Journal of Policy Research, 8(4), 292-309.
https://jprpk.com
https://doi.org/10.5281/zenodo.7635142

Real-Time Extraction and Annotation of Social Media Contents for Predicting National Consumer Confidence Index

Muhammad Ashraf[1], Arslan Ali Raza[2], Muhammad Ishaq[3], Wareesa Sharif[4], Asad Abbas[5], Salman Irshad[6]

Abstract
The advent of web enabled technologies has given birth to new communication platforms such as Facebook, Twitter, YouTube, and blogsites. Online users belonging from variant geographical backgrounds share their opinion, sentiments, and appraisals about number of real-world entities on the social media platforms. These opinion bearing contents have great importance to observer and analysts. These opinionative contents can benefit in the prediction of consumer confidence index (CCI) which is referenced by businesses, governments, and other institutions when they make strategic decision. Social media channels and microblogging sites can have a high volume of data on consumer confidence, analyzing such contents can significantly improve the impact and accuracy of CCI but unavailability of consolidated application for the extraction of user generated content is restricting the further process. However, this study aims to conduct an implementation based comparative literature review to unfold the most valuable mechanisms of text extraction, normalization, and annotation of social media contents from Facebook, Twitter, Youtube, and blogging sites for effective prediction of the CCI. A case study of NACOP (Pakistani National Consumer Confidence Predictor) proposed by Ashraf et al. (2022) about the data of purchasing behavior, consumer price, Job/Employment and personal finance is presented to explore the data extraction tools for the Facebook, Twitter, Youtube, and blogging sites. The experimental evaluation revealed that Facepager, Tagv6, Netvizz, and web scrapper are the optimum extraction APIs for Facebook, Twitter, YouTube, and Blogsites respectively. The study has significant implications to theory and practice.
***Keywords*:** Consumer Confidence Index, Data Extraction, Data Normalization, NACOP, Preprocessing, Sentiment analysis, Social Media Analytics, Facepager, Tagv6, Netvizz, and Web Scrapper

1. Introduction

Social media has now become the key platform of sharing opinions, sentiments, and suggestions. Online users publish their views and suggestions about number of real-world entities. These opinions and sentiment have great importance to analysts and observers. Machine learning approaches and mechanisms are used in digging out the insights of these opinions. Extraction, annotation, and classification of public opinions shared in the form of text can be referred as opinion mining or sentiment analysis (Habib, 2022). Opinion mining can significantly capture consumers' views, opinions, and sentiment for predicting Consumer Confidence Index [hereafter it is called CCI]. CCI is used to compute the optimistic as well as pessimistic approach of consumer regarding economic situation. An optimistic customer may spend more money whereas a pessimistic may spend less money on buying a product due to lack of confidence (Ashraf, Raza, & Ishaq 2022). The CCI information has been used to influence business strategy decisions made by manufacturers, retailers, banks, and governments, relative to other key economic indicators (Investopedia, 2014). A consistently declining trend in CCI would indicate that consumers have a negative economic outlook which implies further decline in economic activity. While a positive CCI trend would be taken as antecedent to growth in economic activity. By analyzing the correlation between different economic indicators, businesses could base their decision of either investing in production capacity or cut labor costs, based on the anticipated economic activity inferred from the level of consumer confidence (Igboayaka, 2015). Similarly, banks can also take measures to curb the anticipated increase in borrowing by consumers and the use of credit cards. The government could come up with measures such as reduction of taxes to encourage spending (Ashraf *et al.*, 2022).

The survey-based method of data collection has been extensively used for computing the CCI, it poses several challenges to industry with regards to measuring the CCI (Ashraf *et al*., 2022; Shayaa *et al.*, 2018). First, *Data Frequency and Timeliness;* the official CCI figures are published quarterly, which does not allow for end users/industry players to make accurate prediction and planning. There is a need to publish weekly or monthly CCI indicator for industry players and end-users to facilitate their strategic business activities (e.g., marketing plans, financial forecasting, investments plans, etc.). Second, *Quality of Respondents;* the most questionnaire surveys are conducted via phone call or via face-to-face interviews by a third party or companies and in most cases, the phone call lands to the similar demographics (i.e., housewives, civil service, & students). Results might be highly biased and may not represent the collective view of the consumer confidence at the national level (Ashraf *et al*., 2022). Third, *Limited Sample Size;* due to the laborious and time-consuming approach via phone calls and face-to-face interviews, the sample size can only be limited to 1,000 – 2,000 people, or households. To extrapolate

[1]COMSATS University Islambad Vehari Campus, Pakistan
[2]COMSATS University Islambad Vehari Campus, Pakistan
[3] COMSATS University Islambad Vehari Campus, Pakistan
[4] The Islamia University of Bahawalpur, Pakistan
[5] COMSATS University Islambad Vehari Campus, Pakistan
[6] COMSATS University Islambad Vehari Campus, Pakistan

CCI values to a larger and diverse population in Pakistan, the result may not be representative or accurate (Ashraf *et al*., 2022). Fourth, *Lack of Alternative Data Source;* there is no alternative consumer confidence metrics that truly reflect the condition of the economy (Ashraf *et al*., 2022). The limitations of consumer confidence survey continue to spur research efforts to clarify and develop evidence on associations that form the core of the economic indicator of consumer confidence (Odendaal, Reid, & Kirsten, 2020). Yet in another regard, the CCI is subjected to rigorous analysis, especially when associations with economic activity are being discussed. One aspect of such dynamic has to do with the emergence of new sources of data such as social media that could be used to measure consumer confidence.

Social media is a potential source of information for measuring the human perceptions that reflect their confidence in the economy (Ashraf *et al*., 2022; Chung, Shin, & Park, 2022). Consumer sentiment data on social media offer greater marginal significance for predicting consumer confidence (Odendaal, Reid, & Kirsten, 2020; Shayaa *et al.*, 2018). Particularly, consumers' activities in Pakistan provide exciting opportunities to leverage on publicly available data on social media sites (Audi *et al.,* 2021; Ashraf *et al*., 2022; Audi *et al.,* 2022). Pakistan has a population of over 212.7 million, 65.5 million labor workforce, $256.66 billion consumer spending, 163 million mobile penetrations, 65.13 million Internet users and 37 million social media users (87.68% Facebook, 4.63% Twitter, 2.82% Pinterest, 2.19% Youtube & 1.96% Instagram). Additionally, the data flow in digital era is growing faster than trader and finance (Hirt and Willmott, 2014). Consequently, social media is a relevant source and more effective to provide a huge volume of sentiment data on consumer confidence compared to conventional survey questionnaires. Further, it is also noticed that consumer sentiment data on social media are not in desirable format, and one need to collect, consolidate, normalize, and preprocess these opinionative contents to compute the effective value of the CCI.

However, this study aims to conduct an implementation-based comparative literature review to disclose the most valuable mechanisms of data extraction, normalization, and annotation of Pakistani consumers' opinionative contents on Facebook, Twitter, Youtube, and Blogging sites for measuring the CCI. A case study of National consumer confidence predictor (NACOP) proposed by Ashraf *et al.* (2022) about the sentiment data of consumers' purchasing behaviour, personal finance, Job/Employment, and consumer price is presented to explore the most suitable mechanisms of data extraction with respect to social media and blogging sites. Consequently, this study is motivated to address the following research questions:

RQ1. How social media channels and blogging sites can be used as source platforms for the prediction of consumer confidence index?

Motivation of RQ1: A brief description of social media networks and blogging sites along with their applications w.r.t the CCI is explored and stated.

RQ2. What are the most suitable mechanisms and tools used in the extraction of user generated contents on Facebook, Twitter, Youtube and Blogsites?

Motivation of RQ2: Experimental evaluation and comparative analysis of existing APIs, tools and mechanisms for Facebook, Twitter, Youtube and blogsites is presented to unfold the best possible mechanisms of sentiment data extractions of consumers' purchasing behaviour, personal finance, Job/Employment, and consumer price as per the defined criteria for measuring the CCI.

2. Social Media Networks as a Source of Data for Consumer Confidence Measurement

RQ1. *How social media channels can be used as source platforms for the prediction of consumer confidence index?*

Social Media is composed of online technology tools which enable communication between different people from around the world via the Internet. According to Margaret (2011), "*Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration*". Social media is a large and well known for its social network and content sharing abilities – text, videos, and pictures (Igboayaka, 2015). It can also be seen as a setting for virtual discourse where people create content, share it, bookmark it and network at an extraordinary degree and speed. It is also used as a real-time snapshot of updates of interest, location, instant memories and so on. The use of social media networks has brought about one of the biggest cultural shifts since the industrial revolution. In 2020, over 3.6 billion people were using social media worldwide, a number projected to increase to almost 4.41 billion in 2025 (Statista, 2021b).

Social media networks are a versatile source of information which could be harnessed for assessing the human perceptions that reflect their confidence in the economy. For example, Facebook and Twitter are popular networks used to express opinion, a state of mind or a perspective which could be mined for the sentiment of users relative to their economic situation (Chung, Shin, & Park, 2022). Youtube is not only providing video contents but also allow users to comment on the video, generating huge sentimental data about the video. Also, blogging sites including News sites, they generate a rich source of opinion towards a subject of discussion, e.g., consumer price, personal finance, unemployment, purchasing behavior, a new government policy, stock market fluctuations, a commercial investment deal and much more (Ashraf *et al*., 2022; Li, Xu, Zeng, Tse, & Chan, 2023). Aggregation of such information can be used to 'know' the economic state of consumers. Furthermore, 60% of the world's population has access to the Internet and now interacts using social media networks (Statista, 2021b). Since its

inception in 1996, social media has managed to infiltrate half of the 7.7 billion people in the world (Brian Dean, 2021). Social media platforms almost tripled their total user base in the last decade, from 970 million in 2010 to the number passing 4.48 billion users in July 2021 (Brian Dean, 2021). Therefore, tapping into this 'new' data source of social media could give a supplementary boost to the conventional methods of data gathered to measure the level of consumer confidence in the economy.

