Finding Influential Users in Social Networking Using Sentiment Analysis

Shaha Al-Otaibi, Amal Al-Rasheed*, Bashayer AlHazza, Hafsa Ahmad Khan, Ghadah AlShfloot, Maram AlFaris, Noura AlFari, Norah AlKhalaf and Nuha AlShuweishi
*corresponding author
E-mail: stalotaibi@pnu.edu.sa, aaalrasheed@pnu.edu.sa, 436004069@pnu.edu.sa, 435200601@pnu.edu.sa, 437004049@pnu.edu.sa, 436000492@pnu.edu.sa, 436001915@pnu.edu.sa, 436005646@pnu.edu.sa, 436003182@pnu.edu.sa

Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Keywords: Instagram, influencers, feature extraction, VADER, sentiment analysis

Received: November 15, 2021

Social networking platforms facilitate sharing of information, ideas, and thoughts by constructing virtual communities. Finding people in certain social networking sites, who really have the power to influence other users, is critical. For example, this search can be focused on the right person who can impactfully support or contradict any opinion or generate more profits when they publish an advertisement for a certain business or product. The aim of this study is to devise a computational method of finding the most influential users on social networking platforms. In this study, we propose a solution named Muatheer, which helps determine who influential users are on Instagram. The study uses techniques of scraping data using Instagram Application Programming Interface (API) and applying the sentiment analysis algorithm to classify comments as either positive or negative. The study also uses the unigram as a feature extraction method. Then, the sentiment analysis result will be combined with other factors to calculate the "influence ratio", which attempts to determine the actual influencer in a specific domain. These experiments were conducted using a publicly available datasets from Instagram, and the proposed algorithm delivered a highly accurate result more than 85 %.

Povzetek:

1 Introduction

Social networking has emerged as a critical function in information technology through which knowledge and thoughts are disseminated across communities. The rapid development of social networking has increased the need to study and discover influencing roles emerging within these platforms. Hence, the challenge of finding influencers in social networking is imperative for numerous areas—from advertising to general wellbeing.

The nature of online social networks changed traditional social influencing and shifted it to modern modes. In traditional social networks, such as radio and newspapers, users with some sort of influence were ordinarily known to commoners. They were, as a result, naturally found easily and consulted for support, contradiction of any opinion, or renunciation of any product. In online social networks, influential users let followers stay blindly devoted to them; therefore, the quest of these unknown influential users now becomes a vital task. Once found, these influential users can be delegated certain tasks.

In the context of this study's focus, the influencer is someone who is able to persuade a lot of other people—such as their followers on social media—to do something, buy products, or use the same things that they do. They are often paid or are given free products in exchange for doing this. They use social networking to promote to their followers' lifestyle choices and commercial products [1].

Typically, users on social networking sites communicate with each other, but some users optimize their strong influencing skills in a certain niche. Hence, it is evident that human beings are entities with extraordinary potential to psychologically cause changes in their environment; as a matter of fact, modeling the behavior of humans is among the foremost patterns to be studied by the research community.

This area of research is related to social influence analysis (SIA), which demonstrates the spread of influence through social networking. It is preoccupied with the questions of who affects whom, who is more influential, and what affects whom. Further, SIA has been applied to a number of fields of analysis, such as marketing, healthcare, recommendations on the Internet, rumor diffusion, among others, that depend on social influencing [2].

Many individuals and organizations face the problem of choosing an influential user who can deliver the exact

content (such as advertising, public opinion, decision-making) to the target audience via social networks. However, most people and organizations rely on some characteristics (for example, the number of followers and viewers) in order to select the right influencer for them; they ignore the role of user-contributed content on social media, which may lead to choosing the wrong person to deliver their content.

The aim of this study is finding right influencers on Instagram to promote products, knowledge, or opinions. Instagram has seen rapid growth in the number of users as well as views since it was launched in October 2010. In spite of being the most popular photo capturing and sharing application, Instagram has attracted relatively less attention from the research community [3].

In this study, we will determine the significant criteria that affect the influencing process and propose a solution that could help discovering influencers using Instagram content, number of followers, and some other factors. The proposed solution can help foster the efficiency of marketing and affect public opinion in the digital environment and can also be used to gather opinions, determine user behavior, fetch information on specific topics, predict trends, influence public opinions, support public decision-making, and provide companies or general entities with the best influencers [4]. It smoothly enables the comparison of influencers in social networking, which will assist individuals, organizations, and business owners with making their choice among influencers easily and clearly.

The rest of this article is organized as follows: Section 2 presents our findings from the literature review to discover this problem and determine the important criteria that affect the influencing role. Then, in section 3, we propose the framework constructed to find influential users on Instagram and develop a prototype that is tested against the collected data to evaluate the algorithm's accuracy in section 4. Finally, the key findings from this research work are summarized in section 5.

2 Background information and related work

In recent years, the term "social media" has become a part of common parlance in society, especially among the young. It is a new form of entertainment, where users can share and exchange their thoughts, personal information, photos, and interests. This media is growing bigger every day with more and more people using it. Today, there are many social media platforms such as Instagram, Facebook, Twitter, Pinterest, and others. The Merriam-Webster dictionary defines "social media" as "forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos)" [5]. Further, social media analytics (SMA) is defined as "the use of analytics-based capabilities to analyze and interpret vast amounts of semi-structured and unstructured data from online sources, also provides insights into customer values, opinions, sentiments, and perspectives" [6], [7].

