing privacy concerns. Massive amounts of data are collected on online social networks, yet the social media platform has strict control over how that data is released. Platforms for social networking typically limit the information that is sent based on social connections, user-based privacy settings, rate limiting, activity tracking, and IP address-based limitations. The social network is far from static; relationships routinely alter, and profiles are continuously updated. The validity of conditions like these is still up in the air, although many social networking sites, including Facebook, forbid the use of screen scrapers and other data mining tools in their terms of service. [2]

2.2.3 Web Scraping versus Twitter API: A Comparison for a Credibility Analysis

Currently, the majority of corporate processes, research, studies, and other activities depend on data extraction from web sources. The goal of data extraction is to gather important information that can be used for a variety of purposes. This has been done using three well-known techniques: web scraping, APIs, and manual extraction. The most efficient and practical methods of data collection are web scraping and APIs [24]. They make it possible to quickly and reliably collect data from diverse website pages and repositories. The information is then saved and kept for later usage and examination. Manual extraction takes longer and is more time-demanding due to human error.

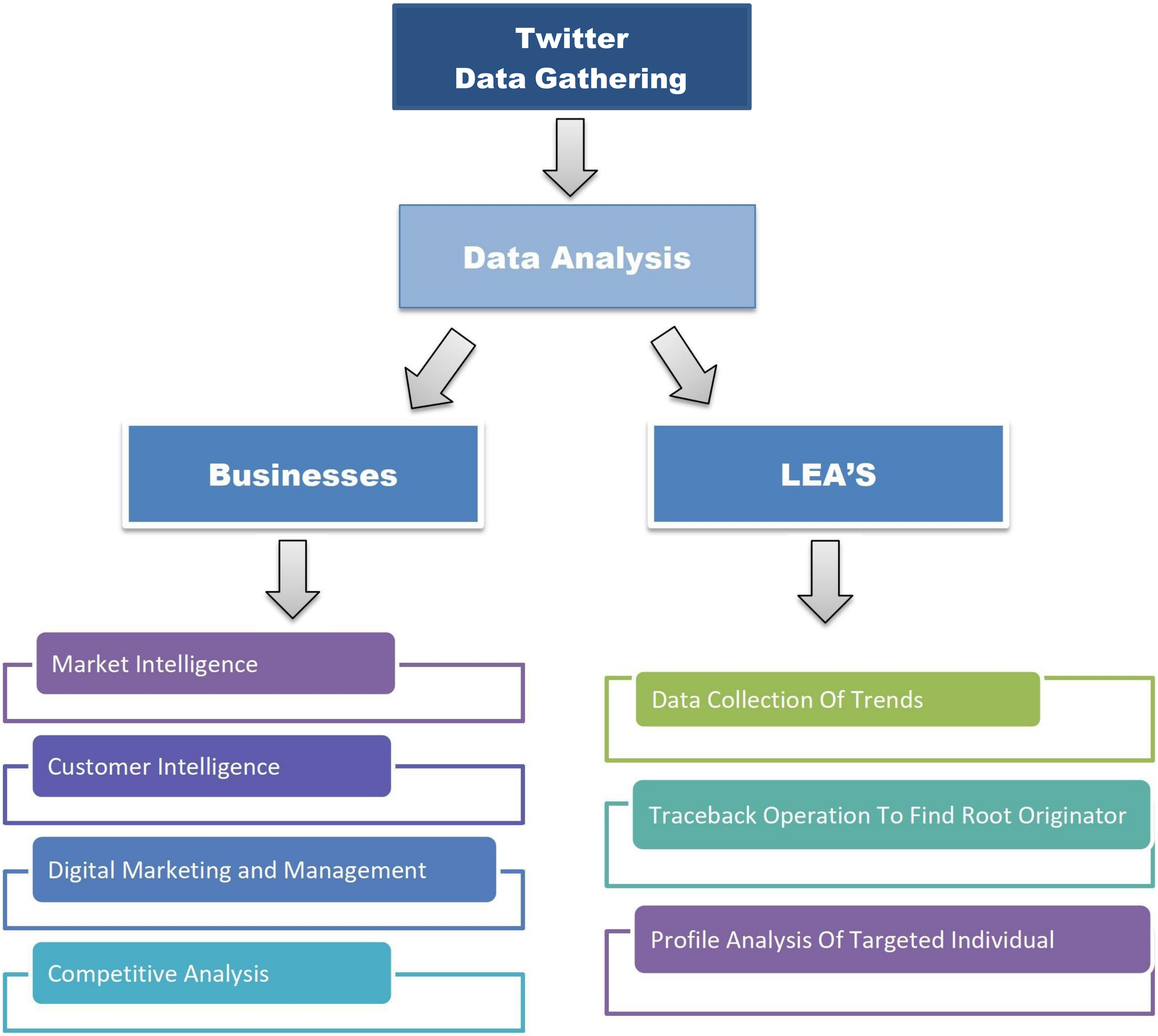
Web scrapers are helpful for swiftly obtaining and analyzing massive amounts of data from a particular website. A bot can therefore retrieve information that is shown in a browser to extract the data and store it in a database for later usage and study [20]. Since it influences when and how the data is shown, the network speed may be a barrier or disadvantage for web scraping. Web scraping is essential and may entail time-consuming programming activities. To avoid copyright infringement, the developers should take into account legal and policy considerations regarding the material they are retrieving [12].

Retweet, like, and date of publication data are all retrieved using the Twitter API. Also, the text is examined to count the words and characters, find stop words, etc. The information offered by Standard API may be sufficient for a project or developer's needs; however, this product's drawback is that access is limited to the previous seven days. Not all of the APIs that social networks provide are practical for data extraction. Yet, it has been shown that the Twitter API is simple to use; by simply knowing the tweet ID or user ID, a range of information can be obtained.

The researchers in [16] use web scraping to quickly get around the Twitter API's date range restrictions to retrieve an infinite number of tweets. The comparison demonstrates that the total number of tweets retrieved via the Twitter API is consistently lower than when using Twitter Scrapy. Although the majority of solutions collect data using the Twitter streaming API, researchers state that there is a limitation when searches go beyond rating intervals and time ranges.

Chapter 3: Proposed Methodology

The social media harvesting process will be determined by the project's specific aims and objectives. It is important to highlight that the methodology must comply with the terms of service of social media platforms and must protect the privacy of users. Because nothing is sure in OSINT, our solution does not entirely rely on AI; we have also performed some manual work.


Twitter
Data Gathering
Data Analysis
Businesses
LEA'S
Market Intelligence
Customer Intelligence
Digital Marketing and Management
Competitive Analysis
Data Collection Of Trends
Traceback Operation To Find Root Originator
Profile Analysis Of Targeted Individual

3.1) Twitter Data Gathering

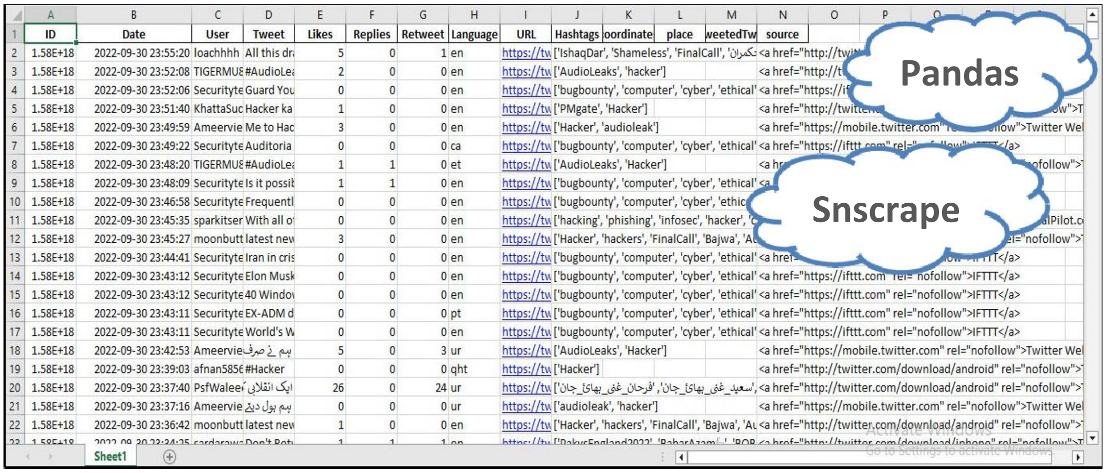
The process of collecting tweets, user profiles, and other information from the Twitter social media network is known as Twitter data gathering. This can be accomplished using a variety of approaches.

Data extraction from Twitter via API has several limitations: -

- ➢ Tweet Limitation (if we use APIs, we can only extract a certain number of tweets)
- ➢ Time restriction (obtaining a large data collection will take a long time).

Data Extracting from Twitter by web scraping: -

Underestimation of the limitations of Twitter API data extraction via web scraping is done using many Python libraries, such as Twitter-scraper and Twint, but these are rejected by Twitter. Because Twitter's revised policies disallowed the old libraries, data extraction from Twitter policies is done using a more updated Python module called **SNSCRAPE**. It can scrape data from public social media accounts without requiring authentication, making it a valuable tool for data researchers who need to collect vast amounts of social media data for the study. It is great for scraping Twitter tweets, user profiles, and hashtags.