For use in the economic and business sector, social media involves the gathering of comments made in online forums, created with the sole purpose of educating people about products, brands, services and general issues in the economy (Blackshaw & Nazzaro, 2004). Mangold and Faulds (2009) defined social media as a wide range of word-of-mouth forums which allows users to be able to express their feeling about a topic or even create a topic for discussion. The communicating parties can connect and hold conversations using formats such as text, audio, video, images, and other multimedia. Social media has become one of the strongest and most powerful tools used online since it allows thousands of users to be reached within the shortest minimum time possible (Odendaal, Reid, & Kirsten, 2020). According to statistics provided by socialnomics.net, on the speed of development of the social media, it took the radio 38 years to reach 50 million users, TV 13 years to reach 50 million users, but it took Facebook just 9 months to reach 100 million users (Igboayaka, 2015).

Using social media has become one of the new methods used by The Economist to gather data on the level of consumer confidence in the economy, the methods including: counting the number of likes and dislikes of a topic, how many comments have been given by consumers on relating topics and gathering how many re-tweets were done by consumers (Igboayaka, 2015). This method of gathering data shows the real consumer sentiments towards a brand or the economy based on the comment provided by the consumer. A study on social media effectiveness showed that 53% of the authenticated Twitter users recommended a company and its products by tweeting about it and 48% of the people actually delivered on their intentions to buy a particular product (Performics, 2010).

Social media networks are rich with expressions of sentiment which is important to the measurement of consumer confidence index. *Sentiment* is known as the feeling or reason for an expression made behind a comment and reference about a particular topic, news, or product (Igboayaka, 2015). Various tools have been made which enable measuring the consumers' sentiment. From a business perspective, the ability of a business owner or a brand to be able to measure or understand the consumer's sentiment behind a comment is very beneficial to the growth of the organization. This helps to determine the state of mind a consumer was in as at the time he/she made the comment (Li *et al.* 2023).

Asur and Huberman (2010) conducted a study on how social media can help predict the future outcome of a new movie before it is released. In this research, Twitter was used to forecast box-office revenues for movies. The result from the research was calculated by analyzing the number of tweets that are currently on the network based on a particular movie. These tweets were divided into positive and negative tweets, Positive tweets are interpreted as representative of customers looking forward to watching the movie based on the preview, while the negative tweets mean that the customers are not interested in watching the movie. A linear regression model for predicting box-office revenues of movies in advance of their release was constructed. The results showed that there is a strong correlation between the amount of attention or discussions consumers have about a movie to be released and its subsequent ranking when the movie is eventually released (Asur & Huberman, 2010). Similarly, research conducted by Chris Barry, Rob Markey, Eric Almquist and Chris Brahm (2011) to examine how the social media has affected the consumers' confidence on a product or the organization. The questions asked on the social media about the product included "*How likely would you recommend [this company or product] to a friend or colleague in social media?*" After the analysis, they stated that: "*Customers who engage with companies over social media are more loyal and they spend up to 40 percent more with those companies than other customers*".

In 2012, Schweidel, Moe, and Boudreaux carried out an analysis of the potential to get brand sentiment from social media conversations. They used data which was collected from different social media domains such as Facebook and Twitter. They proposed the use of a hierarchical Bayesian regression model which was used to measure these sentiments effectively. Based on the research carried out, it has been discovered that social media is one of the fastest and easiest mediums that can be used to get large amounts of consumers reviews which can be used eventually to calculate the overall consumer confidence on a brand or the economy as a whole. Different approaches on what affects consumer confidence were also researched. Among them are unemployment, rising inflation and stock price, but the greatest question remains "*What actually causes a rise and fall in consumer confidence?*" In conclusion, the social media itself can cause a rise and fall in consumer confidence in the world today as information about the economy can easily be known by consumers as quickly as it happens (using blogs, Facebook, Twitter). This affects how consumers feel about the economy unlike the days when such platform was not easily accessible.

Using social media networks as a platform for social intelligence gathering is catching on as a data source in the market intelligence research industry. Companies such as Nielsen, Netbase, Frost & Sullivan, and Mintel enlist their use of content from social media networks, for the purpose of intelligence gathering and marketing (Netbase, 2010; Nielsen, 2013). It is no surprise that a lot of research continues to concentrate on the improvement of analytical methods and frameworks to translate such content into useful information. It is this translation that is pivotal to the usefulness and correctness of whatever information is synthesized from social media networks. For

example, an area of study like sentiment analysis frameworks have originated concepts such as Natural Language Processing (NLP) that can be used as a tool to translate data from social media networks into meaningful insight. However, consumer confidence index and how it is measured have expanded their scope beyond the use of traditional survey-based questionnaires alone and new data sources such as sentiment data on social media present opportunities to gain another perspective on consumer behavior in terms of economic habits (past), situation (present) or intentions (future). At this time, social media networks are relevant sources and more effective compared to cross-sectional nature of survey questionnaires.

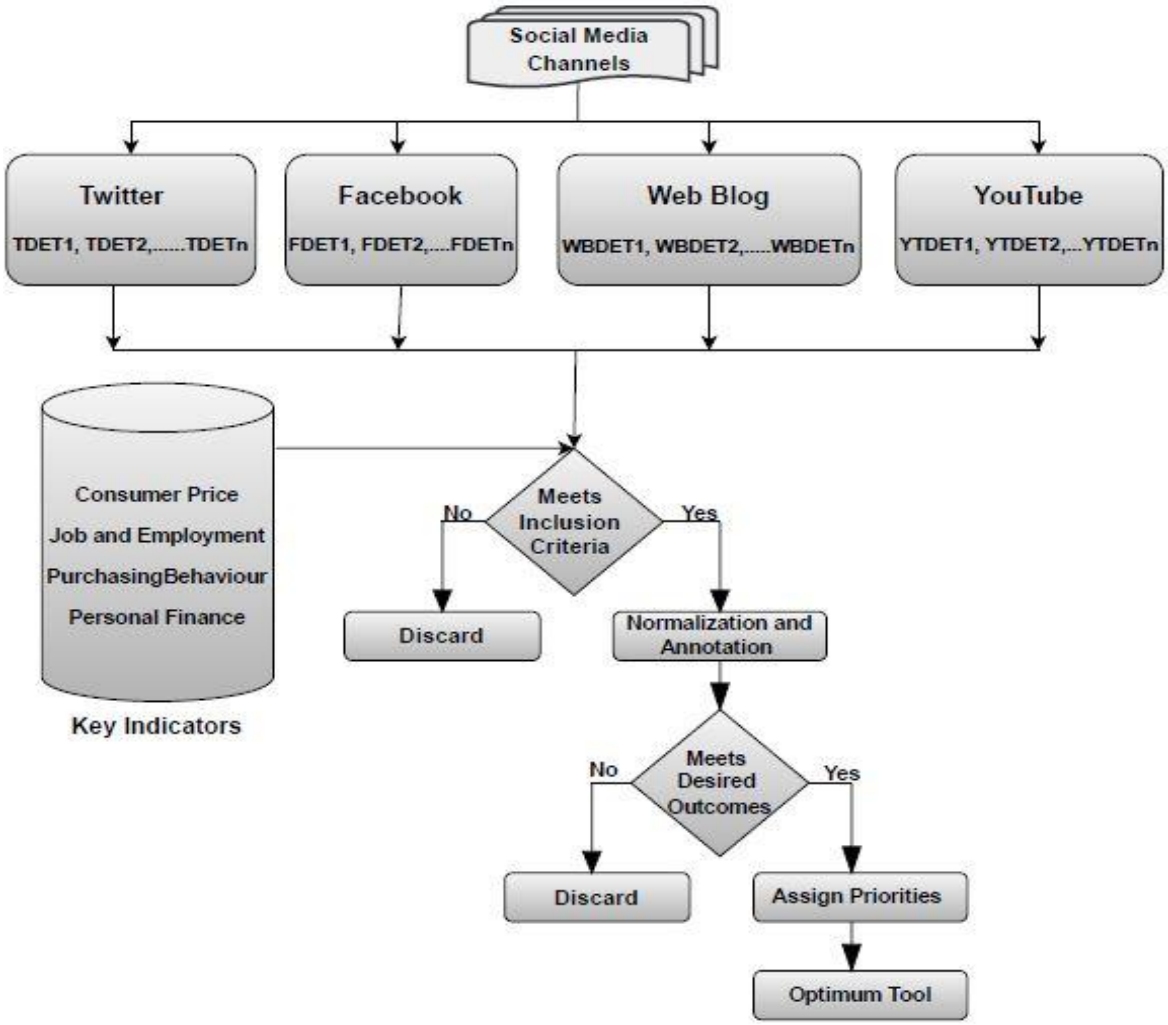

Figure 1: Implementation-based Comparison of Text Extraction Mechanims

3. Methodology

As mentioned earlier this study is motivated to address key research questions of "*What are the most suitable mechanisms and tools used in the extraction of user generated contents on Facebook, Twitter, Youtube and Blogsites?*" However, we aim to conduct an implementation-based comparative literature review to disclose the most valuable mechanism of text extraction, normalization, and annotation of social media contents for effective prediction of the CCI. The study methodology is presented in Figure 1 that explains how implementation-based comparison review is conducted to explore the optimum mechanism of data extraction techniques for Facebook, Twitter, Youtube, and Blog sites based on inclusion criteria for extracting the data of four indicators; purchasing behaviour, job/employment, consumer price, and personal finance, adopted from Ashraf *et al.* (2022) A detailed systematic literature review is first conducted to collect a list of tools and APIs used in the extraction of social media and blogging sites users' opinionative contents for all four channels. Twitter data extraction tools (TDET) are represented as $TDET_1$, $TDET_2$, $TDET_3$ and $TDET_n$. Similarly, $FDET_1$ denotes the Facebook data extraction tool, whereas WPDET presents web blog data extraction tool and at the end YTDET shows YouTube data extraction tools. A list of possible APIs and tools is first checked for inclusion criteria. If a tool satisfies the inclusion criteria it is included for further assessment otherwise it is discarded as irrelevant. The inclusion criteria have three key considerations; first, data must be features/keywords/domain Specific, second, data must be time specific (i.e., when the tweet is created/published), and third, data must be country specific (i.e., Tweet or comment published by Pakistani account/channel). After careful assessment of inclusion criteria, these extraction APIs are

then passed to next phase of text normalization and annotation. If a tool produced quality contents according to desired annotation, it is passed further for priority assignment otherwise it is excluded as irrelevant. A priority list is generated as last step of extraction. Priority list follows the inclusion criteria along with size and multidimensionality of contents generated by each target tool and API.