Social media has created the foundation for the development of several technological breakthroughs and innovations. Even though some technologies started out without social media as their primary application, they have still proven useful to the social media universe: data mining, machine learning (ML), biometric facial recognition, mind reading, radio frequency identification tags, intelligent personal assistants, motion detectors, virtual reality, augmented/mixed reality [8].

Some terminologies related to this type of research will be explained in this paragraph. The SMA framework is described as the method of representing the relationship between each aspect of the research when it pertains to a theory or scientific research: the aim of the research, its methodology, the method of data collection, and analysis [8]. It includes the process of data collection, data pre-processing, data storage, and sentiment analysis as well as considering official names and popular terms used such as "hashtags". Moreover, SMA techniques, such as sentiment analysis, can be used to analyze the collected data, which is benefited by the huge amount of user-generated content online, natural language processing (NLP), computational linguistics, and text analytics. The overall purpose of such techniques is to determine the attitude of a writer with respect to contextual polarity.

Computational science techniques involve the automated sentiment analysis of digital texts that use elements from ML These techniques employ three areas: computational statistics, complexity science, and ML. The last of these areas is divided into supervised learning such as (regression trees, discriminant function analysis, support vector machines) and unsupervised learning (such as (self-organizing maps and k-means).

Today, most of companies use social media as a marketing tool. For example, a company that manufactures hair products will hire an influencer from social media to try their products. Then, the influencer will advertise and promote these products based on their experiment. Therefore, it is important for companies to choose the right influencers for the promotion of their products, because when it works with the wrong influencer, it loses profit. As a matter of fact, it will be easy for companies to choose the right influencers when they possess information about the influencers' specific domain and interests (in sports, fashion, makeup, food, or health).

The following sections introduce the different areas that are related to finding influential users on social media. Different social network analysis techniques are employed in many applications interpreting social media data. Moreover, major statistical packages such as Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS) include dedicated sentiment analysis modules used in such studies. For example, the authors in [9] used sentiment analysis for measuring customer's satisfaction and applied in commercial products and business accounts. They used support vector machine (SVM) as a classification algorithm alongside the unigram as a feature extraction method and then measured

the sentiment of Twitter's data. The experimental result indicates that the unigram features extraction method with SVM classification together bring a high score of accuracy that reaches 87% [9].

Authors in [10] proposed another system named Brandwatch which examines and analyzes a huge volume of data on the web. The process begins with the collection of data from many online sources such as Twitter, Facebook, Flickr, YouTube, Vimeo, and discussion forums. Clients can select the site they desire to inspect. Subsequently, the system cleans out data and removes recurring data and ads, detaching factual mentions of brands from uses of the same words, inspecting and analyzing the site to select its data to enable trends to be followed and tracked. Furthermore, Brandwatch displays data in various ways, including on a digital dashboard [10]. Additionally, the system enables users to discover and follow influencers; they can also promote the brand through the social tool, view the data of influencers on their profiles, give ratings to posts, analyze the number of followers per influencer, and gain insight on the performance of their content. It facilitates tracking the influence of the participant and measures the actual value of its activities [11].

The fact that the social media marketing suite uses artificial intelligence (AI) to understand audiences' behavior is also noteworthy. It enables users to measure social media's impact based on business goals and investment in content. It helps users make decisions in real time, comparing, benchmarking, and analyzing all the information and key performance indicators (KPIs) that are important to make decisions, take actions, and increase the numbers. Additionally, it helps marketers make smarter decisions across their digital channels [12].

In some research, NLP techniques were combined with sentiment analysis to create a new ranking algorithm for better user experience on Instagram. For example, author in [3] applied an algorithm to determine which accounts will be ranked high in search results, based not only on the number of likes or followers but mainly on the sentiment analysis of user comments. The author generates personalized recommendations on Instagram based on both the account's characteristics and the sentiment analysis of user feedback about certain services or products. The data is downloaded and extracted from Instagram by using the Instagram Application Programming Interface (API) open-source project (Instagram PHP Scraper). The requested data is extracted by analyzing JSON objects on the web version of Instagram and stored in database and preprocessed in Python. The implementation part is preceded by research on machine learning algorithms and mining techniques as well as on sentiment analysis [3].

The literature shows also an interest in findings influential people in social networks. Authors in [13] proposed an algorithm for finding influential users of web event in social media. They applied PageRank and HITS algorithms to calculate the influence of users on the integration of two networks: the user behavior networks and association network of words. Additionally, authors in [14] used association learning to identify influential users and to detect relationships between users. They compared their results to the results from degree centrality and PageRank centrality. Additionally, authors in [15] applied association rule learning to find influential bloggers using the Apriori algorithm combined with the Oracle data miner. This was also done to find frequent patterns among bloggers holding blog activities together and discover who influences others based on the rules learned from association rule mining.