ID	Date	User	Tweet	Likes	Replies	Retweet	Language	URL	Hashtags	oordinate	place	weetedTw	source
1.58E+18	2022-09-30 23:55:20	loachhhh	All this dr	5	0	1	en	https://tw	['IshaqDar', 'Shameless', 'FinalCall', 'حکمران				<a href="http://twi
1.58E+18	2022-09-30 23:52:08	TIGERMUS	#AudioLe	2	0	0	en	https://tw	['AudioLeaks', 'hacker']				<a href="http://t
1.58E+18	2022-09-30 23:52:06	Securityte	Guard You	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://if
1.58E+18	2022-09-30 23:51:40	KhattaSuc	Hacker ka	1	0	0	en	https://tw	['PMgate', 'Hacker']				<a href="http://twitte ow">T
1.58E+18	2022-09-30 23:49:59	Ameervie	Me to Hac	3	0	0	en	https://tw	['Hacker', 'audioleak']				<a href="https://mobile.twitter.com" re ollow">Twitter Wel
1.58E+18	2022-09-30 23:49:22	Securityte	Auditoria	0	0	0	ca	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://ifttt.com" rel= </a>
1.58E+18	2022-09-30 23:48:20	TIGERMUS	#AudioLe	1	1	0	et	https://tw	['AudioLeaks', 'Hacker']				<a hr ofollow">
1.58E+18	2022-09-30 23:48:09	Securityte	Is it possit	1	1	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a
1.58E+18	2022-09-30 23:46:58	Securityte	Frequentl	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethic				
1.58E+18	2022-09-30 23:45:35	sparkitser	With all o	0	0	0	en	https://tw	['hacking', 'phishing', 'infosec', 'hacker', 'c				alPilot.c
1.58E+18	2022-09-30 23:45:27	moonbutt	latest new	3	0	0	en	https://tw	['Hacker', 'hackers', 'FinalCall', 'Bajwa', 'A				el="nofollow">
1.58E+18	2022-09-30 23:44:41	Securityte	Iran in cris	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href ow">IFTTT</a>
1.58E+18	2022-09-30 23:43:12	Securityte	Elon Musk	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://ifttt.com" rel="nofollow">IFTTT</a>
1.58E+18	2022-09-30 23:43:12	Securityte	40 Windo	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://ifttt.com" rel="nofollow">IFTTT</a>
1.58E+18	2022-09-30 23:43:11	Securityte	EX-ADM d	0	0	0	pt	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://ifttt.com" rel="nofollow">IFTTT</a>
1.58E+18	2022-09-30 23:43:11	Securityte	World's W	0	0	0	en	https://tw	['bugbounty', 'computer', 'cyber', 'ethical'				<a href="https://ifttt.com" rel="nofollow">IFTTT</a>
1.58E+18	2022-09-30 23:42:53	Ameervie	ہم نے صرف	5	0	3	ur	https://tw	['AudioLeaks', 'Hacker']				<a href="https://mobile.twitter.com" rel="nofollow">Twitter Wel
1.58E+18	2022-09-30 23:39:03	afnan585	#Hacker	0	0	0	qht	https://tw	['Hacker']				<a href="http://twitter.com/download/android" rel="nofollow">
1.58E+18	2022-09-30 23:37:40	PsfWalee	ایک انقلابی	26	0	24	ur	https://tw	['سعید_غنی_بھائی_جان', 'فرحان_غنی_بھائی_جان']				<a href="http://twitter.com/download/android" rel="nofollow">
1.58E+18	2022-09-30 23:37:16	Ameervie	ہم بول دیئے	0	0	0	ur	https://tw	['audioleak', 'hacker']				<a href="https://mobile.twitter.com" rel="nofollow">Twitter Wel
1.58E+18	2022-09-30 23:36:42	moonbutt	latest new	1	0	0	en	https://tw	['Hacker', 'hackers', 'FinalCall', 'Bajwa', 'Au				<a href="http://twitter.com/download/android" rel="nofollow">
1.58E+18	2022-09-30 23:34:25	sardarawa	Don't Rot	1	1	1	en	https://tw	['PakvsEngland2022', 'BabarAzam', 'PCB				<a href="http://twitter.com/download/iphone" rel="nofollow">T

Fig 2: Data Gathering by SNSCRAPE

3.1) Data Analysis

This step entails mining data for insights using techniques such as natural language processing (NLP) and machine learning (ML). The data collected can be used to assess client preferences.

Many strategies are used to investigate the behavior of the data, such as examining the community's response to the hashtag, determining whether the hashtag is controlled by a real or bot account, and examining the people's comments.

Data Analysis can be further divided into two steps: -

3.2.1) Business Dimensions:

I. **Market Intelligence: -**

Insights into consumer interests and behavior gained from Twitter data can be utilized to guide marketing and advertising initiatives. Businesses can create specific marketing strategies that resonate with their target audience by analyzing Twitter data to help them gain a competitive advantage.

II. **Customer Intelligence: -**

Customer intelligence is the practice of utilizing Twitter data to understand customer habits, interests, and viewpoints. Businesses can use customer service to track and address customer questions and concerns, which can help them better understand their consumers, promote customer engagement, and improve marketing initiatives. As a result, consumer loyalty and satisfaction may increase.

III. **Digital Marketing and Management: -**

Twitter data, including hashtags, themes, and keywords, can be utilized to determine the kinds of material that resonate with viewers. Businesses can use this information to improve their content strategy and produce more interesting and appropriate material. Real-time monitoring of the effectiveness of digital marketing initiatives can be done using Twitter data. Businesses can use this data to assess the effectiveness of their efforts and make data-driven decisions.

IV. **Competitive Analysis: -**

A competitor's performance and strategy can be discovered via Twitter data. Businesses might find ways to set themselves apart from rivals and obtain a competitive advantage by analyzing Twitter data. Twitter data can be used to locate and keep track of rivals in a given market or sector. By keeping up with the most recent trends and modifying their strategies as necessary, businesses can benefit from this information. Tweets from Twitter can be used to track client feedback against rival businesses. Businesses can use this data to pinpoint client problem areas and opportunities to enhance their goods and services.

3.2.2) Law Enforcement Agencies (LEA'S):

I. **Data Collection of Trends: -**

Data collection on Twitter trends is the procedure of gathering and examining information on Twitter about new trends, hashtags, and topics that may be important to law enforcement actions. LEAs can utilize this data to spot and follow expected threats, keep track of public opinion, and enhance their general situational awareness.

II. **Traceback Operation to Find Root Originator: -**

When a hashtag or trend on Twitter goes viral, it indicates that a lot of people are using it or talking about it. This frequently results in an increase in the number of tweets and retweets on that subject. This increase in activity can be a sign that a proliferator is spreading the hashtag or trend on purpose to attract attention and manipulate public opinion. Likewise, it may also point to the involvement of a proliferator if the text or content linked to the hashtag or trend looks to be planned out or scripted.

To safeguard public safety and uphold the law, LEAs can identify accounts that may be involved in spreading the hashtag or topic by analyzing the network of accounts that are using it. They can then take the necessary measures.

III. Profile Analysis of Targeted Individual: -

The behavior, interests, and preferences of a target individual can be learned by looking at their Twitter profile. To evaluate the profile of a specific person and to spot patterns and trends in their behavior, look at the substance of their tweets. Take into account elements like the timing and frequency of their tweets, as well as the tone of their communication. Investigating the target person's fan base to learn more about their network and the kinds of individuals who interact with their content. Finding active users who interact with the material of the targeted person or who tweet about them. Understanding who is influencing conversations on Twitter about the targeted person and how their messaging is propagated can help.

3.2) Uniqueness of Our Approach

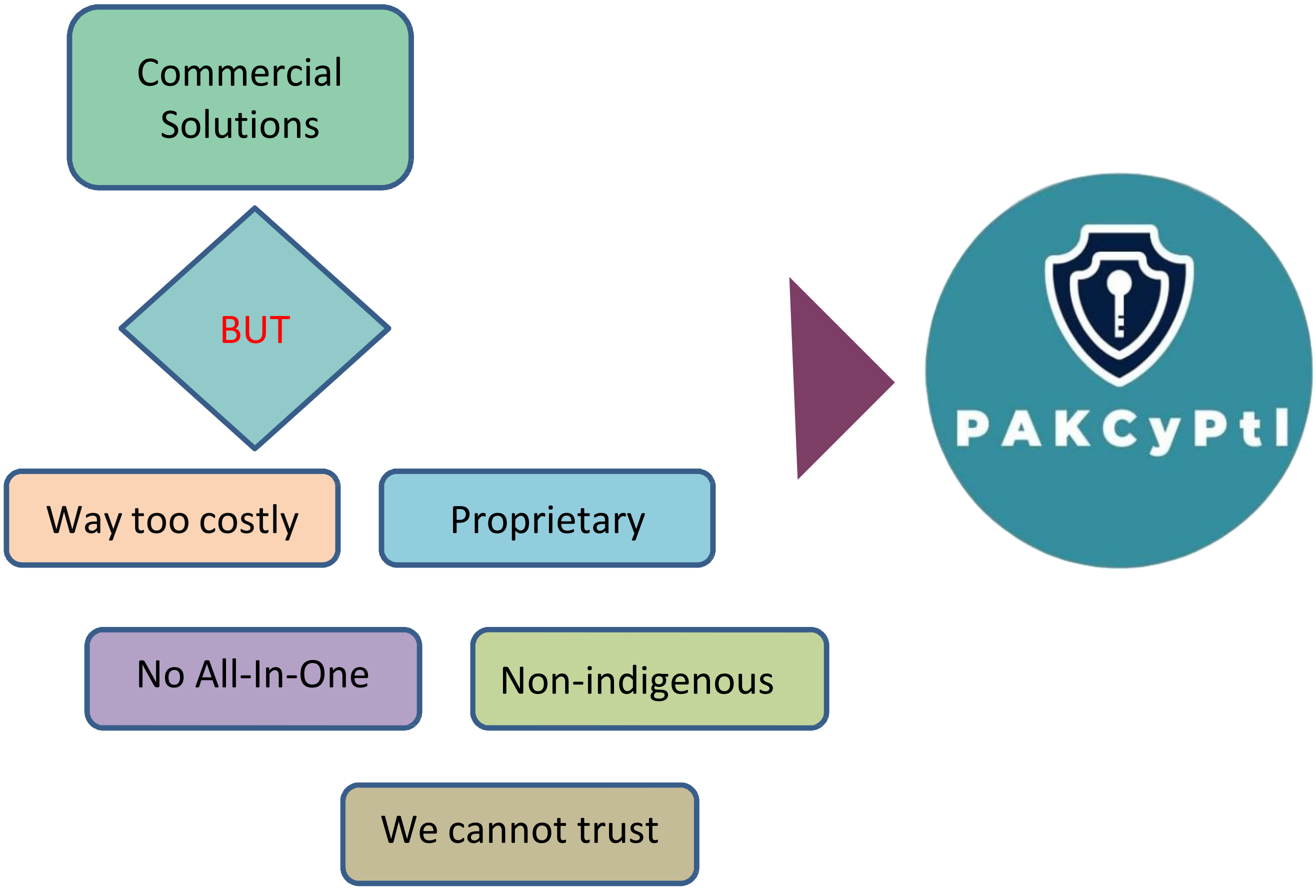


Fig 3: Uniqueness of Our Approach

Several commercial solutions are offered domestically and internationally for the problem statement we stated at the beginning, but there are some drawbacks to those solutions that eventually make our solution the better one in many respects.

(1) They are way out of reach for enterprises and organizations because of their high cost and fluctuating currency conversion rates. For instance, a person making rupees would prefer not to have his business analysis done in dollars.

(2) They are frequently proprietary, which makes them more challenging to use.

(3) No "all-in-one" solution exists. A wide range of data, analyses, and visualizations are available from us. Because we have functionalities ranging from obtaining large data sets to competitive analysis, profile analysis, hashtag analysis, and much more all in one place for ease of access and feasibility, it can be used by businesses as well as law enforcement agencies, researchers, or academic personnel doing some data-related research.