To achieve the target objectives, numerous mechanisms have been employed to unfold the insights of existing extraction APIs and Tools. The scope and constraints for target datasets are defined at compile time. Tweets having formal and informal English and Roman Urdu Opinionative terms are in the scope of target datasets of consumer price, job/employment, personal finance, and purchasing behavior, whereas other contents are excluded. The key assumptions considered in the extraction of target data are elaborated in Figures 2.

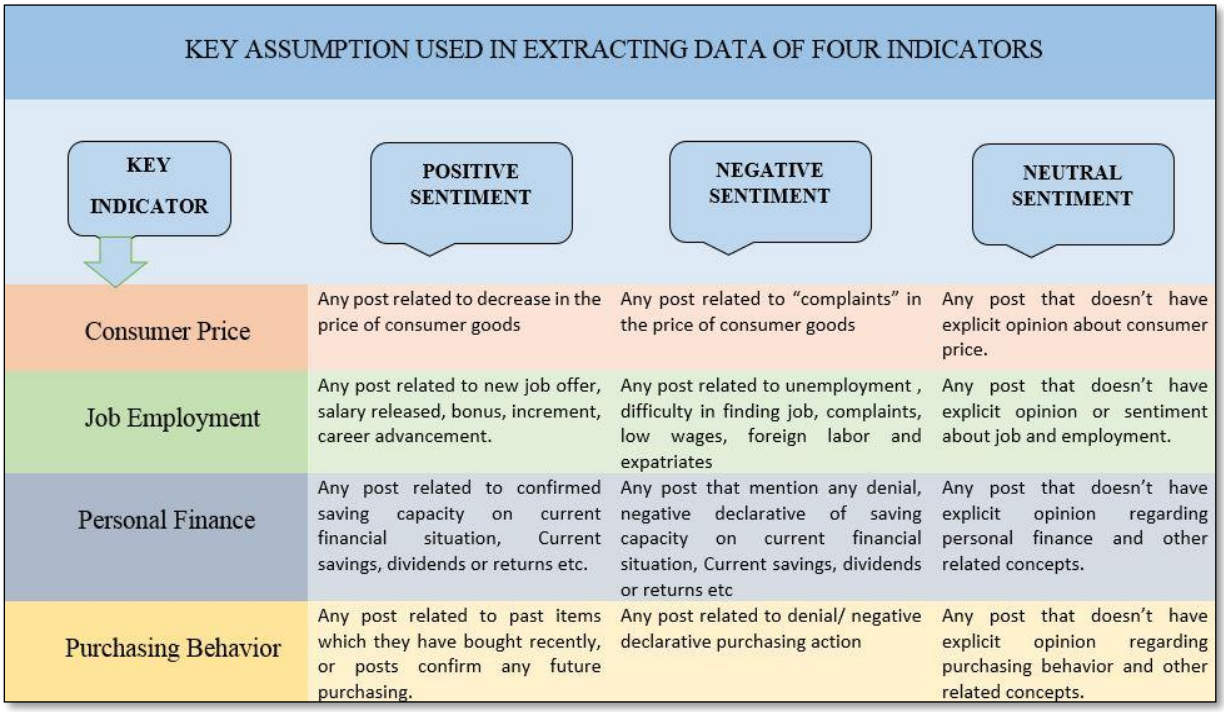

Figure 2: Key Assumptions used in the Extraction of Dataset

A list of keywords is utilized in the extraction of user generated contents. Table 1 presents a subset of keywords used in the orientation of public tweets and comments.

Table 1: A Partial List of Keywords Utilized in the Extraction of User Generated Contents

S#	Domains	Keywords
1	Purchasing Behavior	" Want to pay, purchased, bought; not buy; will pay, will purchase, will buy, I bought, I want to buy, I want to pay, I don't want to pay, I want to buy, I will buy, میں نے خریدا ، میں خریدنا چاہتا ہوں ، میں ادا کرنا چاہتا ہوں ، میں ادا نہیں کرنا چاہتا ، میں خریدنا چاہتا ہوں ، میں خریدوں گا ، will booking, will pay, will purchase, will buy, not buy, want to pay, purchased, bought, want to rent, already bought, want to buy, want to pay, want to rent, booking, don't want to pay, buy, not buy, will buy, بکنگ کریں گے ، ادا کریں گے ، خریدیں گے ، خریدیں گے ، نہیں خریدیں گے ، ادا کرنا چاہتے ہیں ، خریدا ، خریدا ، کرایہ پر لینا چاہتے ہیں ، پہلے ہی خریدا ہے ، خریدنا چاہتے ہیں ، ادا کرنا چاہتے ہیں ، کرایہ چاہتے ہیں ، بکنگ ، ادا نہیں کرنا چاہتے ، خریدیں ، خریدیں ، خریدیں گے ، to take, to buy, purchased, bought, have bought; not buy; will buy, will purchase, will buy, خریدنا ، خریدنا ، خریدنا ، خریدنا ، خریدنا؛ نہ خریدیں خریدیں گے ، خریدیں گے ، خریدیں گے ، want to go, looking for, went to; not going to; will go, maybe go, شاید جانا چاہتے ہیں ، تلاش کر رہے ہیں ، گئے؛ نہیں جا رہا؛ جائیں گے ، جائیں گے ، جائیں گے ، kharidna chahta hon, kharid lia hai, kharidon ga, khrida hai, khridna hai, khridon ga, food, drinks, household, cometics, healthcare, fashion, stationery, stationeries, kharido, roti, kharidh, going for movies, dinner, entertainment, buying groceries, grocery, holiday trip, buy vehicle, car, motorcycle, phone, buy house, home, price, buy, purchase, resturant, restaurant, affordable, ghee, ghi tail ki price, order kia hai, order, order kron ga, order kia tha, buy electronics, furniture, accessories, mobile phone, mobile, smartphone, laptop, computer, main adda karna chahta hon, main pay karta hon, pay kron ga, mobile kharida hai, Let's wait for

		samsung, buy phone, bech k, installment, plot ki qeemat, plot ki price, rent pay laina hai, kraya py laina hai, rent, seat booking, bus ticket, train ticket, airplan ticket, jahaz ki ticket, kharidain gay, kharidain lain, kharidain, ticket, cinema ticket, movie ticket, Accommodation, transportation, Car, motorcycle, House property, real estate karaya, kirayadar, malikmakan, rent pay ghar chahy, rent py ghar, kharid chuka hon, plot for sale, buy plot, trafic challan, traffic challan, motor bike, travel, restaurant chalain, movies dekh chuky, movies dekhne chalain, want to watch movies, want to travel, plot khridna hai, plot kharidain gay, plot chahy, plot khrid lya hai, juice peena hai, movies dekty hain, movies nae dekty, avoid watching movies, stop eating in restaurant, saw movie, ghomane jain gay, doctor se checkup karwana hai, gari wash karwani hai, car wash, ghari wash, gari dulwani, carwash, want to play, want to see, want to learn, want to take away, want to have"
2	Job Employment	"Get a job, new job, hired, hiring, job absorption, promotion, salary increase, big salary, get salary, job application, accepted job, job vacancy, job offer, apply job, jobs offer, permanent jobs, new jobs, نوکری حاصل کریں ، نئی نوکری ، نوکری ، بھرتی ، نوکری جذب ، پروموشن ، تنخواہ میں اضافہ ، بڑی تنخواہ ، تنخواہ ، نوکری کی درخواست ، قبول شدہ نوکری ، نوکری کی جگہ ، نوکری کی پیشکش ، نوکری کی درخواست ، نوکریوں کی پیشکش ، مستقل ملازمتیں ، نئی ملازمتیں, Unemployment, no job, layoff, hard work, opportunities, hard work, laid off, quit, unemployed, jobless, no job, low wages, low salary… بے روزگاری ، کوئی نوکری نہیں ، ملازمت سے برطرفی ، محنت ، مواقع ، محنت ، چھٹی ، بے روزگار ، بے روزگار ، نوکری نہیں ، کم اجرت ، کم تنخواہJob vacancies, Job vacancies, job offers, job search, job search, Jobs vacancy, now hiring, ملازمت کی آسامیاں ، نوکری کی آسامیاں ، نوکریوں کی پیشکشیں ، نوکری کی تلاش ، نوکری کی تلاش ، نوکریوں کی خالی جگہیں ، اب نوکریوں پر۔ "New Jobs offers, nae job, nokri mil gae hai, jobs offer received, job mil gae, tankhawa, job advertisment, permosion, nokri ki application, nokri ki darkhashat, nokri ki paishkash, molazamat, mulazamat, molazmat, mulazmat, molazmin, Salary increment, salary increase, New Positions, Career advancement, career, Jobs satisfaction, nokri chahy, salary revision, bonus, salary increment, career advancement, oppertunity, opertunity, moqa, employee salary, paid employee, labor, labour, workplace, rishwat khor, nokriyan, nokri chahye, sarkari nokri, nokri chohey, nokri pay rakh lain, retrenchment, Low jobs opportunity, Low wages, Jobs dissatisfaction, berozgari, rozgar, working, ujrat, nokri se chotti, low ujrat, kam tankhawa, nokri nae mil rahi hai, unemployment, difficulty in finding jobs, labour, retirement, retired employee, retired officer, government employee, private employee, govt employee, satified employee, motmain molazim, raises in salary, salary raises, raise in salaries, increase in salary, persioners, jobless, unemployed, termination, job termination, berozgar, job advertisment, job dond, hiring, employee, employment, nokri ki talash, nokri khali hai, jobs available, nokri available, job chahy, unemployment bharti jayegi"
3	Consumer Price	"Expensive, cheap, complaint, price up, price down, price has gone up, price has gone down, price dh up, price dh down, complaint, price increase, price decrease, price going up, price going down, price goes up, price goes down, price went up, price went down, price up, price down, prices do not change, prices are the same, prices are stagnant, price x changes, مہنگا ، سستا ، شکایت ، قیمت اوپر ، قیمت نیچے ، قیمت بڑھ گئی ہے ، نیچے ، شکایت ، قیمت میں اضافہ ، قیمت میں کمی ، dh اوپر ، قیمت dh قیمت نیچے گئی ہے ، قیمت بڑھ رہی ہے ، قیمت نیچے جا رہی ہے ، قیمت بڑھ رہی ہے ، قیمت نیچے جاتی ہے ، قیمت بڑھ جاتی ہے ، قیمت نیچے جاتی ہے ، قیمت بڑھ جاتی ہے ، قیمت کم ہوتی ہے ، قیمتیں تبدیل نہیں ہوتی ، قیمتیں ایک جیسی ہوتی ہیں ، قیمتیں مستحکم ہوتی ہیں ، قیمت ایکس تبدیل ہوتی ہے, mahnga hai, manga hai, sasta, shakyat, qimat barh rahi hai, qeemat barh rahi hai, qeemat, qimat barh gae, qimatt, mahngae, mahngai ho gae hai, mehngae, mehngai, mehngaiee, mahngaiee, mehangae, mehangai, mahangai, mahangae, cheap price, rupaya, rupai, mehngi tareen, petrol, desiel, rupy, rupees, rupee, pay kro, pay karna, diesel, subsidies, sasti, sastai, rupe, bijli mahngi, bijli, mahngai mukao march, mahngai bachao, paise, sasta, mahnga, mehnga, paisay, paisy, low price, high price, mehngai march, mehngi, herchez, cheese, cheez, utility store, qeemtain, awam rul gae, khurdonosh, sasti, bijli k bil, bijli k bilon, salander, slander, LPG, FPA, qemat, car fuel, high price, overpriced, saving, oil price, cheap plot, plot cheap, petroleum price, pertrolleum product, perices of petroleum, rising prices, increasing prices, decreasing prices, raise prices, increase price, mehngi kardi, mehngi krdi, prices barha di, increase the prices, increased the prices, price hike, prices record hike, prices increase, ruppe increase, rupee increase, rupee decrease, rupe, rising inflation, increasing inflation, inflation, stock price"