The discovery of influential bloggers to understand social activities introduced and provided unique opportunities for sales and advertising. The marketing influential value (MIV) model, for instance, evaluates influential strength and identifies influential bloggers in the blogosphere. In [16] three dimensions of blog characteristics were analyzed; network-based, content-based, and activeness-based factors, and then, the artificial neural network (ANN) algorithm was utilized to discover potential bloggers. Based on official evaluations, the experimental results show that the proposed framework outperformed two social network-based methods—out-degree and betweenness centrality algorithms—and two content-based mechanisms—review rating and popular author approaches.

2.1 Sentiment analysis

Sentiment analysis is an important part of this research, and it is an effective area of study in the field of NLP that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text [17], [18]. Many sentiment analysis approaches rely greatly on an underlying sentiment lexicon, which is a list of lexical features that are generally labeled according to their semantic orientation as either positive or negative. Indeed, sentiment in this context is about attitudes, emotions, feelings—it speaks of subjective impression rather than of facts. It aims to determine the attitude expressed by the text's writer or speaker with respect to the topic or the overall contextual polarity of a document [19]. It is being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of people's behaviors. Hence, their beliefs and perceptions of reality, and the choices they make, are largely conditioned on how others see and evaluate the world. For this reason, when people need to make a decision, they often seek out the opinions of others. This is true not only for individuals but also for organizations. With the explosive growth of social media (particularly through reviews, forum discussions, blogs, and social networks) on the web, individuals and organizations are increasingly using public opinions in these media for their decision-making [20].

However, finding and monitoring opinion sites on the web and extracting the information contained in them remain a formidable task because of the proliferation of diverse sites. Each site typically contains a huge volume of opinionated text that is not always easily interpreted in long posts on life. The average human reader will have difficulty identifying relevant sites and accurately

summarizing the information and opinions contained in them. Moreover, it is also known that human analysis of textual information is subject to considerable biases—for example, people often pay greater attention to opinions that are consistent with their own preferences. People also have difficulty—owing to their mental and physical limitations—producing consistent results when the amount of information to be processed is large. Automated opinion mining and summarization systems are, thus, needed, since subjective biases and mental limitations can be overcome with an objective sentiment analysis system [20].

Sentiment analysis is divided into specific subtasks: sentiment context, sentiment level, sentiment subjectivity, sentiment polarity, and sentiment strength [9] [21]. A good example of sentiment analysis algorithm that we plan to use in our project is Valence Aware Dictionary and Sentiment Reasoner (VADER), which is a lexicon and rule-based sentiment analysis tool that is specifically adapted to sentiments expressed in social media and works well on texts from other domains. It is sensitive to strength of emotion and polarity [22].

It is interesting to note that sentiment analysis was examined on data from Twitter. Polarities of tweets were determined by two ways: a group of ten people and VADER sentiment analysis. In total, 527 tweets from ten different companies were examined. At the conclusion of this study, the obtained results were compared to see if there was a significant similarity among methodologies. The results showed that there was no significant difference among human and VADER sentiment. The inherent nature of social media content poses serious challenges to practical applications of sentiment analysis [22].

3 Muatheer framework

The development process of our proposed solution Muatheer started by determining system requirements from all aspects such as data collection, software requirements, hardware requirements, functional requirements, and non-functional requirements. We started with data collection, which is an important aspect of any type of research. In fact, inaccurate data collection can affect the results of a study. The main data collection strategies used were brainstorming**,** literature review, similar tools analysis, interviews, and questionnaires. The proposed solution was analyzed through a set of diagrams showing the structure of the system, explaining the relationships and interactions of the objects, sequencing the steps performed by the system, and analyzing the processes. Figure 1 represents the framework of Muatheer.

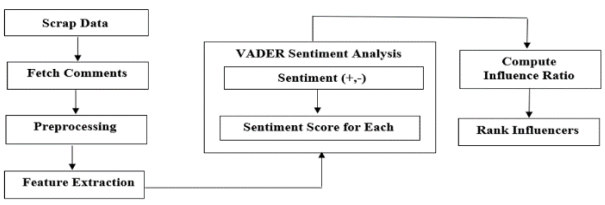

Figure 1: The framework of the proposed solution Muatheer.

In this section, we offer a brief description of the steps that were actually followed to implement Muatheer's website. The following software programs were used, and steps were followed in the implementation:

- XAMPP includes MySQL server and Apache, which we used to store user accounts and analysis reports. The database was managed using the PHP MyAdmin module in XAMPP.
- We created a developer account on Instagram to fetch data using the steps explained in [23]. To implement the Instagram API, we used the Instagram-PHP-Scraper library [24].
- The Muatheer website was implemented using PHP, HTML, CSS, JavaScript, and JSON.
- VADER algorithm was used with a lexicon dictionary from the VADER library [25].

The following paragraphs explain the implementation details of Muatheer.

3.1 Scraping Data Using Instagram API

The Instagram API scraper is the most appropriate for scraping data. The API works as a point of connection between a specific development environment and developers, so that the latter can benefit from the services of this environment without having to build everything from scratch. In general, the goal of the API is to hide the details, perform encapsulation, and highlight procedures to use the code.