(4) The solutions offered internationally are not indigenous. Therefore, it becomes challenging, particularly for law enforcement or government organizations, to trust those third-party tools and analyze sensitive data or information on them, so there is always a trust component. Because of this, there was an urgent need for a locally applicable indigenous remedy. Additionally, compared to a person sitting thousands of miles away from a certain location, a local developer would understand things and problems much better and integrate features according to them in his or her solution.

PAKCyPtl steps in to help in this situation.

3.3) Novelty

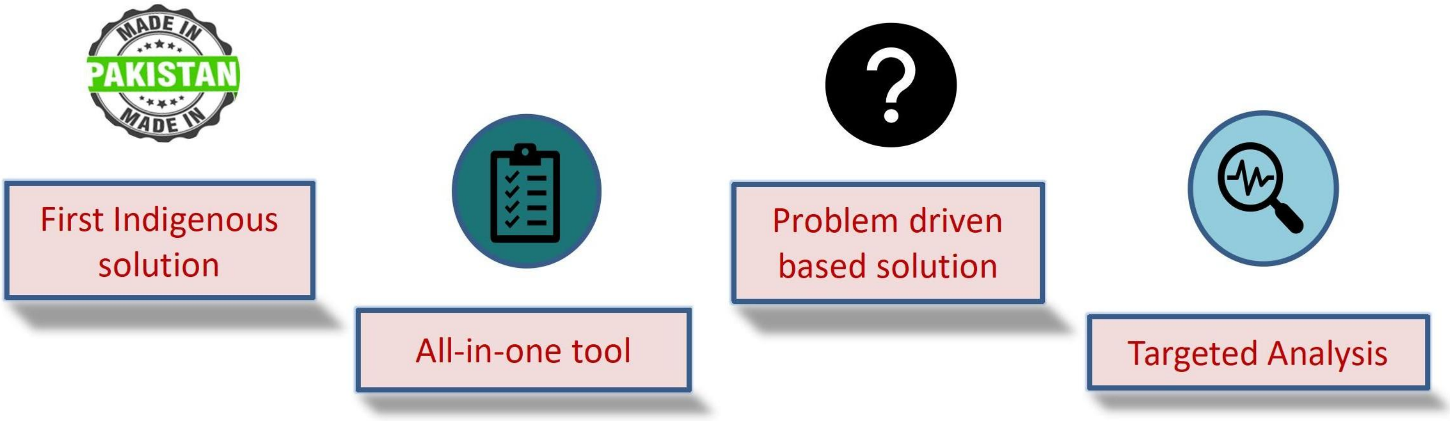


Fig 4: Novelty of our Project

An AI-based social media harvesting solution’s unique selling point is its capacity to automate, streamline, and enhance the collection and analysis of social media data while also offering deeper and individualized analytics capabilities. Additionally, our methodology can be adjusted to meet the needs of the user and certain research goals. This makes it possible to collect and analyze data in a more focused and relevant manner, which can result in more useful insights and well-informed decision-making.

Our project provides indigenous solution which means it does not involve third parties instead it is developed locally in the same country as the end users. Also, the solution we are providing is different from others in that it satisfies all of your needs on a single platform, like data gathering, data analysis, integration of open-source (OSINT) tools, making decisions based on data analysis, hashtag analysis, profile analysis, and sentimental analysis.

By providing complete, comprehensive data analytics on a single platform, our web page assists businesses that want to analyze social media and law enforcement agencies that would like to identify anti-state components, respectively.

3.4) Value Proposition

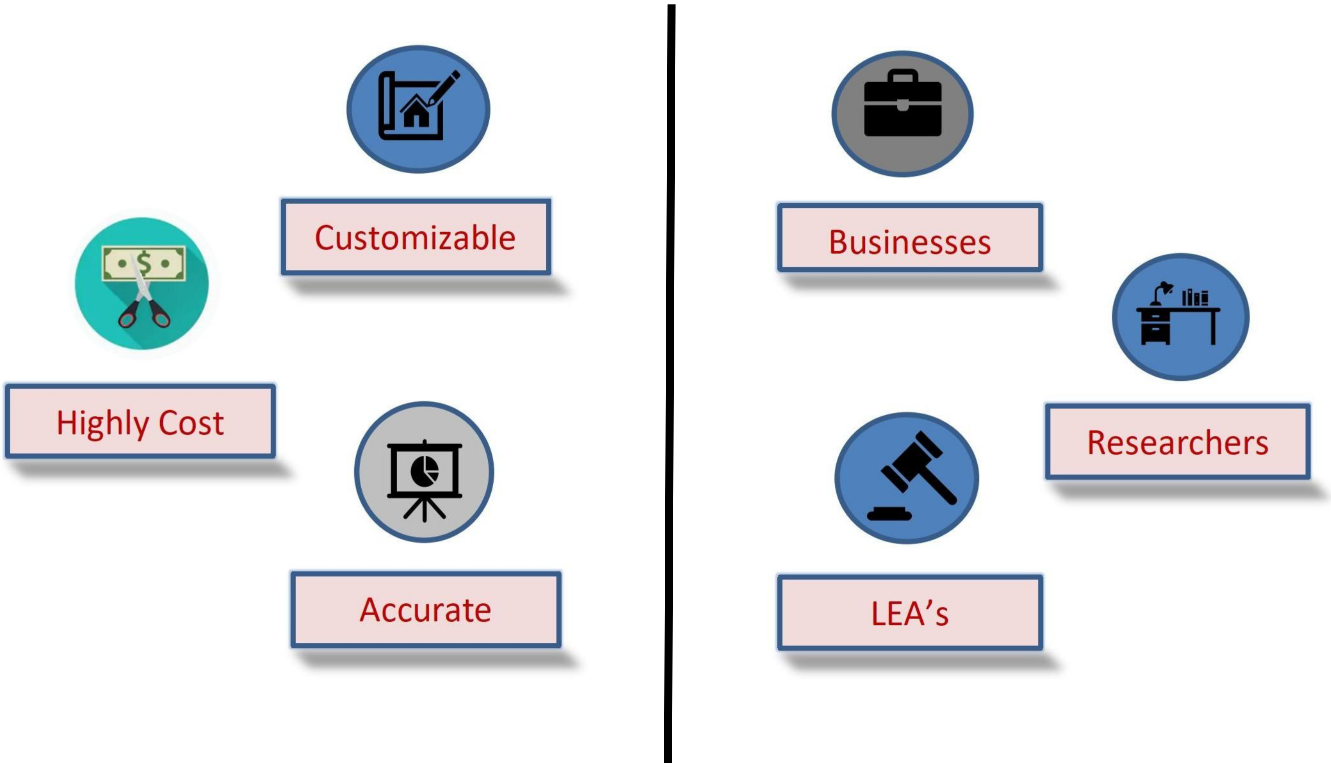


Fig 5: Our Project's Value Proposition

Our project's value proposition is that it can analyze social media data more quickly, accurately, and insightfully than standard manual analysis techniques. As a result, enterprises, organizations, and other stakeholders may be able to make better decisions, gain a better understanding of customer feedback, and achieve better results. When compared to manual analysis techniques, an AI-based solution may automate most of the data collection and analysis process, saving both time and money.

Our proposed tool has several distinguishing characteristics that offer it a competitive advantage over other tools of a similar kind. It is a useful tool for companies, researchers, and anybody else wanting to understand Twitter trends and sentiment because it offers more focused analysis, precise insights, and customization choices. Overall, it will be a cost-effective solution.

Chapter 4: Experimental Setup

4.1) Practical & Technical Working

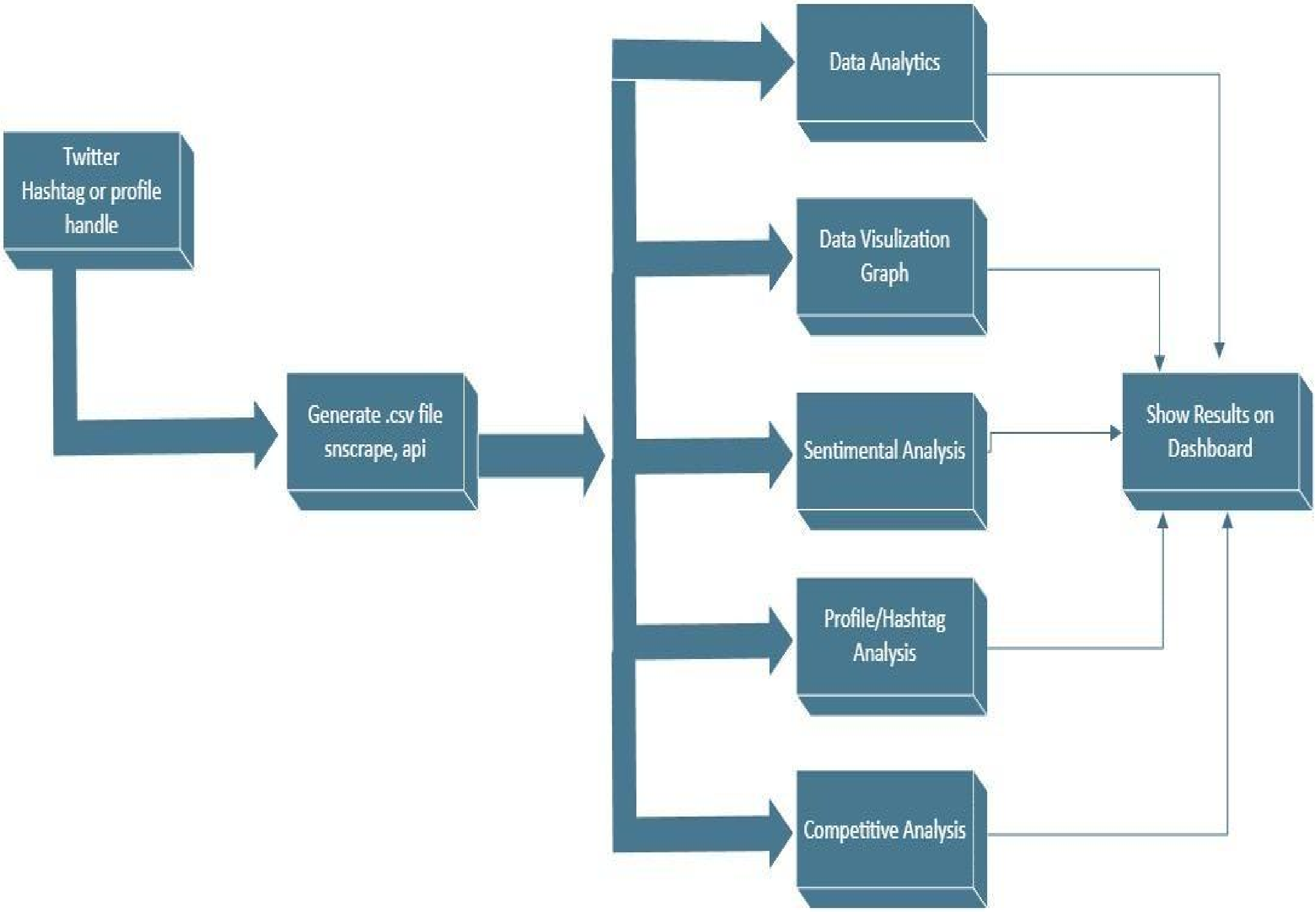


Fig 6: Flow Diagram of our Technical Working

A Python package called **Snscrape** enables users to quickly collect data from Twitter, including tweets with a certain hashtag or from a particular profile handle. To locate tweets with a specific hashtag, we first searched for the hashtag on Twitter and copied the URL from the address bar into the Python query. The URL can then be used to retrieve tweets with the relevant hashtag using Snscrape.