4	Personal Finance	"Can dividend, investment return, profit return, receive return, return on capital, save in, invest in, dividend in, good returns, savings, have savings, invest, invested, got profit, get profit, investment, keep money, savings, receive dividends, income, ریٹرن ، ریٹرن وصول کریں ، کیپیٹل پر ریٹرن کریں انویسٹ کریں ، ڈیویڈنڈ کریں ، اچھا ریٹرن ، بچت کریں ، بچت کریں ، انویسٹ کریں ، انویسٹ کریں ، منافع حاصل کریں ، منافع حاصل کریں ، انویسٹمنٹ کریں ، پیسہ رکھیں ، بچت کریں ، ڈیویڈنڈ حاصل کریں ، آمدنی, profit return, munafa, monafa, paisa bank main rakhain, bank main jama karwado, bank deposit, bank saving, insurance, bond, prize bond, profit and loss account, profit, loss, cash prize, share, company share, bachat, mahana bachat, amdani, amdni, bank loan, bank profit, mutual fund, stock, pasa, easypasa, qamai, saving, pension, pension nai mil rahi, pension mil gae hai, invest krain, bachat, monthy bachat, monthly saving, dividend, saving account, stocks, share, bonds, Blacklisted, get blacklisted, can't make a loan, no saving, don't save money, don't invest, no savings, not investing, didn't invest, bankrupt, bankrupt, bankrupt, x have savings, xsave money, x labur, x invest, x save money, don't want to invest, didn't save, no money to invest, بلیک لسٹ ، بلیک لسٹ ہو جاؤ ، قرض نہیں بنا سکتے ، بچت نہیں ، پیسہ نہیں بچا ، سرمایہ کاری نہ کرو ، بچت نہ کرو ، سرمایہ کاری نہ کرو ، دیوالیہ ، دیوالیہ پیسہ بچائیں ، سرمایہ کاری نہیں کرنا چاہتے ، xسرمایہ کاری ، xپیسہ ، x save، ، ایکس بچت ہے ، بچت نہیں ، سرمایہ کاری کے لیے پیسے نہیں۔, blacklist, blacklist ho gaya, bankrpt, paisa jamma nai kia, sarmayakari, srmyakari, loan, qarza, will pay back loan, qarza daina hai, qarza wapis kardon ga, financial situation, debt, qarza daina hai, returns kam ho gae hain, want to invest, wnt to invest, want to invst, want to save money, want to save money, not want to invest, interested to invest, not interested to invest, consider to invest, not want to buy stocks, want to buy stocks, want to buy stocks, intersted to buy stocks, want to buy shares, keep money, save money, to invest, to buy share, to buy stocks, open an account, سرمایہ کاری کرنا چاہتے ہیں ، سرمایہ کاری کرنا چاہتے ہیں ، سرمایہ کاری کرنا چاہتے ہیں ، پیسہ بچانا چاہتے ہیں ، پیسہ بچانا چاہتے ہیں ، سرمایہ کاری نہیں کرنا چاہتے ، سرمایہ کاری کرنے میں دلچسپی نہیں ، سرمایہ کاری کرنے میں دلچسپی نہیں ، سرمایہ کاری کرنے پر غور کریں ، اسٹاک خریدنا نہیں چاہتے ، اسٹاک خریدنا چاہتے ہیں ، اسٹاک خریدنا چاہتے ہیں ، اسٹاک خریدنے کے لیے انٹرسٹڈ ، شیئرز خریدنا چاہتے ہیں ، پیسہ رکھیں ، پیسہ بچائیں ، سرمایہ کاری کریں ، شیئر خریدیں ، اسٹاک خریدیں ، اکاؤنٹ کھولیں, bank main jama karwana chahta hon, pensioners, pensioner, I want to invest"

Several scrapper and extraction APIs are available for the extraction of user generated contents. It is not possible to apply a unified tool to extract data from all four channels due to varying nature and contents of the channels. A huge volume of experiments employed to identify the most suitable extraction APIs and tools as per nature of channels and scope of desired data. Following sections present channel wise comparative analysis of text extraction APIs and tools.

3.1. Facebook Data Collection Tools

Facebook has become world's largest social network platform, with more than 2.74 billion monthly active users worldwide which has been increasing 12% every year (Tonak, 2020). Facebook considerably tensed the access to its Application Programming Interface (API). The API served for the last 10 years as a core tool by which researchers gather behavioral and digital trace data from Facebook (Abdulla et al., 2018; Mancosu & Vegetti, 2020). On Facebook, independent users and third parties can easily get information related to user profiles, comments, and the reactions of publicly available posts by using the API. The API denoted the best approach by using which Facebook authorized third parties to acquire data from Facebook (Mancosu & Vegetti, 2020).

There are several Facebook data extraction tools are available, the most popular are Comments Piker, Export Comments, Instant Scrapper and Facepager. Each tool has varying capabilities with respect to characteristics of the required data. As per nature of our required data, Facepager is the most authentic Facebook data scraping tool. These data collections are briefly discussed as follows:

Comment Picker: Comments Picker tool provides comments data which is extracted via scrapping. The post comments data can be extracted in both excel and CSV file with data-time, comments including replies, likes and content. Moreover, comments data can be downloaded from own post. The extracted post comments data from Facebook can be used further for multiple purposes, such as sentiment analysis.

Export Comments: Facebook comments can be extracted free of cost from "Export Comments" website in both format such as CSV and EXCEL. There is a limit with "Export Comments" tool which provide no more than 500 comments free. The comments can be extracted with help of post URL without Sign up. In URL address bar, put URL of specific post to extract comments data along with the poster's name, the date and time, and the number of likes.

Instant Data Scraper: Instant Data Scraper is a computerized data extraction tool for website. The benefit of Instant Data Scraper is to extract data automatically from any website rather than Facebook post only. The Instant

Scraper provides comments data from unlimited posts. It automatically scrolls complete page and extract all comments from website. Moreover, it is not only extract data from web pages but also exports it as Excel or CSV files. It works with heuristic Artificial Intelligence analysis of HTML structure to extract comments data. It is more convenient other than "Export Comments" and "Comments Piker" tools. To use this tool, no need any code and scripts. It provides free extension for comments extraction. There are multiple features are available in Instant Scraper tool such as extraction and detecting of dynamic data. In addition, it provides support for extracted data with copy and paste. Extracted data can be downloaded in both spreadsheet and CSV file. The column of spreadsheet can be renaming and filtering.

NETVIZZ API: Netvizz API tool help us to extract data from the Facebook and it also gives some visualizations in graph which help us for comparison of the data. It uses Facebook API for data extraction and the data can be extracted by selecting date. It is also freely available with no limit. It can extract likes and reactions of respective comments. But, in depth reply comments not extracted by Netvizz API. There are no options to reset presets according to user interest. Moreover, sometime Netvizz tool API does not respond well.

Octaparse: Octaparse help us to get social media data, most notably commentary tracks, from Youtube, Twitter, Facebook, Amazon, Instagram, and other website. This tool helps us a lot to get data easily without any restrictions. It uses API of respective social network and web. It also extracts data according to date by selecting nodes (likes, reactions and reply comments also). It is also easy to use. There are no options to reset presets according to user interest. Moreover, it is not freely available for users.

3.1.1. Facebook Graph API

The Graph API is a primary way for apps to read and write to the Facebook social graph. It helps to get the data into and out of Facebook platform. All of Software development kits (SDKs) and products interact with the Graph API in some way, and other APIs are extensions of the Graph API, so understanding how the Graph API works is crucial.

The Graph API is named after the idea of a "social graph" - a representation of the information on Facebook. It is composed of nodes, edges, and fields. Typically, we can use nodes to get data about a specific object, use edges to get collections of objects on a single object, and use fields to get data about a single object or each object in a collection.