- Instagram Basic Display API allows app users to access basic profile information and videos on their Instagram accounts. It can be used to access any type of Instagram account but only provides read access to key data.
- Instagram Graph API allows professional accounts—belonging to companies and creators—to use the app to manage their Instagram presence. It can be used to obtain and publish their media, manage, and respond to comments on their media, specify the media in which these elements have been reported by other Instagram users, and obtain basic metadata and metrics about other businesses and creators on Instagram [24]. To create a profile of Muatheer on the Instagram developer API website, we followed the following steps:
  - Login to Instagram through the following URL and create a new account: www.instagram.com.
  - Enter the developers' Instagram page via this link: developers.instagram.com. Then, get started by clicking on "Get Started" and activating the account through an email.
  - Create the app by clicking on "My Apps" and choose "Create App".
  - Choose "Setting" from the side menu and then choose "Basic".
  - Add a platform from the current interface and add information about Muatheer.
  - Add a product to the application by choosing it from the product list (developer side menu).

- Choose the product and adjust its settings by selecting “Set Up”.
- Click on “Create App” and name the application to the product with the same name as the previous application of the Instagram account. Figure 2 shows the created profile on the Instagram developer API website.

The library called Instagram-Scraper scrapes publicly available data from Instagram posts on profiles, hashtags, and place pages. It also extracts links from photos, comments, and detailed information about the Instagram pages. Additionally, it supports search queries [24]. The REST APIs provide programmatic access to read, search, and perform more functions on Instagram data, such as user profiles and posts. Instagram’s endpoints represent a variety of REST-based web service URLs for accessing much of Instagram’s overall functionality. The REST API identifies Instagram applications and users using OAuth; the responses are in JSON format [26] [27]. Basically, we used the following three GET web services from Instagram:

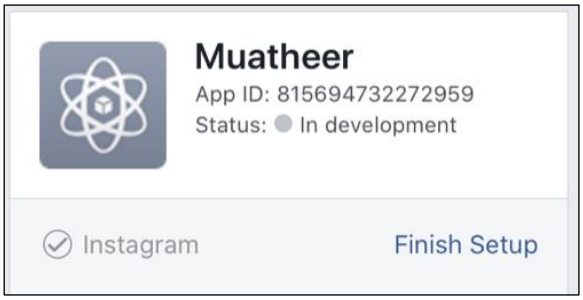

Figure 2: Muatheer profile on the Instagram developer API website.

Comment			
"Comment"	"created_at", "id_str", "text"		
	"entities"	"hashtags"	"text"
User			
"User"	"user_id", "username", "Full_Name", "Biography", "Profile_Pic_Url", "Media_Count", "Followed_By_Count", "Follows_Count"		

Table 1: Comment data model.

Query parameters	**Keyword**	Search term of 500 characters maximum.
	Count	Number of accounts to return.
	result_type	What type of search results to receive and values include collection of top recent posts with user accounts.

Table 2: GET CurrentTopMediasByTagName web service details.

- GET CurrentTopMediasByTagName: This returns a collection of top posts that have comments matching the keyword in the query [27] [28].
- GET Users: This returns profile information about users specified by the user_id parameter in the query [28].
- GET Comments: This returns a collection of comments posted in the user’s post indicated by the user_id parameter in the query [29].

The first GET request was used for keyword analysis to fetch comments containing a search term and results to get just the top media, while the following two GET requests were used for account analysis. The response of these web services includes three JSON models: Comment and User, Comment model, and full properties list. The properties that we used are represented in Table 1.

3.2 Fetch comments

Users can search Instagram for any keywords to find influencers in the same domain, which are related to the keywords. The first step of keyword analysis is to submit a query to Instagram web service GET CurrentTopMediasByTagName. Table 2 explains query parameters.

After specifying the required parameters, a query will be sent to Instagram API to retrieve search results. The request to the Instagram web service is accepted only if the authentication via access tokens is successful. The search query will return the result data in the JSON tree format, which is converted into an array object and then saved in a PHP session for the next step, the preprocessing. Before preprocessing, the raw result array is filtered to only the top recent posts including user accounts. For connection with Instagram API, we used an Instagram-PHP-Scraper library [30] [24].

3.3 Preprocessing

Preprocessing is the step needed to clean the data from noise, standardize it, and convert it to a structured format before extracting distinct features from it. In this work, we used external resources to preprocess the data and provide prior score for some of the most commonly used words.

- **Dictionary**: We used a lexicon of English words list as given in [30] and rated each English word using a compound score. The compound score was computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between –1 (most extreme negative) and +1 (most extreme positive).
- **Stop words removal**: This removes a list of words such as a, an, is, the, with, and, or I, you, among other, which occur in high frequency in a sentence, do not carry any sentiment information, and do not affect the meaning of the sentence. For this reason, they are

better removed. We used the stop words list given in [30].

- **Converting upper case to lower case:** We used case-sensitive analysis and considered the case change of two similar words as different [30].
- **Special character and digit removal:** Digits and special characters do not convey any sentiment have been removed. We used digit removal as given in [30].