The tweets have been converted to a CSV file and placed in a Data Frame for later analysis after being retrieved. Overall, Snscrape enables Python users to quickly gather Twitter data based on particular hashtags or profile handles, facilitating effective data analysis and research.

The following analysis is performed after conversion to a CSV file.

4.2) Data Analysis

We have performed data analysis from both a Business and an LEA's perspective.

From a business perspective, we have examined customer attitudes, preferences, and behavior. Businesses can learn what their customers are saying about their goods and services by gathering and studying tweets that are relevant to their brand. They can use this to find fresh prospects for innovation, as well as to enhance their marketing strategy and client service.

Our data analysis can also be utilized to keep track of competitors and industry trends, which enables firms to learn more about the marketing tactics, product remarks, and customer involvement of their rivals.

From the perspective of a LEA, we have found propaganda networks, false profiles, coordinated campaigns, as well as the main sources of that propaganda. This can assist law enforcement agencies (LEAs) in identifying prospective suspects, reasons, connections, and ways to stop criminal conduct before it happens, all of which will help them develop cases against the perpetrators and bring them to justice.

4.3) Data Visualization Graph

Data analysis requires the use of data visualization. Using graphs and charts, you can gain insights into the data's trends, patterns, and relationships. The Python libraries used for data visualization include **Matplotlib** and **Plotly**.

For producing static visualizations, such as line plots, scatter plots, bar charts, histograms, and more, **Matplotlib** is a favored package. It offers a broad range of customization possibilities for visualizations and can be applied in a variety of situations, including data analysis, corporate intelligence, and scientific studies.

Interactive visualizations are made using **Plotly**. Compared to Matplotlib, it offers enhanced functions like panning, zooming, and hover effects. Web-based apps can use Plotly to produce dynamic visualizations.

We have generated the following graphs to display on our dashboard: -

- **Time-series graphs** are often used to analyze trends and patterns since they exhibit data across time. These can also be used to display the number of tweets associated with a given hashtag or topic over time.
- **Bar graphs** are used to compare different types of data. To compare the number of tweets related to different hashtags or subjects, bar charts can be utilized for data analysis.
- **Word clouds** are visual representations of text data, with the size of each word signifying its frequency of occurrence in the data. They can be used to display the most frequently used terms or hashtags associated with a certain topic.
- **Network graphs** are used to visualize the relationships between various things. They can be used to show the relationships between individuals or hashtags within a particular conversation or topic.
- **Heat maps** are used to show the concentration of data in a given area. They can be used to depict how tweets about a particular subject or hashtag are distributed geographically.
- **Pie charts** are very helpful for classifying data that can be simply displayed, and they can be used to spot patterns and trends in the data. For instance, you may create a pie chart to

display the percentage of tweets that reference various hashtags or keywords when examining tweets connected to a particular event or topic.

- It is possible to see how the frequency of tweets connected to a specific topic or hashtag changes over time in a **line graph** by connecting the data points with a line.

4.4) Hashtag Analysis

Both businesses and Law Enforcement Agencies (LEAs) can benefit greatly from hashtag analysis on Twitter. The usage of hashtags has grown in popularity as a tool for users to group and categorize their tweets about a specific subject, making it simpler for other users to discover and join the conversation.

Hashtag research can provide a plethora of information about a company's performance and how its clients view it. This study might offer insightful information about how well a company performs on social media. Businesses can track brand awareness, customer reaction, influencer identification, competitor analysis, and social media activity by tracking relevant hashtags. For instance, a chain of restaurants may track hashtags relating to food and dining to learn which dishes or dining experiences are most popular among customers. They can then apply this data to their menu selections or marketing initiatives.

Law enforcement organizations can track hashtags associated with illegal activities like drug trafficking or cybercrime to obtain information and find potential perpetrators. To monitor the situation and take appropriate action, they can also follow hashtags relating to public safety issues, such as protests or crises. For instance, to learn about the location and scope of a protest, keep an eye out for any threats or acts of violence, and prepare an appropriate reaction, a police department would follow hashtags associated with that event.

The graphs that we generated to display it on our dashboard under hashtag analysis are as follows:

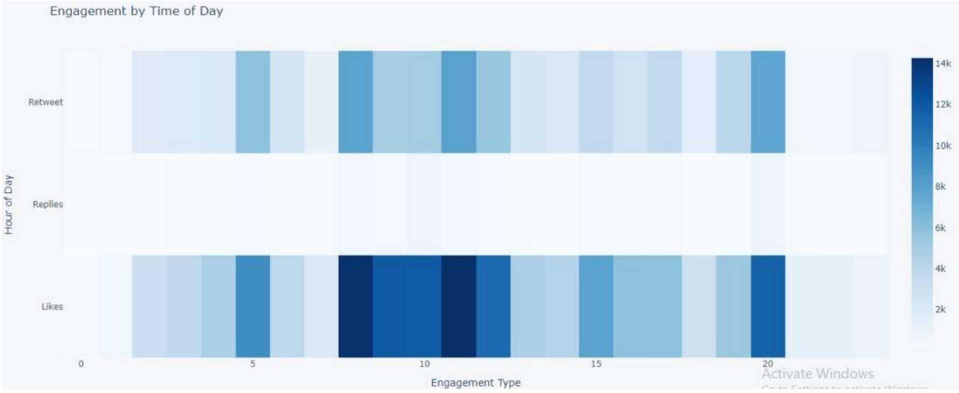
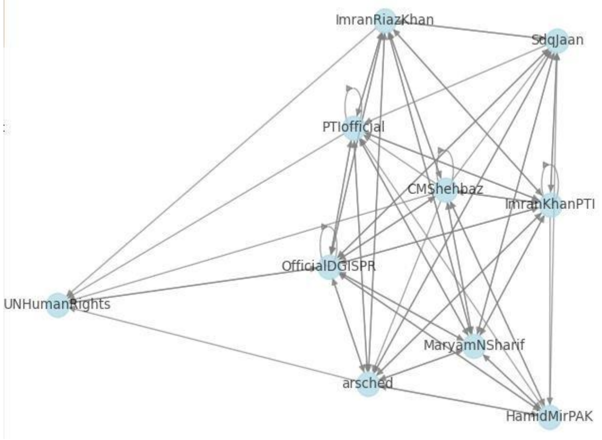


Fig 7: Graph of Heatmap of Retweets, Likes, and Replies



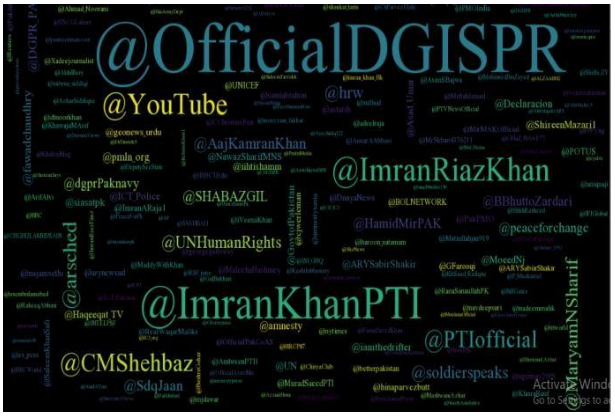


Fig 8: Influential Mentions

Fig 9: Generation of a Work cloud

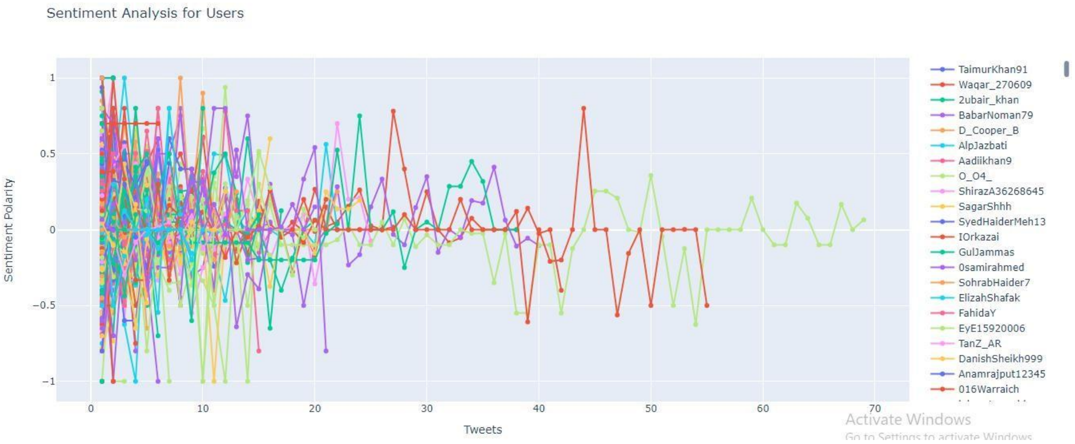


Fig 10: Sentiment Per User

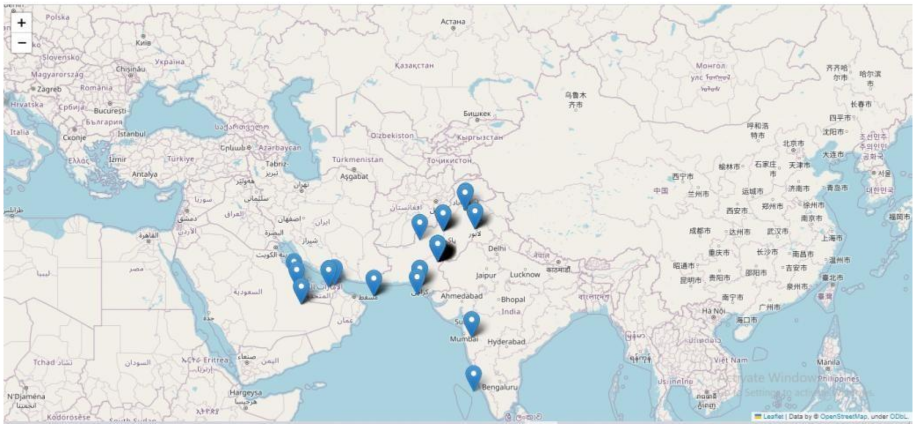
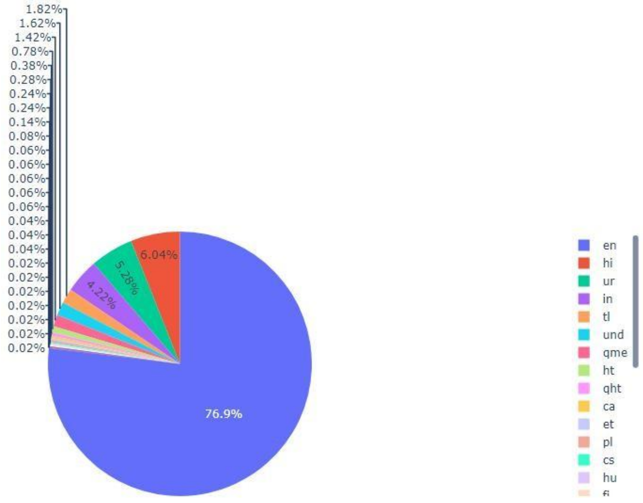