An access token is an opaque string that identifies a user, app, or page and can be used by the app to make Graph API calls. When someone connects with an app using Facebook Login and approves the request for permissions, the app obtains an access token that provides temporary secure access to Facebook APIs. Access tokens are obtained via a number of methods. The token includes information about when the token will expire, and which app generated the token. Because of privacy checks, the majority of API calls on Facebook need to include an access token.

An efficient implementation provided by Facebook is Facepager which is used to scrap data on Facebook. The data extracted by Facepager is further processed and analyzed to produce results (Abuein & Shatnawi, 2019).

3.1.2. Facepager

Facepager is a data extraction tool which retrieve data automatically from Facebook. Facepager is developed by Jakob Jünger and Till Keyling in 2019 as a data extraction tool which retrieve data automatically from Facebook. Facebook Graph API works at backend of the Facepager to get data extraction. Facepager is usually used for publicly available data on Facebook. The data can be exported into CSV file (Leung et al., 2018). Facepager extract data of Facebook page names groups and give permission to collect multiple post, likes and other type of data. For data collection, Facepager is used in this research. The Facepager application tool requires Facebook page key as an input which provides access to all posts along with their comments, pictures, and videos. The available review of posts can be explicated in comma-separated format (Nazir et al., 2019).

The Facepager provides number of different presets which can help to get data on limited time periods. These presets may be prolonged with different parameters to get other type interested data. To extract comments from the Facebook pages, we adhere to following steps:

1. Login to Facebook in Facepager Application tool and get Token access.
2. Facepager allows to user to create new database which store the data by clicking on the new database button and give it a filename.
3. Find out the Facebook Page ID which is numeric in nature. The Facebook Page ID is extracted from the URL "*https://lookup-id.com*".
4. The numeric code will be generated against given URL address which is mentioned above and put into "Add Nodes" tab for comments extraction.
5. Choose the presets of Facebook and then apply the GET Facebook Posts which automatically set the Parameters "/<page-id>/posts" in the Facepager. Presets give opportunity to explicate comments data by specific date which can be edit with the help of parameters according to our requirement time period.
6. To get Facebook page comments, we have to first extract the post by selecting the Presets "Get Facebook Page" and then Select all the post and resource parameter to "/<post-id>//comments" then click on "Fetch Data".

7. For more deep results to extract reply comments select all the comments and then select the resource parameter to "/<post-id>//comments" which give the reply comments of the selected comments by clicking on Fetch data.
8. Data extracted in CSV or Excel by selecting the Seeds and then click on "Export Data".

Facepager outperform the other existing extraction APIs with the provision of flexible set, reset options, date specification, extraction of comment replies, and free access are the key advantageous aspects of Facepager. In addition to these Facepager is responsive and fast in comparison with other tools available in the market. Facepager also provides the flexibility to the users to extract the relevant data by setting the parameters which are convenient for the users.

The two mechanism, comment picker and Facepager were dominant on Facebook data extraction but among these two Facepager reached to optimistic level in terms of flexibility and coverage of contents

A comparison of Facepager with Comment Piker is presented in Table 2.

Table 2: Comparison of Facepager and Comment Piker Tools

Features	Facepager Tool	Comment Piker Web Tool
Availability	Freely available tool placed in GitHub and other web portal download links.	Comment Piker is just a web portal and has a subscription fee to use for more than 14 days.
Responsiveness	It used the latest updated Facebook API's which are fast and responsive for developers to use them.	This web tool is used the Facebook Comments plugin which is embedded on an external website to extract comments.
Commenting	It allows both getting Public Page comments and personal Page comments.	This tool allows only to get personal page information, has no permission to extract comments from the public page.
Scope of extraction	Facepager extracts both comments (Text and pictures) and reactions of the comments which is better to know the feeling of the people.	Comment Piker only extracts the comment which is in text form.
Time Specification	It allows the user to extract information from Facebook with a time limit as required by the user.	Comment Piker has not an option to set the required time limit by the user.
Speed of Extraction	Fast, easy and responsive for users to extract data from Facebook.	Due to the web-based tool, it is not much fast and responsive as compared to the Facepager tool.
Customization	Facepager gives flexibility to users for the setting of the required data columns which are useful for users instead of getting irrelevant data.	It does not allow the users, to select the output data column which is useful for the users.

3.2. Youtube Comments Extraction Tools

Youtube is a famous social media platform. It was launched in 2005 by Chen, Hurley and Karim for online video sharing. It is owned by Google in 2006. Since its launch, it has grown as video-sharing website by gaining meteoric popularity. According to Statista (2021), the number of Youtube users was projected from 28 million to 33 million from the year 2021 to 2025. From so many Pakistani users on Youtube, the analysis of the comments posted by users will also help us in the current project.

User's comments data from the Youtube channel are extracted by using different tools with/without using APIs. The most popular freely available tools used for the extraction of the data from Youtube are NetVizz-Youtube Data Tool, Facepager and Youtube Comments Suite.

3.2.1. NetVizz-Youtube Data Tool

NetVizz-Youtube Data Tool (YTD Tool) was introduced by Bernhard, Matamoros-Fernández, and Coromina, (2018). This tool is freely available and have utilized in many investigations (Bernhard et al., 2018; Jünger and Keyling, 2019) for downloading of data from Youtube. This tool uses 'Youtube API-V3' to extract data from this social media platform. This tool has interesting features for extracting Youtube data in terms of Search Youtube Channel, Channel Network, Information about Channel, Video Network, Video Info and Video List. The most important feature of this tool is to extract data of specific country. As our project is relevant of Pakistani users' comments, therefore this tool allows us to extract and download data related to Pakistan. The country code for Pakistan is "PK". Another feature is time duration that is, Youtube data from a particular time period. For this project, data from 01-01-2021 will be extracted for the predefined keywords of the current project.

From the Youtube Platform, YTD Tool is utilized. The working steps are summarized as:

1. Selection of Keywords

2. By entering the selected keyword, country code "PK" and time duration (e.g., 01-01-2021 to 31-12-2021) in the 'Video List' part of YTD Tool.
3. Video list after searching the data related to the desired keyword generates a file having information of channel title, corresponding channel Id, Video title, corresponding Video Id, date of publishing, comment count, view count and like & dislike count
4. Video Id extracted by the corresponding tool is related to our predefined keywords and location
5. Finally, the Video ID from the above step are used on Video Info tool to extract the comments on this particular video. This step generates four files namely: first file has basic information and statistics about the video; second file contains comments; third file shows comment count and comments author, and fourth file has a comment network.

3.2.2. Facepager for Youtube Comments

Facepager tool was introduced by Jakob Jünger and Till Keyling to extract available data from websites and social media platforms like Twitter, Facebook and Youtube (Keyling, 2015). This tool uses web scraping and APIs for extracting data. It may be used for Windows, Mac OSX and Linux operating systems and can be download from Facepager Source Code Site (2022).

The brief introduction of the steps involved for data extraction from Youtube using Facepager are: creating a blank database by clicking "New database"; logging to user's Goggle account for access token; adding the name/names of Youtube urls in the "Add Nodes" section; setting the desired parameters like "Base Path, Resource, Object ID, Time etc." in the "Setup query" section; calling "Fetch data" after selecting the specific node or multiple nodes in "Nodes View" section; the extracted data may be seen in "Data View" section; column setup may be used to clear or add the columns according to the desire of the user and finally data can be exported into CSV files by clicking "Export data". The detailed usage of how to extract comments from Youtube with Facepager is presented in Youtube Video (Facepager, 2020).

3.2.3. Youtube Comments Suite (YCS)

Matthew Wright introduced "Youtube Comments Suite (YCS) v1.4.7". It is developed in JAVA language. This suite can be used on both operating systems: windows and Linux. It is available in complied form and in source form. This suite also provide opportunity to download comments from Youtube without Youtube API. YCS can be downloaded from GitHub site (GitHub, 2022).

In the *search Youtube section*, user may enter desired keyword. Searching may be performed by choosing any parameter from *relevance*, *date*, *title*, *video count*, *view count* and *rating*. The optional arguments for search are *location* and *normal*. On choosing *location* argument, this suite provides current Google Location with distance range from 1 KM to 1000 KM. In other words, this YCS will search Youtube in this distance range. The search result of 3798 videos from Youtube are listed. First select these videos and then add in the *Group Manage* section.

The comparison of the above stated tools for downloading comments from Youtube is presented in Table 3. As the Facepager gives the best response to extract data from the Youtube channel but it does not allow the users to extract data with our predefined keywords. Moreover, it does not give the best option for the selection of the location. As in our research project, the user's comments required from the Youtube channel are to be extracted for predefined keywords and location. While the corresponding YTD Netvizz tool provides the flexibility to the users to extract Youtube comments data according to their requirements. For the current research project, to extract the user's comments from the Youtube social media, Netvizz -Youtube Data Tool (YTD Tool) is utilized. As depicted in Table 3 Netvizz YTD is dominant among other YouTube comments' extraction tools with the provision of keywords-based searching of country as well as time specific contents.

Table 3: Comparison of Youtube Comments Downloading Tools

S. No.	Name of Tool	Search videos by Key Words	Need user's Google API	Search for Specific country/Region	Search videos For Specific time-period	Download comments from single Youtube URL	Download Comments from multiple Youtube URLs
1	YTD Tool	Yes	No	Yes	Yes	Yes	Yes
2	Facepager	No	Yes	No	Yes	Yes	Yes
3	Youtube Comments Suite	Yes	No	Yes (but Google location from 1 to 1000 KM range)	No	Yes	Yes

3.3. TWITTER DATA EXTRACTION TOOLs and APIs

Twitter is a microblogging service that allows 280 characters for sharing and posting expressions, opinions, and suggestions. These short messages are referred as tweets. A tweet can be factual information, opinion, expression, sentiment, or emotion. The accessibility to post, likes and retweets is allowed to registered users only whereas unregistered users are allowed to read the posts. Twitter has now become the key source for extraction and analysis

of user generated contents to big data analysts and observers. Below we have provided tool wise description and comparative analysis of extraction APIs, Scrappers, and tools.