After building these lingual and sentiment dictionaries, the preprocessing of comments began. We followed the same preprocessing steps that were suggested by [31] and [30]:

- Remove extra whitespaces.
- Tokenize each comment, i.e., split them into an array of separate words.
- Remove punctuations, all digits, and stop words.

3.4 Feature extraction

A feature is a piece of information that can be used as a characteristic capable of assisting in solving a problem. The quality and quantity of features are very important for the results generated by the selected model. Given below are the most common types of features extracted:

- **Unigram Features:** One word is considered at a time, and it is decided whether the considered word is capable of being a feature.
- **N-gram Features:** They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward.
- **Bi-gram**: This is a sequence of two adjacent elements from a string of tokens, which are typically words. A bi-gram is an n-gram for $n = 2$.

In comments training, consisting of positive and negative comments, we can split each comment into words and add each word to the feature bag. Adding individual (single) words to the feature bag is referred to as the "unigrams" approach. Table 3 shows the unigram approach and the resulted bag of words.

In unigram features, each feature is a single word found in a comment. Each comment will be a combination of each of these feature words. Based on this pattern, a comment is labeled as either positive or negative.

3.5 VADER sentiment analysis

The proposed algorithm utilizes a combination of rule- and lexicon-based approaches that are performed with help of VADER, which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed on social media; it works well on textual data. VADER combines qualitative analysis. It is sensitive to the strength of emotion and polarity. Sentiment analysis using VADER relies on a dictionary, which, depending on emotion intensity, is also known as the sentiment score, and maps lexical features [22] [32]. The reason for choosing VADER as a classification algorithm in this problem is based on the research works and the benchmark presented in [19] which recommends VADER after it showcased its high performance. This study compared VADER and its effectiveness with eleven typical state-of-practice benchmarks including the Linguistic Inquiry and Word Count (LIWC), the Affective Norms for English Words (ANEW), the General Inquirer (GI), SentiWordNet (SWN), and machine learning oriented techniques relying on Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) algorithms [19].

VADER relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text [19].

The score of a text is obtained by aggregating the intensity of each word in the text. For accurate sentiment analysis, the weight is set to each word that could be positive, negative, or neutral. Sentiment score is in the range between –1 and 1. If VADER score > 0.1, it is taken as positive; if it is < –0.1, it is taken as negative; finally, if it is between –0.1 and 0.1, it is taken as neutral. As for the weight, for example, the positive word "happy" was transformed into a sentiment score of 0.52. If an adverb, such as "so", was added to the sentence, the score increased by 0.61 over the score attained for a word such as "happy". On the other hand, a word with a negative connotation, such as "sad, had a sentiment score of –0.48. Thus, the weight of each word indicated the sentiment score more accurately. The compound score calculated the sum of all the lexicon ratings, which have been normalized between –1 (most extreme negative) and +1 (most extreme positive) [31]. An example of applying the VADER sentiment score on Instagram comments is presented in Table 4.

The feature extraction method is applied in both training and testing phases. In training, the comments are

Comment (Positive)	The FOOD; was… delicious and flavorful
Pre-processed	food delicious flavorful
Unigram Feature Bag	{ food, delicious, flavorful }
Comment (Negative)	You are a… liar and CHEATER
Pre-processed	lair cheater
Unigram Feature Bag	{liar, cheater }

Table 3: Unigram approach—bag of words.

Text of Instagram's comment	**Compound score**	**VADER polarity**
Free Amazon gift cards	0.967	Positive
Lying and cheating are bad characteristics	0.8889-	Negative

Table 4: Applying VADER sentiment score on Instagram comments.

labeled with "+1" for positive and "–1" for negative. The classifier will use labeled comments to learn from them and builds its learning model. In testing, each new unlabeled comment will be compared to the bag of words generated from labeled comments. The classifier based on VADER algorithm and supported by VADER library will be implemented as given in [33].

3.6 Keyword frequency

Keyword frequency refers to counting the number of times the searched keywords are repeated in Instagram comments. Therefore, if the frequency of keywords in influencer comments is high compared to all posts initiated by the account, this gives an indicator that the account is more interested and specialized in the specified domain related to those searched keywords. Then, a normalized number can be computed that determines whether the influencer is more interested and active in the same domain of the searched keyword or not.

3.7 Influence ratio

Finally, the influence ratio is calculated to discover real influencers in the searched domain using specific keywords. From the questionnaire and interviews that were conducted with users who were interested and attracted to certain Instagram influencers, we concluded that most of users trusted influencers when they realized that they are experts on a certain topic or domain; they display organized content on their account, have many followers, high number of likes, and high percentage positive comments. Equation 1 calculates the ratio of influence by dividing 100% by a set of parameters according to their importance in selecting the influencers. The weight of each parameter was determined based on the importance of the parameter and information gathered previously from people; the weights are as follows:

- 30% frequency of keyword
- 30% positivity
- 20% number of likes
- 20% number of followers

$$\textbf{Influence Ratio} = (0.3 \times F_{Keywords} + 0.3 \times Positivity + 0.2 \times N_{Like} + 0.2 \times N_{Followers}) \quad (1)$$

All parameters were normalized and then applied to the previous equation. Subsequently, the high frequency of the searched keywords was compared to the total number of words in the influencer's posts. It is an important factor in the measurement of influence, as it gives a clear indication that the influencer is more interested in this domain. The second important factor is the positivity of the comments, which gives an indicator of the influencer's good impact on their followers and the trust they garner. Finally, the number of followers who like their posts compared with their total number of followers also indicates the interest of the followers on the content posted by the influencer. Finally, the high percentage of followers who are involved in the extracted comments compared with the total of followers also support the trust on that influencer.