Fig 11: Coordinates on Map



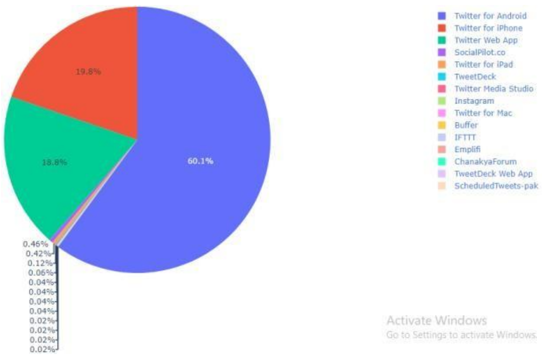


Fig 12: Pie Chart of Languages

Fig 13: Percentage of Different Sources

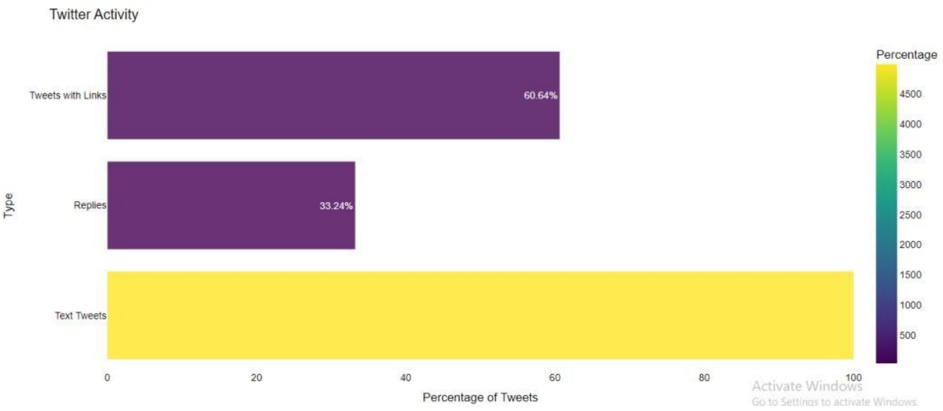


Fig 14: Text Tweets, Tweets with Links, Replies

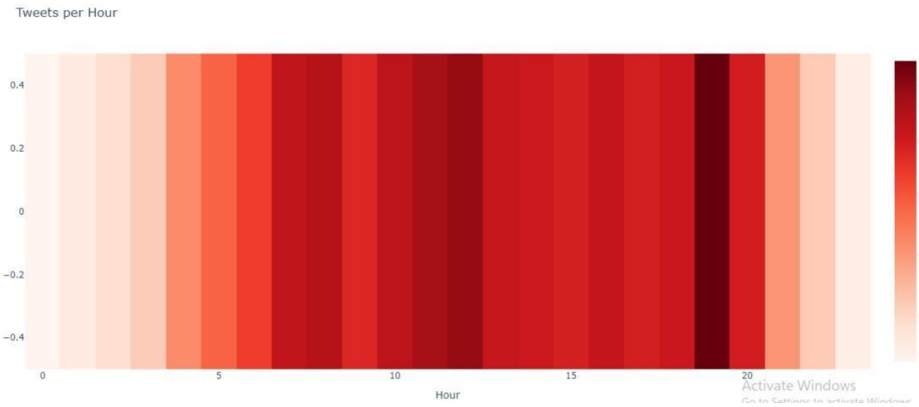
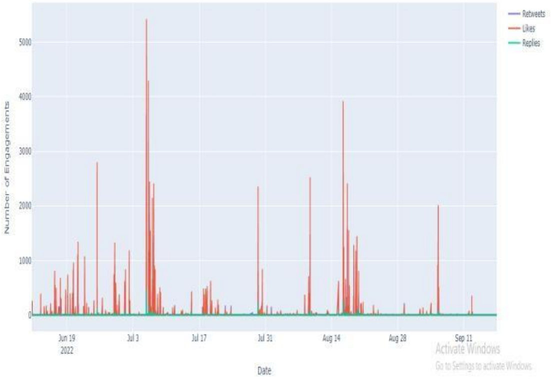


Fig 15: Heatmap of Tweets Per Hour



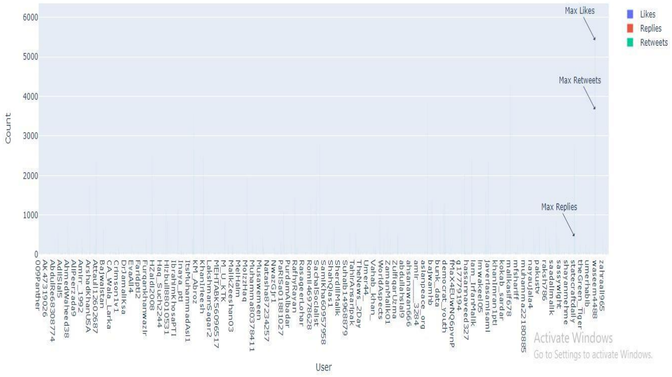


Fig 16: Twitter Engagement Over Time

Fig 17: User Engagement Metrics

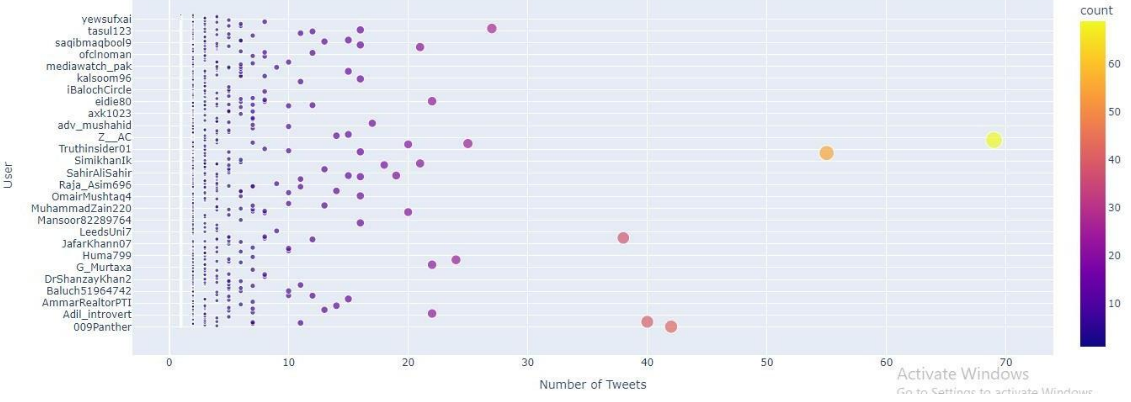


Fig 18: Number of Tweets by User

4.5) Sentimental Analysis

Businesses and LEA’s can benefit greatly from the information that sentiment analysis on Twitter can offer, which will help them make better decisions, run more efficiently, and maintain public safety.

Sentiment analysis is useful for businesses because it allows them to track and comprehend customer feedback, market trends, and brand perception. They can monitor the tone of tweets mentioning their name and goods to see any potential problems as well as positive and negative feedback. This might make it easier for businesses to respond to unfavorable reviews and enhance their goods and services. They can learn about the strengths and weaknesses of their rivals by keeping an eye on the sentiment of tweets mentioning them, which could benefit them in creating competitive strategies.

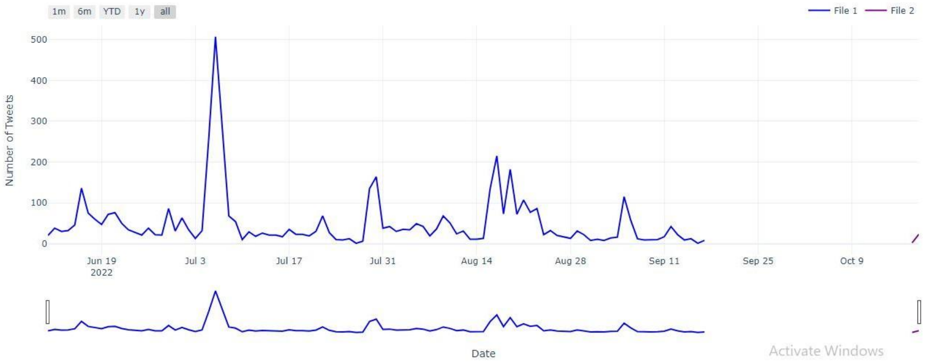
LEA’s can track public sentiment, identify potential risks, and acquire intelligence by using sentiment analysis. They can learn more about public opinion and potential responses by examining the sentiment of tweets around a specific event or subject. This could assist them in preparing for and handling any demonstrations, riots, or other forms of instability in society.

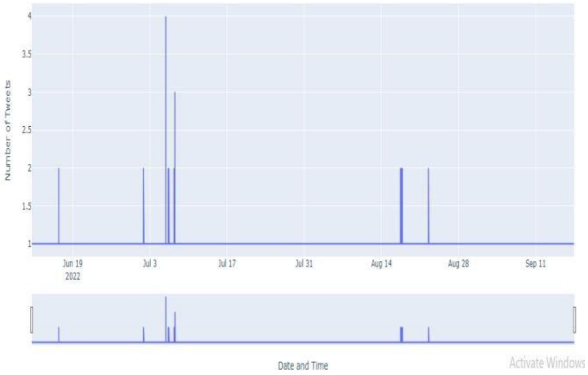
Also, LEAs can obtain information on unlawful organizations and operations by examining social media posts relating to criminal behavior. This information can help them create and carry out efficient law enforcement tactics.