3.3.1. TWEEPY

Tweepy is one of the popular used open-source Python based package that gives us a precise way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details, such as data encoding and decoding, HTTP requests, result pagination, OAuth authentication, rate limits and streams. For using the Twitter API, we just need, Consumer_key, Consumer_secret, Access_key, and Access_secret.

Tweepy gives access to most of Twitter's functionality, some of the advantages of using Tweepy for data extraction are creditability of data, direct interaction without any intervention of third-party dependency and thorough documentation whereas duration restriction and requirement of additional preprocessing are the limitations of Tweepy.

3.3.2. ORANGE

The ORANGE data mining tool help us to extract data from Twitter by using the Twitter API key. It also gives research results against hash tags and keywords, and an option to extract tweets for specific language. Further it is easy to use, freely available keyword-based data extraction tool but it does not provide facility to extract date specific data. We performed further preprocessing for date specific selection of twitter contents.

3.3.3. TWINT

Twint is also one of the widely used progressive Twitter scrapper scripted in Python that allows for extracting Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search operators to let us scrape opinionative tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers. Twint also makes special queries to Twitter allowing us to scrape a Twitter user's followers, tweets a user has liked, and who they follow without any authentication, API, Selenium, or browser emulation. Requirements, advantages, and limitations are as follows:

To get Twint up and running following decencies must be installed:

1. Python 3.6 or higher	2. aiohttp	3. aiodns
4. beautifulsoup4	5. cchardet	6. dataclasses
7. elasticsearch	8. pysocks	9. pandas
10. aiohttp_socks	11. schedule	12. geopy
13. fake-useragent	14. py-googletransx	

Twint is advantageous in terms of free availability, and it doesn't require access token, API and rate limit but it has few limitations such as its documentation is insufficient, and it doesn't provide metadata and country specific contents. Further data extraction mechanism of Twint is comparatively complicated.

3.3.4. TAGS: Twitter _Archiving _Google _Sheet

TAGS is One of the widely used publicly available Google Sheet pattern which allows us to setup and run automatic set of search results from Twitter. The TAGV6 is prominent among others as it has interactive interface, and it is free tool used in extracting country as well as time specific keywords-based information without any explicit access code. Following steps are followed to extract data using TAGS:

- The first step is to setup Twitter access, it requires users to login to their Gmail and Twitter account
- Next just simply enter the search term under Enter term and enter the value of 18000 (Tag V6.1 can extract maximum of 18000 tweets in a single go) under number of tweets.
- Then go to tags and select "Run now!"
- Once the script is finished go to file and select download/comma separated values(.csv)

As TAGS is the only tool that can provide the information about city/country of the user. So, we write a function that can do filtration based on this information. Comparative analysis shown above in Table.4 clearly demonstrates that TAGS is capable to reach the target objectives. It is observed that TAGV6 has interactive easy to use interface that makes it possible to extract country as well as time specific data without any explicit coding.

3.4. Web Blogs Scrapping Tools

A blog is a discussion or information available on the World Wide Web consisting of distinct, often informal diary-style text entries. Posts are usually presented in reverse chronological order, so that the most topical post appears first, at the top of the web page. The prevailing mechanisms used in weblog data extraction are selenium and web scraper.

3.4.1. SELENIUM

Selenium is an umbrella project for a range of tools and libraries that enable and support the automation of web browsers. It provides extensions to emulate user interaction with browsers, a distribution server for scaling browser allocation, and the infrastructure for implementations of the W3C WebDriver specification that lets us write interchangeable code for all major web browsers.

Blogging Sites don't have any official API, so we are using a selenium bot to extract relevant information from the website. Selenium directly interacts with the HTML structure of a website.

Table 4: Comparative Analysis of Tools used to extract Text from *TWITTER*

S. No.	Tools/APIs	Introductory Features	Pros	Limitations
1	Tweepy	Official API	• Official API may reflect the originality of contents. • Contents verification have already been assisted	• Can provide only data of last seven days • Extracted contents may not contain geolocation • Text Size Limit • Expensive to Use
2	Twint	Python based Script to extract keywords-based data.	• Twint doesn't have rate limitations. • It doesn't require any API key or access token. • It doesn't have monthly quota. • Twint can extract data of any date in the past. • Twint is totally free as compared to Tweepy which is sometimes expensive based on the services.	• Official documentation is insufficient and not very extensive • The process to get it up and running is quite complicated. • It doesn't provide any information about the country from where the tweet was published. • Twint uses beautiful soup to scrape data which makes the overall process slower as compared to Tweepy. • It doesn't provide enough metadata
3	Tagv6	TAGS is a free Google Sheet template and work like scrapper.	• It doesn't require any API key or access token • It provides information about city/country of user • It comes with an interactive and easy to learn GUI i.e., It doesn't require code to write.	• It has a rate limit of 18000 tweets.

3.4.2. WEB SCRAPER

Web Scraper is freely available tool for both web-based and cloud-based data extraction. Data scraping from the blogs Web Scraper needs to access the target website or blog. For this, it is integrated into developer tools of the browser.

For the data scraping of the web, the 1st thing is to create a new sitemap which is consists of the sitemap name and URL of the website from which the scraping is started. For multiple search terms, the start URL can be increased by clicking on the "+" tab. After that, by creating the sitemap, the selector will be added which is required. For extracting text from the web text selector is added and for URL links the link Selector is used. By clicking on the data preview, we can check out the selected data is accurate or not. Moreover, the Web Scraper tool allows exporting data in both the Microsoft Excel and CSV files with accurate format without any impurities.

4. Results and Discussion

Experimental evaluation is performed to extract the valuable contents for the prediction of consumer confidence index. Data is extracted from four key platform namely: Facebook, Youtube, Twitter and blog for the CCI indicators considered in NACOP. Figure 3 presents the NACOP model which comprises of two major sections; first is real time extraction and annotation of social media contents and second is sentiment analysis and CCI prediction.

This study contributes to the first step of model presented in Figure 3 and a huge collection of user generated contents about purchasing behavior, job employment, consumer price and personal finance are extracted from all four social networking platforms. Secondly these contents are cleaned, annotated, and categorized according to CCI subjected to multiple data science approaches to ensure relevancy of the outcome in predicting the consumer sentiment and the public mood. As mentioned earlier four social networking platforms are considered for extraction and annotation of user generated contents and a total of 14 different tools, APIs and scrapers are assessed and compared as shown in Table 5.

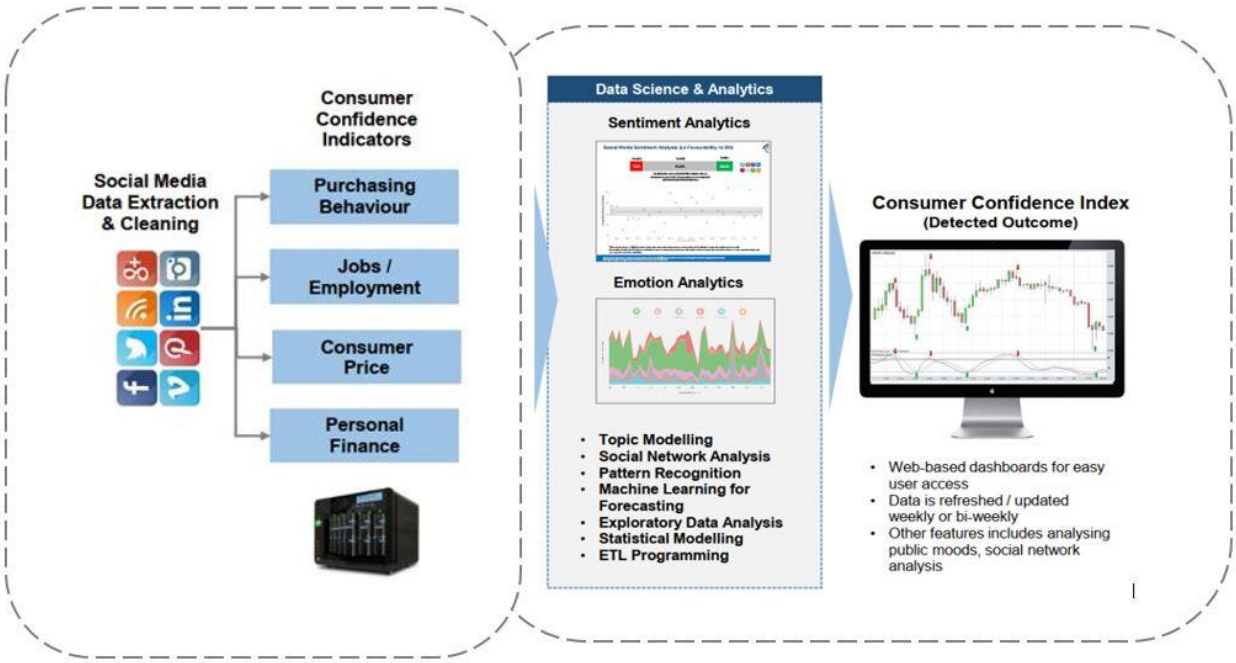

Figure 3: NACOP Model

Table 5: Comprehensive comparison of existing data extraction tools and APIs

S.NO	TOOLS	DESCRIPTION	LIMITATIONS
		1. FACEBOOK DATA EXTRACTION TOOLS	
1	Comment Picker	This web tool is used the Facebook Comments plugin which is embedded on an external website to extract comments. They offer choices for excluding comments based on the number of friends who have been tagged, comments from the same user, like on the post, or a certain text.	It requires subscription This tool provides no permission to extract comments from the public page. This tool does not provide data with a time frame.
2	Export Comments	Users don't have to be the owner of Youtube, TikTok, Facebook, Instagram, or any other social media page that published the post. They can assist in exporting the comments whether it is a status, image, video, or URL as long as it is accessible to the public. You can export all of the accessible comments by just pasting the post's URL.	Provide data for just one post at a time. Extract only text comments This tool does not provide data with a time frame.
3	Instant Data Scraper	Data is extracted from web pages by Instant Data Scraper and exported as Xls or CSV files. Any website can use the automated data extraction technology known as Instant Data Scraper. It utilizes AI to determine which information on an HTML page is the most important and allows saving it to an Xls or CSV file (XLS, XLSX, CSV).	It does not provide any good formatting data. Difficult to tackle or use this tool. It is difficult to use for many posts comment extraction at a time.
4	NETVIZZ API	Netvizz provides export data in common file formats from many areas of the social media platform using the data collection and mining tool Netvizz on Facebook. With the help of Netvizz, friendships, groups, and webpages can have their demographic, post-demographic, and relational aspects numerically and qualitatively examined	This tool provides useful results, but this API is not in working condition since 2019. In-depth reply comments not extracted by Netvizz API.
5	Graph API	The methodology for transferring data to and from the Facebook network is through the Graph API.	It does not provide help to extract public