Therefore, the influence ratio can help determine the actual influencer, the higher influence ratio, and the greater chance that the user will be chosen as an influencer in the search domain. The search result will produce ranked Instagram accounts using their influence ratio with the help of some information such as posts, accounts information, number of likes, number of comments, frequency of keyword, number of followers\following, and positivity as shown in Figure 3.

4 Experimental results

In this section, the experiment was applied to the search results of the keyword "Coffee", which yielded some accounts; the lowest influence ratio appears in Figure 3. Such a result with a low influence ratio demonstrates that the shown account has low influence in the domain of "coffee". Table 5 presents selected raw of comments of this account to explain the details of the selection process.

Id	**********
Username	coffee**********
Full name	Coffee**********
Posts	180
Followers	1690
Follows	18
Positivity	1.2
Frequency of keywords	23
Influence ratio	3.95 %

Figure 3: The testing result for low influence account.

1	#coffeetime #coffeelover #coffeeaddict #coffeeshop #coffeebreak #coffeegram #coffeelovers #coffeelove #coffeeholic #coffeelife #coffeemug #coffeeoftheday #coffeeart #coffeecup #coffeetable #coffeeporn #coffeehouse #coffeebean #coffeesesh #coffeeshots #coffeebeans #coffeeplease
2	Art! ❤☕
3	Thanks dude @coffee.xxxxx
4	😍
5	Wow
6	This has such a cool sci-fi look to it, I love it

Table 5: Account result.

- From these comments analyzed by the VADER sentiment analysis, let us take one: "Thanks dude @coffee.xxxxx".
  - The word "Thanks" in the lexicon equals 1.9; the remaining words are neutral.
  - Here, we will calculate the score of each positive, negative, and neutral.
  - Neutral = (1+1) / (1+1+2.9) ~ 0.408; positive = 0.592; negative = 0
- The frequency of the keyword "coffee" in this comment is 1 and in the first comment is 22.
- Calculate the compound score as shown in Figure 4 by aggregating the words' scores.

Figure 5 demonstrates that the following account, which is a famous account in the domain of "coffee", appears in the top of search result. It shows that the influence ratio is high (82.58) in the selected domain.

Using a combination of qualitative and quantitative methods, a gold standard list of lexical features (along with their associated sentiment intensity measures) should be first constructed and empirically validated. The lexical features should be specifically attuned to sentiment in microblog-like contexts. Then, these lexical features are combined with the consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity

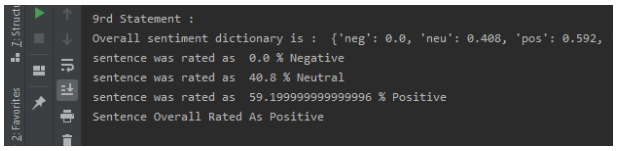

Figure 4: VADER sentiment coding result.

Id	**********
Username	coffee******®
Full name	Coffee**********
Posts	1825
Followers	294K
Follows	2
Positivity	91.0
Frequency of keywords	1882
Influence ratio	82.58 %

Figure 5: The testing result for a high influence account.

Tool/Technique	F1 Accuracy
VADER	0.96
LIWC	0.63
ANEW	0.60
GI	0.69
SWN	0.67
NB	0.84
ME	0.83
SVM	0.83

Table 6: F1 accuracy of VADER and eight other highly regarded sentiment analysis tools/techniques on a corpus of over 4k tweets.

Finally, a set of accuracy tests was applied to examine the effectiveness of the selected sentiment analysis algorithm and the generated system as a whole. The VADER algorithm's effectiveness is compared based on the study presented in [19], as shown in Tables 6. The typical state-of-practice benchmarks include LIWC, ANEW, GI, SWN, and machine learning-oriented techniques rely on NB, ME, and SVM algorithms.

By comparing the above results in Table 6, we can observe that the VADER algorithm shows better performance than other methods, as it accomplishes F1 score equal to 0.96. Additionally, the accuracy was tested by applying a set of metrics such as Precision, Recall, Accuracy, and Error Rate.