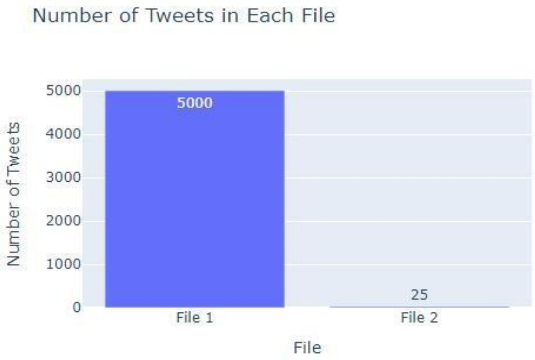
4.6) Competitive Analysis

Monitoring competitors’ social media activity to acquire insights into their strategies, strengths, and shortcomings is part of competitive analysis. Competitive analysis on Twitter can assist businesses in identifying competitors, analyzing competitors’ social media activity, tracking opponents’ brand image, and identifying opportunities. Businesses can identify competitors and analyze their social media strategy by monitoring Twitter activity relating to their sector or line of products. This includes the type of material they share, the frequency of their tweets, and their engagement rates. From the viewpoint of law enforcement, a competitive analysis can be used to track organized crime networks and operations. Monitoring social media activity related to illegal organizations or groups, as well as tracking the social media activity of individuals who may be involved in unlawful conduct, can fall under this classification.

We conducted a competitive study of tweets on Twitter and got the graphs below: -





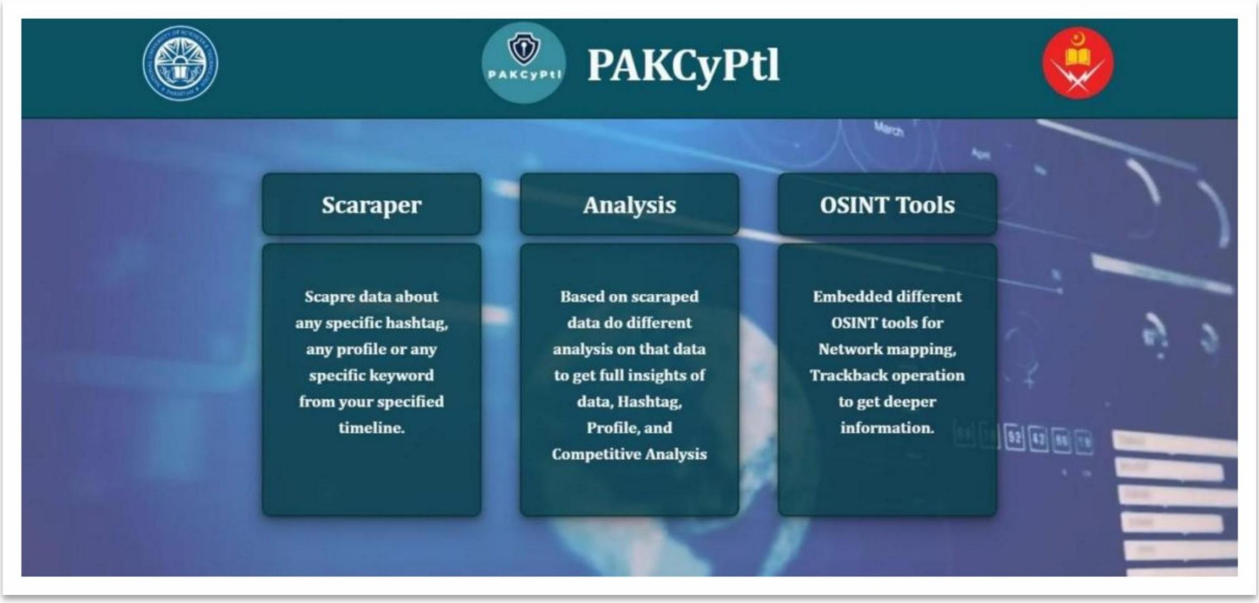


4.7) **Trackback Operation**

In our dashboard or GUI, we have incorporated different open source (OSINT) tools like Talkwalker, Hoaxy, Twitonomy that may be used to capture and analyze data from Twitter, offering essential insights for performing a trackback operation. The operation requires the use of data scraping tools to collect tweets containing specified keywords, hashtags or mentions that you want to track. These can be on a particular topic, event, or brand.

4.8) **Graphical User Interface (GUI) Building**

All of the data scraping, above-mentioned analysis findings, and open source (OSINT) tools have been integrated into a GUI that we have developed.



Scapre data about any specific hashtag, any profile or any specific keyword from your specified timeline.

Based on scaraped data do different analysis on that data to get full insights of data, Hashtag, Profile, and Competitive Analysis

Embedded different OSINT tools for Network mapping, Trackback operation to get deeper information.