		Apps can use this HTTP-based API to upload photographs, handle ads, post new stories, and do a broad range of other functions programmatically.	data directly from Facebook.
6	Facepager	Facepager freely available tool to download social media data, most notably commentary tracks, from Youtube, Twitter, Facebook, and Amazon. The tool comes with several presets that may be expanded with parameters to obtain the data of interest. Using the Facebook API directly is too difficult to tackle and extract data. So, we used the tool Facepager although in the backend it used the Facebook API, to extract data from Facebook.	Extracting a large dataset becomes slow sometimes.
		OPTIMUM DATA EXTRACTION TOOL OF FACEBOOK "FACEPAGER"	
2.	YOUTUBE COMMENT EXTRACTION TOOLS		
7	Youtube Comments Suite	Youtube Comment Suite is used to collect comments from a variety of videos, albums, and channels for preservation, general query, and activity display. Realize the Communities > Comments tool's capabilities, as well as additional features.	It is difficult to utilize for non-technical people because it required Java 11+ extension.
8	NetVizz Youtube	This tool is freely available and has been utilized in many investigations for downloading data from Youtube. This tool uses 'Youtube API-V3' to extract data from this social media platform. This tool has interesting features for extracting Youtube data in terms of Search Youtube Channel, Channel Network, Information about Channel, Video Network, Video Info, and Video List. The most important feature of this tool is to extract data from the specific country	It does not provide output files in CSV format directly.
9	Youtube Facepager	Facepager freely available tool to download social media data, most notably commentary tracks, from Youtube, Twitter, Facebook, and Amazon. The tool comes with several presets that may be expanded with parameters to obtain the data of interest. Using the Facebook API directly is too difficult to tackle and extract data. So we used the tool Facepager although in the backend it used the Facebook API, to extract data from Facebook.	It does not offer keyword-based Youtube research. It won't offer country specific data.
		OPTIMUM COMMENT EXTRACTION TOOL OF YOUTUBE "NETVIZZ YOUTUBE"	
3.	TWITTER DATA EXTRACTION TOOLS		
10	Twint	Twint is an open-source Python tool used for Twitter scraping, which means we can use it to retrieve data from Twitter without utilizing the Twitter API. Twint has several features that distinguish it from most Twitter scraping APIs, including Twitter API has a restriction of 3200 (last) tweets that may be downloaded, whereas Twint has almost no limit and can download practically all tweets.	It is difficult to utilize for non-technical people because it required a bit of coding skills.
11	Tweepy	Tweepy is an open-source Python program that provides a very straightforward way to use Python to access the Twitter API. Tweepy contains a set of classes and functions that represent Twitter's models and API endpoints, and it handles numerous implementation details transparently, such as: • Encoding and decoding of data • Requests for HTTP • Pagination of results • Authentication using OAuth • Rate Ends • Streams	It is difficult to utilize for non-technical people because it requires a python library. Does not provide a deep level search.
12	Tagv6	TAGS is a free Google Sheet template that enables you to set up and manage automated Twitter search result gathering.	Provide results of last fifteen days.

		It provides research both by keyword and also by location of the people. It generates an automated dataset which can easily be preprocessed. OPTIMUM DATA EXTRACTION TOOL OF TWITTER "TAGV6"	
4.	WEB BLOGS DATA SCRAPPING TOOLS		
13	Selenium	Selenium was not created with web scraping in mind. Selenium is a web driver made to generate web pages for web application test automation. Since many websites operate on JavaScript to provide dynamic content for web pages, Selenium is excellent for web scraping.	Selenium demands your team's skill and management of resources. Selenium does not provide reporting-based test visibility.
14	Web Scrapper	This is freely available for both web-based and cloud data extraction. Data scraping from the blogs Web Scraper needs to access the target website or blog and by using the developer it helps easy to scroll all over the pages of the website and extract the data into an output file in CSV. OPTIMUM COMMENT EXTRACTION TOOL OF YOUTUBE "WEB SCRAPPER"	This tool does not have any mobile application yet.

Figure 4. presents the dataset statistics, it shows that a massive volume of user generated contents according to the keywords of purchasing behavior, job employment, consumer price and personal finance are accessed through Weblogs, Twitter, YouTube, and Facebook. A total of 312142 tweets covering (January 2021 to August 2022) are collected according to keyword-based information of each behavioral indicator. 75076 tweets are from Facebook; out of which 17004 (23%), 24984 (33.3%), 28371 (37.8%) and 4717 (6.2%) tweets are related to the indicators of Purchasing Behavior, Employment, Consumer Price, and Personal Finance respectively. 53088 tweets are from Youtube; out of which 18296 (34.5%), 10761 (20.3%), 15032 (28.3%) and 8999 (17%) tweets are related to the indicators of Purchasing Behavior, Employment, Consumer Price, and Personal Finance respectively. 168883 tweets are from twitter; out of which 30682 (18.2%), 47734 (28.3%), 57293 (34%) and 33070 (19.6%) tweets are related to the indicators of Purchasing Behavior, Employment, Consumer Price, and Personal Finance respectively. 15095 tweets are from Weblog; out of which 4219 (28%), 3293 (28.3%), 4375 (29%) and 3208 (21.3%) tweets are related to the indicators of Purchasing Behavior, Employment, Consumer Price, and Personal Finance respectively. These contents are then normalized and annotated according to subjective information of each indicator. We further decomposed this data into month as well as channel wise categories to make our analysis more relevant and transparent. Secondly these contents are cleaned, annotated, and categorized according to CCI subjected to multiple data science approaches to ensure relevancy of the outcome in predicting the consumer sentiment and the public mood. A total of fourteen different tools, APIs and scrapers are assessed and compared for Facebook, Twitter, Youtube, and Blogging sites. Based on implementation-based comparison of the data extraction tools with respect to inclusion criteria and the target platforms presented in Table 5, the most suitable data extraction tool for Facebook is Facepager out of Comment Picker, Export Comments, Instant Data Scraper, NETVIZZ API and Graph API, for Twitter is Tagv6 out of Twint, Tweepy, Tagv6, and Facepager, for Youtube is NetVizz Youtube out of Youtube Comments Suite, NetVizz Youtube, and Youtube Facepager, and for Web blog is Web Scrapper out of Selenium and Web Scrapper.

5. Implications and Future Research Area

The study has significant implications to theory and practice. This study provides an implementation-based comparative literature review to unfold the best possible techniques of text extraction and indicator wise annotation for consumer confidence index to disclose the true picture of economy. Real-time extraction has been performed and optimized mechanisms are unfolded to contribute to the field of big data analysis and CCI. The major contributions of this review are: (a) a classification of fourteen tools according to the social media sites and blogsites, and (b) an extensive evaluation of the tools performed through several experiments on real data extracted mainly from Twitter, Facebook, Youtube, and different blogsites.

The experimental evaluation revealed that Facepager, Netvizz YouTube, Tagv6 and web scrapper are the optimum extraction APIs for Facebook, YouTube, Twitter, and Blogsites respectively. These extractors are not only capable of mining the user generated contents from Facebook, YouTube, Twitter, and Blogsites. but it also considers the keywords based informative tweet, date of publishing of tweet, region of the opinion holder and indicator wise annotation. The identification of platform relevant data extraction tools can save time and effort of practitioners in selecting the data extraction tools for extracting the quality data of users' generated comments on social media

sites and blogsites. The study also contributes to social media data extraction literature by examining and evaluating the fourteen different data extraction tools based on specified criteria and proposing the most optimal data extraction tools of Facepager, Netvizz YouTube, Tagv6 and web scrapper for Facebook, YouTube, Twitter, and Blogsites respectively. These tools can provide quality inputs for measuring the consumer confidence.

As all studies have limitations, this study also has a few limitations that to be noted for future studies. First, this study has evaluated the extraction tools with respect to four indicators, future studies may expand the generalizability of these data extraction tools to explore the multidimensionality of the users' generated contents. Second, this study focusses individual data extraction tools, future research can use the APIs in consolidated manner as a single platform. Third, this study aims to evaluate the existing data extractions and decide the optimal tool with respect to social media channels and inclusion criteria. We encourage future researchers to use these optimal data extractions for extracting the data and using it further for computing the CCI.