Precision (P) is defined as the ratio of relevant items selected to the number of items selected [34], as shown in Eq. (2). In the proposed algorithms, it tells us about when it predicts positive and how often it is correct. N_{rs} refers to the number of really positive and predicted positive, and N_s refers to the number of all selected comments. It can be applied to Negative and Neutral comments in the same way shown in Table 5.

$$P = \frac{N_{rs}}{N_s} \tag{2}$$

Recall (R) is defined as the ratio of relevant items selected to the total number of relevant items available as shown in Eq. (3) [34]. In our algorithms, it gives us an idea of when it is positive, how often it predicts. Further, N_{rs} refers to the number of really positive and predicted positive, and N_r refers to the number of all positive comments. It can be applied to Negative and Neutral comments in the same way shown in Table 5.

$$R = \frac{N_{rs}}{N_r} \tag{3}$$

Accuracy is defined as the ratio of the correct predicted items to the total number of items available as shown in Eq. (4) [34]. In our algorithms, it gives us an idea about the total corrected predicted items including Positive, Negative, and Neutral. $N_{rs(Positive,\ Negative,\ Neutral)}$ refers to the number of corrected predicted items including Positive, Negative, and Neutral, and N refers to the number of all available items.

$$Accuracy = \frac{N_{rs(Positive,Negative,Neutral)}}{N} \tag{4}$$

Moreover, we can investigate the Error Rate (Err.), which represents the ratio of the items that are predicted incorrectly to the total number of items available as shown in Eq. (5).

$$Err. = 1 - \frac{N_{rs(Positive,Negative,Neutral)}}{N} \tag{5}$$

4.1 Training/Testing Dataset Collection

Typically, when you separate a dataset into a training set and testing set, most of the data is used for training and a smaller portion of the data is used for testing. There are 4000 publicly available datasets from Instagram with

sentiment analysis. The VADER sentiment analysis applies into 174 comments. It computes the valence score of the sentence. If the score is positive, VADER adds a certain empirically obtained quantity for every exclamation point (0.292) and question mark (0.18). If the score is negative, VADER subtracts. Then, the above equations were applied to the results, and the calculated metrics are shown in Table 7.

Table 7 shows that the overall accuracy of the VADER sentiment analysis is 85.05%, which is high, and the error rate is 14.94%, which is low. Also, the result is depicted in Figure 6.

5 Conclusion

Social media accelerates the creation and sharing of information via virtual communities and networks. Many individuals and organizations fall into the problem of choosing the right influential user who can present advertisements or any public opinion. This research examined influencers on Instagram and studied important criteria that affects the influencing process by applying sentiments analysis with a set of criteria encapsulated in the influence ratio. The criteria of discriminating users include the number of likes, number of followers, frequency of keywords and posts' positivity. Muatheer used the sentiment analysis algorithm combined with the unigram as a feature extraction method. This facilitated the function of finding the influential users on Instagram, and the comparison of influencers on social media smoothly helps public organizations and business owners choose influencers easily and clearly. The evaluation of the proposed solution was done by applying the experiment using 4000 comments on Instagram as the training dataset and 174 comments as a test. The accuracy was 85.05%, and the error rate was 14.94%. Moreover, Precision and Recall were applied to measure system achievement, which yielded high results.

Matrix Metrics						**Error Rate**	**Accuracy**
N= 174		Predicted Class				14.94%	85.05%
		Positive	Negative	Neutral	Precision		
Actual Class	Positive	60	6	4	85.71%		
	Negative	6	70	2	89.74%		
	Neutral	3	5	18	69.23%		
	Recall	86.95%	86.41%	75%			

Table 7: The metrics matrix of the experimental result.

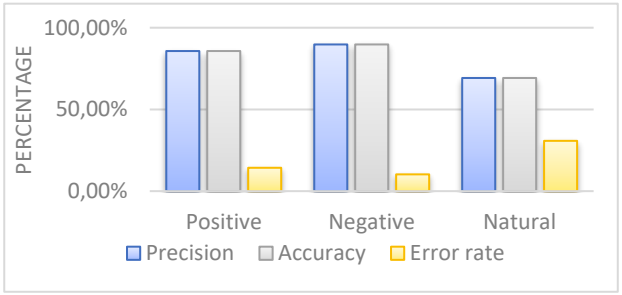

Figure 6: Metrics for different types of comments.

References

[1] C.-W. C. Ki, L. M. Cuevas, S. M. Chong and H. Lim, "Influencer marketing: Social media influencers as human brands attaching to followers and yielding positive marketing results by fulfilling needs," Journal of Retailing and Consumer Services, vol. 55, p. 102133, 2020. https://doi.org/10.1016/j.jretconser.2020.102133

[2] P. J. Carrington, Models and Methods in Social Network Analysis, United States of America: Cambridge University, 2005.

[3] D. Salomon, "Moving on from Facebook: Using Instagram to connect with undergraduates and engage in teaching and learning," College & Research Libraries News 74, no. 8 (2013): 408-412., vol. 74, no. 8, pp. 408-412, 2013. https://doi.org/10.5860/crln.74.8.8991

[4] S. Kaur and R. Mohana, "Prediction of sentiment from macaronic reviews.," Informatica , vol. 42, no. 1, pp. 127-137, 2018.

[5] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson and T. Seymour, "The history of social media and its impact on business," Journal of Applied Management and entrepreneurship , vol. 16, no. 3, pp. 79-91, 2011.

[6] C. Holsapple, S.-H. Hsiao and R. Pakath, "Business Social Media Analytics: Definition, Benefits and Challenges," in Twentieth Americas Conference on Information Systems, Savannah, GA, USA, 2014.