Fig 19: Dashboard

Results

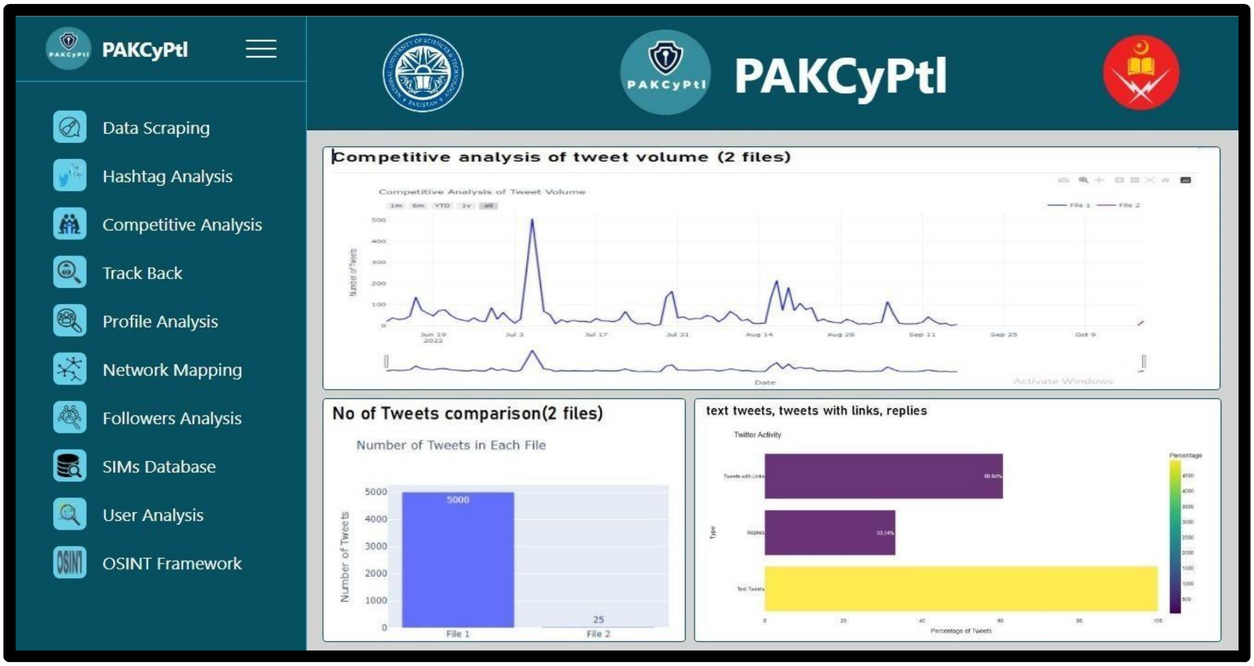


Fig 20: Displaying of Competitive Analysis

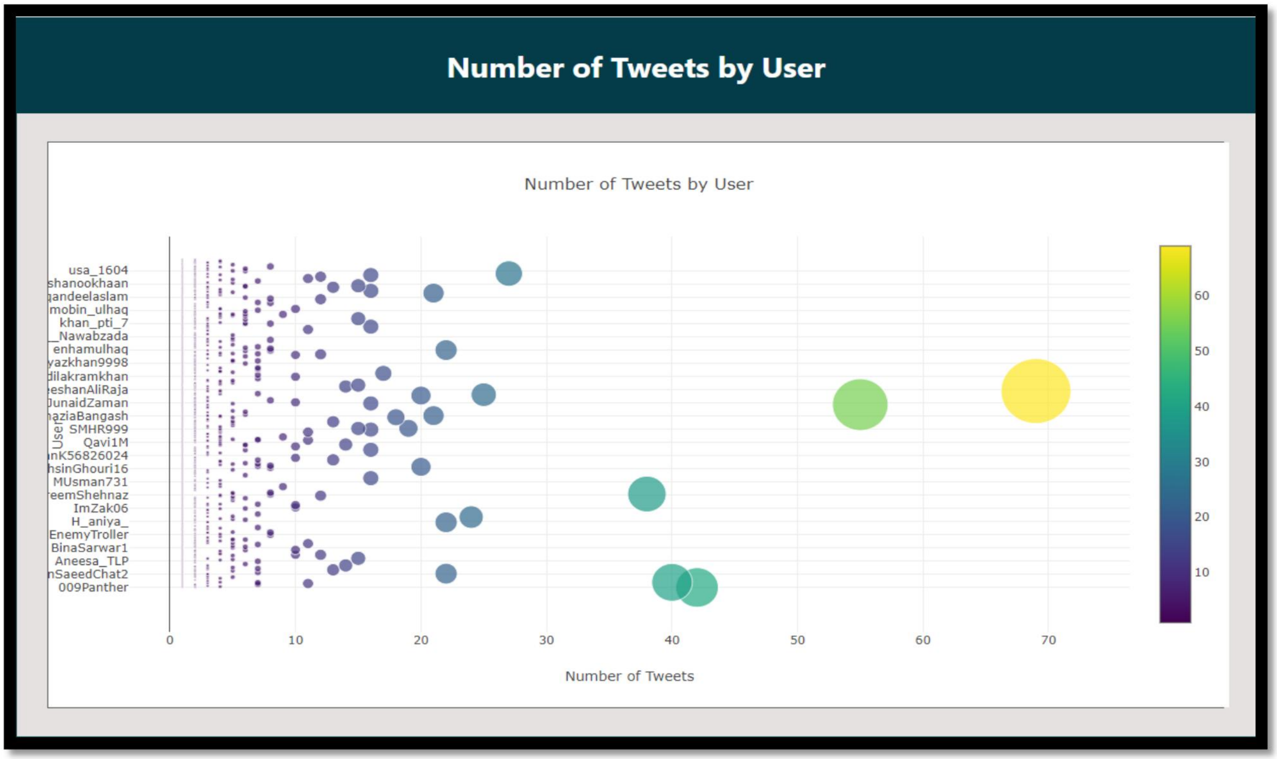
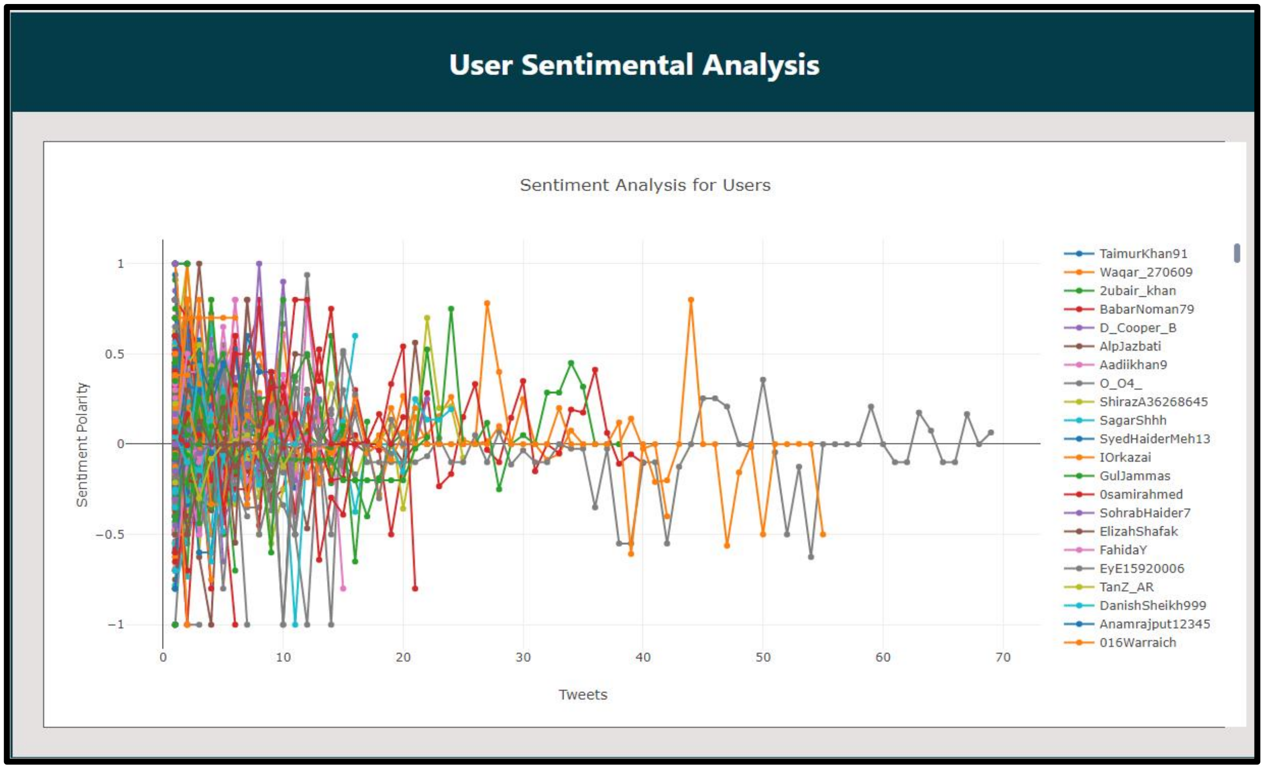
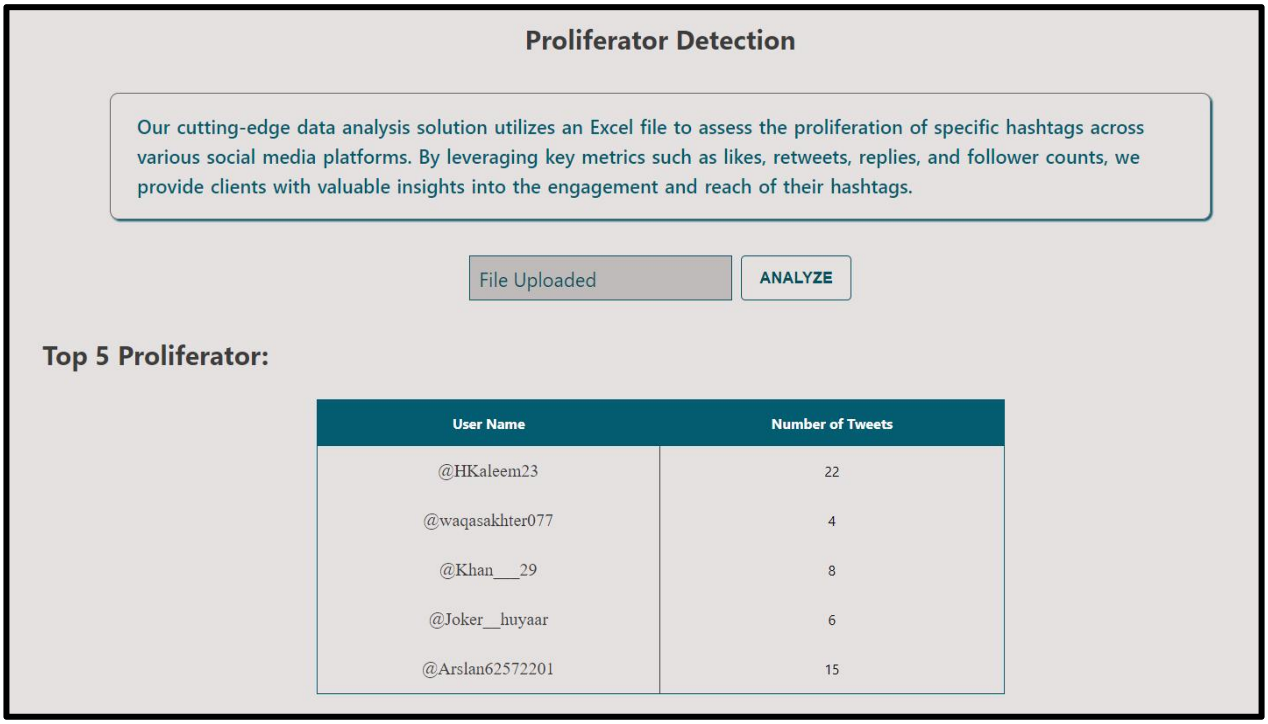


Fig 21: Displaying of Number of Tweets by User



User Sentimental Analysis

Fig 22: Displaying of Sentiment Analysis for Users



Proliferator Detection

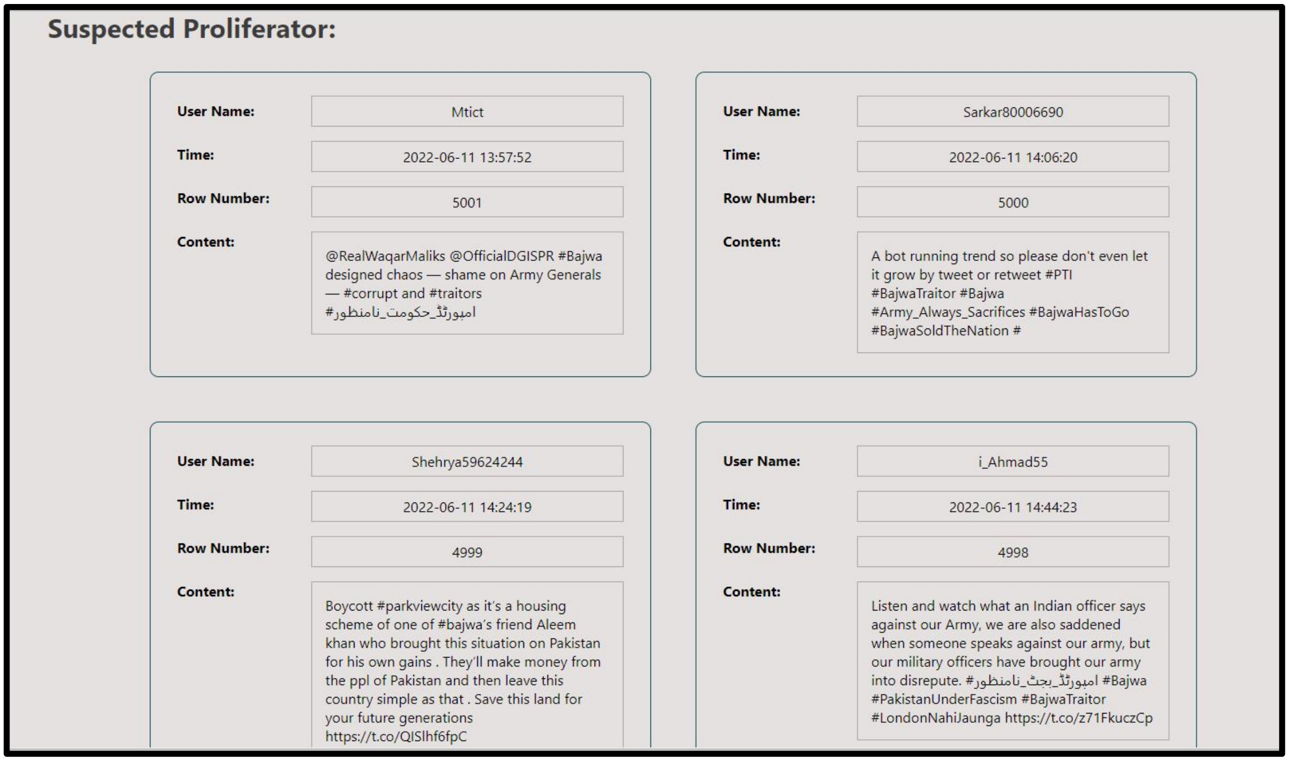
Our cutting-edge data analysis solution utilizes an Excel file to assess the proliferation of specific hashtags across various social media platforms. By leveraging key metrics such as likes, retweets, replies, and follower counts, we provide clients with valuable insights into the engagement and reach of their hashtags.