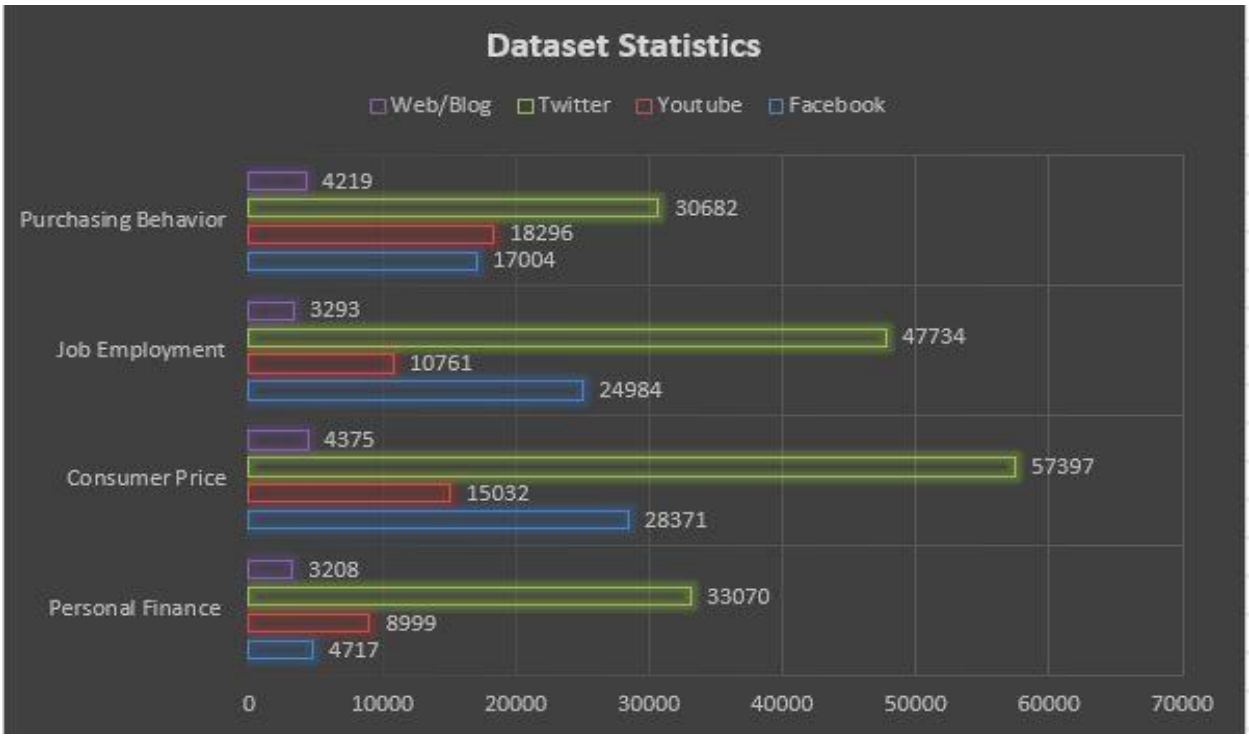

Figure 4. Dataset statistics

6. Conclusions

The CCI is referenced by businesses, governments, and other institutions when they make strategic decision. Ashraf, Raza, and Ishaq (2022) proposed an approach to social media analytics for predicting national consumer confidence predictor as "NACOP" with the aim of rolling-out a viable commercial product/solution in the form of NACOP web-based portal. NACOP utilizes big data and microbloging sites to predict the national consumer confidence. Microbloging sites and social media channels can have a high volume of data on consumer confidence, analyzing such contents can significantly improve the impact and accuracy of CCI but unavailability of consolidated application for real-time extraction of user generated content is restricting the further process. This study proposes a comparative analysis to unfold the best possible mechanism of real-time text extraction and indicator wise annotation for consumer confidence index to disclose the true picture of economy. A case study of four indicators of purchasing power, personal finance, Job/Employment and consumer price is presented to explore the best possible mechanism of user generated data extraction on Facebook, Twitter, Youtube, and Blogsites. The experimental evaluation revealed that Facepager, Netvizz YouTube, Tagv6 and web scrapper are the optimum extraction APIs for Facebook, YouTube, Twitter, and Blogsites respectively. It is concluded that above mentioned APIs can be used effectively in consolidated manner as a single platform. We must encourage future researchers to actively participate to compute the CCI using social media contents.

References

Akhilesh Ganti (2020). *What Is the Consumer Confidence Index (CCI)?* https://www.investopedia.com/terms/c/cci.asp (Accessed on December 13, 2021).

Ashraf, M., Raza, A. A., & Ishaq, M. (2022). *A novel approach of social media analytics for predicting national consumer confidence index*. Bulletin of Business and Economics (BBE), 11(2), 220-234.

Asur, S. & Huberman, B. A., (2010). *Predicting the Future with Social Media.* Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference. IEEE/WIC/ACM, pp. 492 – 499.

Audi, M., & Ali, A. (2019). The advancement in Information and Communication Technologies (ICT) and economic development: a panel analysis. *International Journal of Innovation, Creativity and Change,* 15 (4), 1013-1039.

Audi, M., Ali, A., & Al-Masri, R. (2022). Determinants of Advancement in Information Communication Technologies and its Prospect under the role of Aggregate and Disaggregate Globalization. *Scientific Annals of Economics and Business*, *69*(2), 191-215.

Barry, Chris; Markey, Rob; Almquist , Eric; Brahm, & Chris, (2011). *Putting Social Media to Work.* http://www.bain.com/Images/BAIN_BRIEF_Putting_social_media_to_work.pdf (Accessed on December 14, 2021)

BBC, (2014). *Boom and Bust.* http://www.bbc.co.uk/bitesize/higher/history/usa/boombust/revision/1/ **(**Accessed 23 April 2015)

Blackshaw, P. & Nazzaro, M., (2004). Consumer-Generated Media (CGM) 101: Word-of-Mouth in the Age of the Web-Fortified Consumer. *A Neilsen Buzz Metrics White Paper*, Spring.

Bloomberg News, (April 2003). Consumer Confidence Shows a Substantial Gain. *The New York Times*, p. 8.

Brian Dean (2021). *Social Network Usage & Growth Statistics: How Many People Use Social Media in 2021?* https://backlinko.com/social-media-users (Accessed on September 6, 2021)

Cheng, Yao (2020). The Determinants, Implications and Interaction of Consumer Sentiment, Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/13769/

Chung, S., Shin, D., & Park, J. (2022). Predicting Firm Market Performance Using the Social Media Promoter Score. *Marketing Letters*, 1-17.

Curtin, R. T., (2002). *Surveys of Consumers: Theory, Methods, and Interpretation.* Washington DC, s.n

Hirt, M., & Willmott, P. (2014). Strategic principles for competing in the digital age. *McKinsey Quarterly*, 5(1), 1-13.

Igboayaka, J. V. C. E. (2015). *Using social media networks for measuring consumer confidence: Problems, issues and prospects* (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

Investopedia, 2014. *Capital Investment.* http://www.investopedia.com/terms/c/capital-investment.asp (Accessed on December 13, 2021).

Keyling, T., & Jünger, J. (2015). *Observing online content. In Political Communication in the Online World* (pp. 183-200). Routledge

Li, X., Xu, M., Zeng, W., Tse, Y. K., & Chan, H. K. (2023). Exploring customer concerns on service quality under the COVID-19 crisis: A social media analytics study from the retail industry. *Journal of Retailing and Consumer Services*, 70, 103157.

Mangold, G. W. & Faulds, D. J., (2009). Social media: The new hybrid element of the promotion mix. *Elsevier*, 52(4), p. 357–365.

Margaret, R., (2011). Capex (capital expenditure). http://whatis.techtarget.com/definition/CAPEX-capital-expenditure (Accessed on April 11, 2021).

McKinsey & Company, (2012). *The social economy: Unlocking value and productivity through social technologies.* https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-social-economy (Accessed December 12, 2021).

Mueller, E., (1963). Ten Years of Consumer Attitude Surveys: Their Forecasting Record. *Journal of the American Statistical Association*, 58(304), pp. 899-917.

Netbase, (2010). *How Does Netbase Achieve the Best Accuracy for Understanding Consumers Online?* https://neurorgs.net/wp-content/uploads/docs/NetBaseR6.pdf (Accessed on January 5, 2022).

Nielsen, (2013). *Consumer Confidence Concerns and Spending Intentions around the World.* https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/nielsen-global-consumer-confidence-q1-2012.pdf (Accessed on December 11, 2022).

Odendaal, H., Reid, M., & Kirsten, J. F. (2020). Media-Based Sentiment Indices as an Alternative Measure of Consumer Confidence. *South African Journal of Economics*, *88*(4), 409-434.

Pendery, D., (2009). Three top economists agree 2009 worst financial crisis since great depression; risks increase if right steps are not taken. *Business Wire News.* *http://www.predella.it/archivio/indexa0f3.html?option=com_content&view=article&id=197&catid=81&Itemid=108 (Accessed on December 13, 2021)*

Performics, (2010). *Social Networking Study: Facebook Use Continues to Rise; Brand Participation and Engagement Heavily Welcomed by Social Networkers.* http://www.performics.com/social-networking-study-facebook-use-continues-to-risebrand-participation-and-engagement-heavily-welcomed-by-social-networkers/ (Accessed on April 12, 2021).

Petev, I. D., Pistaferri, L., & Saporta-Eksten, I. (2011). Consumption decisions are crucial determinants of business cycles and growth. Personal consumer. *The Great Recession*, 161.

Roberts, I. & Simon, J., (2001). *What Do Sentiment Surveys Measure?* http://www.rba.gov.au/publications/rdp/2001/pdf/rdp2001-09.pdf (Accessed on December 13, 2021).

Schweidel, D. A., Moe, W. W. & Boudreaux, C., 2012. Social Media Intelligence: Measuring Brand Sentiment from Online Conversations. *Marketing Science Institute*, pp. 12-100.

Shayaa, S., Ainin, S., Jaafar, N. I., Zakaria, S. B., Phoong, S. W., Yeong, W. C., ... & Zahid Piprani, A. (2018). Linking consumer confidence index and social media sentiment analysis. *Cogent Business & Management*, *5*(1), 1509424.

Shayaa, S., Al-Garadi, M. A., Piprani, A. Z., Ashraf, M., & Sulaiman, A. (2017, December). Social media sentiment analysis of consumer purchasing behavior vs consumer confidence index. In *Proceedings of the International Conference on Big Data and Internet of Thing* (pp. 32-35).

State Bank of Pakistan (2021). Consumer Confidence Survey. https://www.sbp.org.pk/research/CCS.asp (Accessed on December 14, 2021).

Statista (2021a). Most popular social networks worldwide as of July 2021, ranked by number of active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (Accessed on September 8, 2021).

Statista (2021b). Number of global social network users 2017-2025. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Accessed on September 7, 2021).

The Conference Board of Canada, (2014). *Consumer Confidence.* http://www.con263ferenceboard.ca/topics/economics/consumer_confidence.aspx (Accessed on April 6, 2021).

The Conference Board, (2011). *Consumer Confidence Survey Technical Note - February 2011.* https://www.conference-board.org/pdf_free/press/TechnicalPDF_4134_1298367128.pdf (Accessed on December 14, 2021).