[7] W. Etaiwi, D. Suleiman and A. Awajan, "Deep Learning Based Techniques for Sentiment Analysis: A Survey," Informatica, 45(7)., vol. 45, no. 7, pp. 89-95, 2021. https://doi.org/10.31449/inf.v45i7.3674

[8] M. Vedanayaki, "A study of data mining and social network analysis," Indian Journal of Science and Technology, vol. 7, no. S7, pp. 185-187, 2014.

[9] S. Al-Otaibi, A. Alnassar, A. Alshahrani, A. Al-Mubarak, S. Albugami, N. Almutiri and A. Albugami, "Customer Satisfaction Measurement using Sentiment Analysis," International Journal of Advanced Computer Science and Applications, vol. 9, no. 2, pp. 106-117, 2018. https://doi.org/10.14569/ijacsa.2018.090216

[10] R. Stair and G. Reynolds, Principles of Information Systems, USA: Cengage Learning, 2020.

[11] M. Tremelling, "5 Things You Didn't Know You Could Do with Pixlee," [Online]. Available: https://www.pixlee.com/blog/5-things-you-didnt-know-you-could-do-with-pixlee/. [Accessed 5 10 2019].

[12] N. Wirth, Hello marketing, what can artificial intelligence help you with?, International Journal of Market Research, vol. 60, no. 5, pp. 435-438, 2018. https://doi.org/10.1177/1470785318776841

[13] Q. Ma, X. Luo and H. Zhuge, "Finding influential users of web event in socialmedia," Concurrency Computat Pract Exper, vol. 31, 2019. https://doi.org/10.1002/cpe.5029.

[14] F. Erlandsson, P. Bródka, A. Borg and H. Johnson, "Finding Influential Users in Social Media Using Association Rule Learning," Entropy , vol. 18, 2016. https://doi.org/10.3390/e18050164

[15] Shazad, B., Khan, H.U., Farooq, M., Mahmood, A., Mehmood, I., Rho, S. and Nam, Y., "Finding Temporal Influential Users in Social Media Using Association Rule Learning," Intelligent Automation And Soft Computing, vol. 26 , no. 1 , pp. 87-98, 2020. https://doi.org/10.31209/2019.100000130

[16] Li, Y.M., Lai, C.Y. and Chen, C.W., Discovering influencers for marketing in the blogosphere, Information Sciences, vol. 181 , no. 23, pp. 5143-5157, 2011. https://doi.org/10.1016/j.ins.2011.07.023

[17] S. Chawla and M. Mehrotra, "Impact of emotions in social media content diffusion," Informatica, vol. 45, no. 6, pp. 11-28, 2021.

[18] J. Li, Y. Wang and J. Wang, "An Analysis of Emotional Tendency Under the Network Public Opinion: Deep Learning," Informatica , vol. 45, no. 1, p. 149–156, 2021. https://doi.org/10.31449/inf.v45i1.3402

[19] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, vol. 8, no. 1, pp. 216- 225, 2014.

[20] B. Liu, "Sentiment Analysis and Opinion Mining," in Web Data Mining pp , Illinois - USA, University of Illinois at Chicago, 2011, pp. 459-526. https://doi.org/10.1007/978-3-642-19460-3_11

[21] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," Ai & Society, vol. 30, no. 1, pp. 89-116, 2015. https://doi.org/10.1007/s00146-014-0549-4

[22] A. Kirlić and Z. Orhan, "Measuring human and Vader performance on sentiment analysis," Invention Journal of Research Technology in Engineering & Management (IJRTEM), vol. 1, no. 12, pp. 42-46, 2017.

[23] I. API, "Instagram Developer," Instagram, [Online]. Available : https://www.instagram.com/developer/. [Accessed 9 Jan 2020].

[24] Riyam, "Instagram-php-scraper," Github, [Online]. Available: https://github.com/postaddictme/instagram-php-scraper. [Accessed 31 feb 2020].

[25] Abusby, "Github," [Online]. Available: https://github.com/abusby/php-vadersentiment. [Accessed 25 Feb 2020].

[26] Instagram, "Libraries," [Online]. Available: https://www.instagram.com/developer/libraries. [Accessed 3 March 2020].

[27] Repat, "instagram-php-scraperTag," [Online]. Available: https://github.com/postaddictme/instagram-php-scraper/blob/master/examples/getCurrentTopMediasByTagName.php.

[28] Instagram, "Endpoints," [Online]. Available: https://www.instagram.com/developer/endpoints/users. [Accessed 14 March 2020].

[29] Instagram, "Comment," [Online]. Available: https://www.instagram.com/developer/endpoints/comments. [Accessed 14 March 2020].

[30] Abubsy, "Php-vadersentiment," [Online]. Available: https://github.com/abusby/php-vadersentiment. [Accessed 20 March 2020].

[31] Cjhutto, "VaderSentiment," [Online]. Available: https://github.com/cjhutto/vaderSentiment. [Accessed 14 March 2020].

[32] C. H. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Association for the Advancement of Artificial Intelligence, Atlanta, 2014.

[33] Abubsy, "Aadersentiment," Feb 2020. [Online]. Available : https://github.com/abusby/php-vadersentiment.

[34] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in the 23rd international conference on Machine learning, Pittsburgh Pennsylvania USA, 2006. https://doi.org/10.1145/1143844.1143874