File Uploaded ANALYZE

Top 5 Proliferator:

User Name	Number of Tweets
@HKaleem23	22
@waqasakhter077	4
@Khan___29	8
@Joker__huyaar	6
@Arslan62572201	15

Fig 23: Proliferator Detection



Suspected Proliferator:

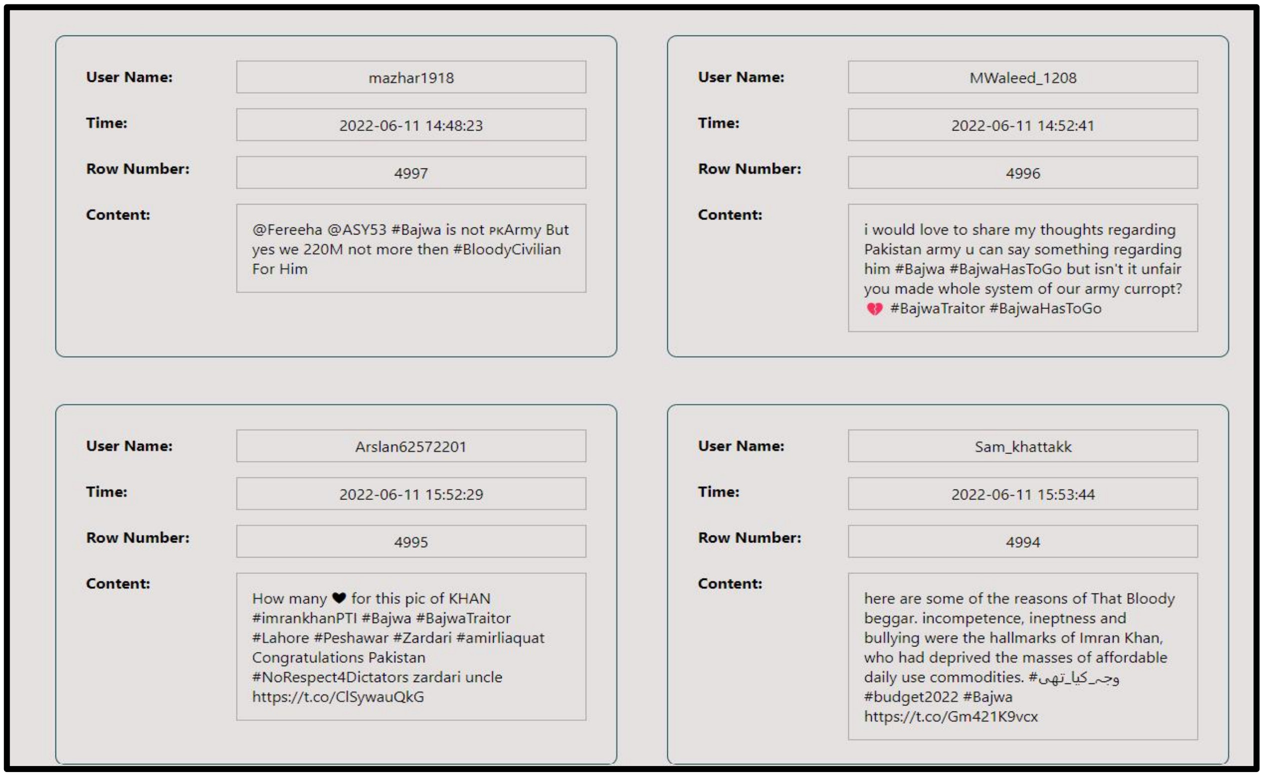
User Name: Mtict
Time: 2022-06-11 13:57:52
Row Number: 5001
Content: @RealWaqarMaliks @OfficialDGISPR #Bajwa designed chaos — shame on Army Generals — #corrupt and #traitors #امپورٹڈ_حکومت_نامنظور

User Name: Sarkar80006690
Time: 2022-06-11 14:06:20
Row Number: 5000
Content: A bot running trend so please don't even let it grow by tweet or retweet #PTI #BajwaTraitor #Bajwa #Army_Always_Sacrifices #BajwaHasToGo #BajwaSoldTheNation #

User Name: Shehrya59624244
Time: 2022-06-11 14:24:19
Row Number: 4999
Content: Boycott #parkviewcity as it's a housing scheme of one of #bajwa's friend Aleem khan who brought this situation on Pakistan for his own gains . They'll make money from the ppl of Pakistan and then leave this country simple as that . Save this land for your future generations https://t.co/QISlhf6fpC

User Name: i_Ahmad55
Time: 2022-06-11 14:44:23
Row Number: 4998
Content: Listen and watch what an Indian officer says against our Army, we are also saddened when someone speaks against our army, but our military officers have brought our army into disrepute. #امپورٹڈ_بجٹ_نامنظور #Bajwa #PakistanUnderFascism #BajwaTraitor #LondonNahiJaunga https://t.co/z71FkuczCp

Fig 24: Suspected Proliferator



User Name: mazhar1918
Time: 2022-06-11 14:48:23
Row Number: 4997
Content: @Fereeha @ASY53 #Bajwa is not ᴘᴋArmy But yes we 220M not more then #BloodyCivilian For Him

User Name: MWaleed_1208
Time: 2022-06-11 14:52:41
Row Number: 4996
Content: i would love to share my thoughts regarding Pakistan army u can say something regarding him #Bajwa #BajwaHasToGo but isn't it unfair you made whole system of our army curropt? 💔 #BajwaTraitor #BajwaHasToGo

User Name: Arslan62572201
Time: 2022-06-11 15:52:29
Row Number: 4995
Content: How many ❤ for this pic of KHAN #imrankhanPTI #Bajwa #BajwaTraitor #Lahore #Peshawar #Zardari #amirliaquat Congratulations Pakistan #NoRespect4Dictators zardari uncle https://t.co/ClSywauQkG

User Name: Sam_khattakk
Time: 2022-06-11 15:53:44
Row Number: 4994
Content: here are some of the reasons of That Bloody beggar. incompetence, ineptness and bullying were the hallmarks of Imran Khan, who had deprived the masses of affordable daily use commodities. #وجہ_کیا_تھی #budget2022 #Bajwa https://t.co/Gm421K9vcx

Chapter 5: Conclusion

We have built an Open-source intelligence (OSINT) tool (web application) to gather and analyze data from publicly available sources (Twitter) to meet specific intelligence requirements.

Our project offers an indigenous solution, which means it has been developed locally, in the same nation as the end customers, without the involvement of outside parties. The approach we're taking is unique in that it addresses all of your requirements on a single platform, including data collection, data analysis, integration of open-source (OSINT) tools, making decisions based on data analysis, hashtag analysis, profile analysis, and sentimental analysis.

The tool we've proposed offers more focused analysis, exact insights, and customization options than other tools of a similar kind, giving it a competitive advantage over them. Overall, it will be a financially wise choice.

Businesses and Law Enforcement Agencies (LEAs) can use AI-based social media harvesting as a valuable tool for mining insightful information from Twitter data. Businesses can use Twitter data to better understand customer behavior, market trends, and brand sentiment with the assistance of AI algorithms and data analysis tools, which can then be used to inform marketing and business initiatives.

LEAs can use Twitter data to track criminal behavior, keep tabs on public opinion, and spot possible risks to public safety. To spot patterns, trends, and potential dangers, LEAs can examine massive amounts of Twitter data with the aid of machine learning algorithms and natural language processing technologies.

Chapter 6: Future Work

Future milestones that need to be achieved to commercialize this project are the following.

- By keeping up with new trends, incorporating new data sources, improving data accuracy and relevance, enhancing user experience, offering more customization options, and ensuring compliance with data privacy laws, we can adapt our current solution to meet market demands and trends. The solution can maintain its competitiveness and accommodate the changing needs of enterprises and LEA's by giving priority to these areas.
- We want to be getting adopted by the LEA's. To do this, we must first learn about the demands set by LEA's and the difficulties they have in monitoring social media. This could require completing market research, speaking with LEAs directly, and staying current with new developments in the industry.
- Upgrading our project to include other social media platforms as well. This will enable us to better assist our clients and keep on top of new trends in social media analysis.
- We will be working on similar projects for other businesses by creating an approach that focuses on generating a solid reputation and forming connections with a varied variety of clients across all industries.
- We will need to create a deployment strategy that makes sure our system is dependable, scalable, and user-friendly before we launch our product to businesses and LEAs. By offering an efficient service that satisfies the needs of our customers, we will be able to reach our future milestones.

References and Work Cited

1. SOCIAL MEDIA HARVESTING by Man-pui Sally Chan, Alex Morales, Mohsen Farhadloo, Ryan Joseph Palmer, and Dolores Albarracin. [online] Available at: <https://www2.psych.ubc.ca/~schaller/528Readings/Chan2019.pdf>
2. Big Data, Collection of (Social Media, Harvesting) by Hai Liang and Jonathan J. H. Zhu, *The International Encyclopedia of Communication Research Methods.* [online] Available at:<https://www.researchgate.net/publication/320929016_Big_Data_Collection_of_Social_Media_Harvesting>
3. What is Open-Source Intelligence? by Ritu Gill. [online] Available at: <https://www.sans.org/blog/what-is-open-source-intelligence/>
4. Social Media Data Mining: What It Is, How It Works, and How to Use It by Aušrinė. [online] Available at: <https://whatagraph.com/blog/articles/social-media-data-mining>
5. How to scrape millions of tweets using Snscrape: A comprehensive guide to scraping tweets coherent with Twitter by Rashi Desai. [online] Available at: <https://medium.com/dataseries/how-to-scrape-millions-of-tweets-using-snscrape-195ee3594721>
6. Analyzing Twitter Data for Deeper Insights. [online] Available at:< https://www.nabler.com/digital-analytics/articles/analyzing-twitter-data-for-deeper-insights/>
7. How Twitter takes business intelligence and market research to new heights by Barbara Kremers. [online] Available at:<https://www.buzztalkmonitor.com/blog/how-twitter-takes-business-intelligence-market-research-to-new-heights/>

8. How to conduct competitor analysis benchmarking on Twitter by Sirarpi Sahakyan. [online] Available at:<https://sociality.io/blog/competitor-analysis-benchmarking-on-twitter/>
9. Social media and its role for LEAs by Petra Bayerl and Babak Akhgar, *Cyber Crime and Cyber Terrorism Investigator's Handbook (pp.197-220).* [online] Available at:<https://www.researchgate.net/publication/288205288_Social_media_and_its_role_for_LEAs
10. How to Track and Analyze Your Twitter Hashtags by Sam Lauron. [online] Available at: < https://www.rivaliq.com/blog/twitter-hashtag-analytics/>
12. Scraping for big data without breaking the bank: Using Twitter API for academic research by Hannah L. Ford, *Western Region American Association for Agricultural Education Conference*
13. Using Osint to Gather Information About a User from Multiple Social Networks by Nilesh Sambhe, Piyush Varma, Arpan Adlakhiya*, Aditya Mahakalkar, Nihal Nakade and Renuka Lakhe, Vol. 13 No. 2 (2020)
14. Web Scraping versus Twitter API: A Comparison for a Credibility Analysis by Irvin Dongo, Yudith Cadinale, Ana Aguilera, Fabiola Martinez, Yuni Quintero, Sergio Barrios, *The 22nd International Conference on Information Integration and Web-based Applications & Services*
15. Detecting Cyber Security Related Twitter Accounts and Different Sub-Groups: A Multi-Classifier Approach by Mohamad Imad Mahaini∗ and Shujun Li†, *International Conference on Advances in Social Networks Analysis and Mining*

Thesis

ORIGINALITY REPORT

10%	3%	1%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Franklin University Student Paper	3%
2	Submitted to Columbia College of Missouri Student Paper	2%
3	Submitted to Swinburne University of Technology Student Paper	2%
4	Submitted to Riverside College Halton Student Paper	1%
5	content.yudu.com Internet Source	1%
6	www.builtinnyc.com Internet Source	1%
7	ir.jkuat.ac.ke Internet Source	1%

Exclude quotes	Off	Exclude matches	O
Exclude bibliography	On		