Privacy in the Age of Social Media: An Investigation into the Threats Posed by Data Scraping

Changmyeong Oh
Computer science and Engineering
Korea University
Seoul, Republic of Korea
kemarine@korea.ac.kr

***Abstract*—*Social networking services have rapidly established themselves as essential components of modern society. Due to their web-based operations, these services are inevitably exposed to web scraping. However, existing research on web scraping primarily focuses on ethical aspects, often overlooking the specific impacts of scraping and analyzing data from social networking services on individual privacy. In this study, we investigate potential data privacy threat hypotheses that may emerge from the analysis of data on social networking services. We test these hypotheses by scraping and analyzing posts from a specific user (a celebrity). Our findings suggest that extracting and analyzing frequently used words from data scraped from social networking services can potentially enhance the accuracy of re-identification, posing a threat to individual privacy. This study concludes by raising awareness of the potential data privacy threats posed by scraping social networking services. We highlight additional possibilities for threats and call for further research to develop strategies for securing safety from such risks.*

***Keywords*—*Web scraping, Data privacy, Privacy threats, Re-identification, Data analysis, Social Networking Service*

I. Introduction

Social Networking Services (SNS) are becoming an essential part of life for modern humans. Twitter, a notable SNS, announced in the first quarter of 2023 that it had 298.9 million Monthly Active Users (MAU) (Meta Platforms, 2023). The unrestricted nature of content creation on such widely used SNS platforms allows for a diverse range of posts to be uploaded to their servers, spanning from objective facts to personal daily routines. Owing to the platform characteristics of SNS, which primarily operate through web and app interfaces, these sites are vulnerable to scraping technology that can extract posts based on specific categories. Through scraping and subsequent application of various data analysis techniques, it's possible to discern the characteristics of these posts. In this context, we identified the potential for personal data leakage through the scraping of specific users' posts. While existing studies have suggested the possibility of ethical issues arising from scraping, they do not clarify how scraping data from SNS can specifically be used for extracting personal information.

In this work, we scrape and analyze tweets from celebrity accounts on Twitter, a popular SNS, to present and verify potential scenarios of how such data could be used for personal data leakage or re-identification. This research emphasizes the need for scraping prevention technology and raises awareness among SNS users about the risk of personal data leakage, thereby contributing to the protection of personal data privacy in the SNS environment. Moreover, it underscores the need for additional research into how various unused data analysis techniques in this study could be employed for potential breaches of personal information.

II. Materials and Methods

A. Materials

We selected three users(celebrities) and collected up to 5000 of their tweets from the most recent ones using the snscrape library in Python. The reason for choosing celebrities as samples is because they offer high accessibility to additional data needed in the process of verifying hypotheses about re-identification and privacy leakage scenarios. The identifiers used to collect tweets were as follows: the date of creation, and the content. The date was in the format of YYYY-MM-DD.

We used there pandas library and re library in Python to preprocess the data by removing special characters, garbled languages during collection, and mentions that point to specific users. Additionally, to facilitate data analysis, we utilized the vader_lexicon in Python's nltk library to conduct sentiment intensity analysis. The analyzed file consists of three identifiers: datetime, text, and sentiment. Sentiment scores range from -1 to 1, with higher values corresponding to more positive content. Moreover, for word frequency analysis, we combined all the preprocessed tweets, split them into words based on spaces, and saved the frequency of these words. This file consists of two identifiers: word and frequency.

B. Methods

We used the source code for the data collection and preprocessing processes. Simplified code follows:

```
# Import necessary libraries
import pandas as pd
import re
import snscrape.modules.twitter as sntwitter
from collections import Counter
from nltk.sentiment.vader import
SentimentIntensityAnalyzer
import nltk

# Define key functions
def scrape_twitter(user_id, max_tweets):
    # This function uses the snscrape library to collect tweets
```

```
from a specific user.

def process_dates(df):
    # This function processes date data in the DataFrame.

def count_dates(df):
    # This function calculates the frequency of each 'year-
month'.

def preprocess_text(text):
    # This function removes special characters and Korean
text, and applies sentiment analysis.

def count_words(df):
    # This function calculates word frequency.

# Download necessary resources
nltk.download('vader_lexicon')

# Use functions to collect and process data
df = scrape_twitter('userID', 5000)
df = process_dates(df)
df_date_count = count_dates(df)
df['Text'] = df['Text'].apply(preprocess_text)
df_word_counts = count_words(df)

# Save the results to CSV files
df.to_csv('sentiment.csv', index=False)
df_date_count.to_csv('date_count.csv', index=False)
df_word_counts.to_csv('word_counts.csv', index=False)
```

We collected tweets from a specific user using the snscrape library (see scrape_twitter function in Simplified Code). The tweet data was processed to simplify the dates (see process_dates function in Simplified Code), and we calculated the frequency of each 'year-month' (see count_dates function in Simplified Code). Next, we removed special characters and Korean text from the tweets and performed sentiment analysis (see preprocess_text function in Simplified Code). We also calculated the frequency of each word in the tweets (see count_words function in Simplified Code). The processed data, date counts, and word frequencies were saved to CSV files for further analysis. We analyzed the processed data from three perspectives and set up scenarios for personal information breaches and re-identification attacks:

1) **Date of tweets**: We disregarded the day of the data and organized the number of tweets by month to identify periods when many or few tweets were posted. We proposed a scenario that if there is information about a particular person's schedule, it can be used in a linkage attack. We verified this by investigating events related to the celebrity during periods of fewer tweets.
2) **Frequency of used words**: We excluded words like articles, pronouns, common verbs(be, have, etc.), and auxiliary verbs and sorted the most frequently used words to identify patterns. We hypothesized that this analysis could reveal information about the target's interests, hobbies, and job. Especially if it reveals information about the job, we proposed a scenario where it can be used in re-identification. We verified this by investigating the relationship between the celebrity's job and the high-frequency words.
3) **Sentiment analysis**: We used nltk's vader to quantify the positivity or negativity of a tweet. We gauged the individual's emotional state through the average of these numbers, calculated the ratio of non-zero tweets and positive and negative tweets, and evaluated the reliability of the average value through comparison with the overall average. We also proposed a scenario where this information could be used in a linkage attack when linked to disease data, based on the hypothesis that it could be used as a tool to estimate the probability of depression. To test this hypothesis, we analyzed data from a celebrity who mentioned they were suffering from depression and a celebrity who did not.

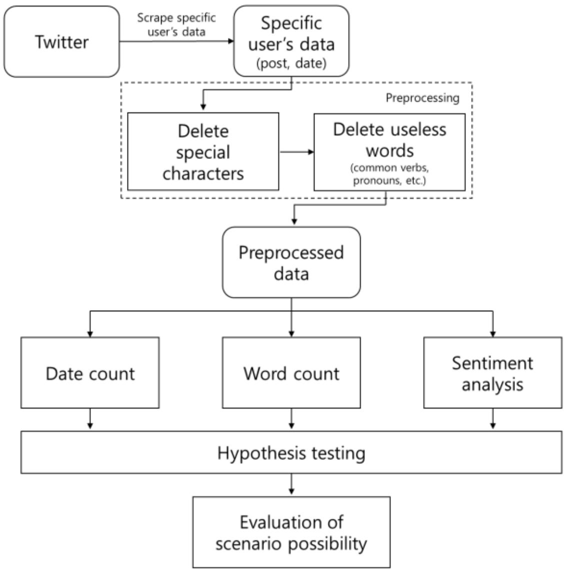
Twitter
Scrape specific user's data
Specific user's data (post, date)
Preprocessing
Delete special characters
Delete useless words (common verbs, pronouns, etc.)
Preprocessed data
Date count
Word count
Sentiment analysis
Hypothesis testing
Evaluation of scenario possibility

Figure 1. Flowchart of the overall research process

III. RESULTS

Before discussing the results, We will refer to the three users selected during the data collection phase as User 1, User 2, and User 3. They are all recognizable actors (with a minimum of 100,000 Twitter followers), each possessing a Twitter account. User 1, selected to prove hypothesis 3, has publicly acknowledged suffering from depression.

A. Date of tweets

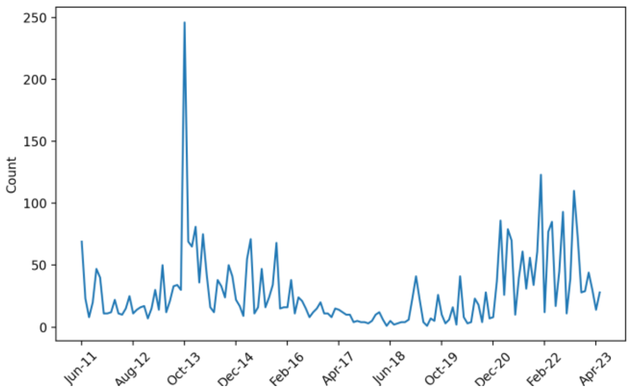

Figure 2. Monthly tweet frequency of User 1

In the case of the first user, excluding the initial period of starting to use Twitter, we can see that the number of tweets noticeably decreased between February 2016 and October 2018. However, there were no notable events in the news during that period, other than reports about the works in which the actor was appearing, so we could not derive meaningful results. We can also see that a considerable number of tweets were written in October 2013, but during that period too, there was only news about the works the actor was involved in. We could not find any strong correlation between the frequency of tweets by the first user and their schedule.

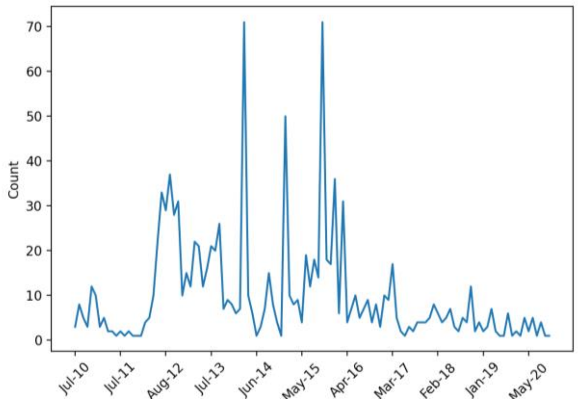

Figure 3. Monthly tweet frequency of User 2

In the case of second user, although there were fluctuations between 2013 and 2016, they were generally active in tweeting, but the number decreased sharply after 2017. When we looked at the news during the active writing period, we mainly found articles about the campaign the user participated in. This will be covered in the word frequency section later, but we could only derive that they were engrossed in the campaign. When we looked at the news after 2017, when the tweets were less frequent, we mainly found a 2023 interview about the reason the actor had taken a 5-year hiatus. This suggests that they also stopped their main activities during the period when they were less active on Twitter. Therefore, we only found cases that could support that there is no correlation between the schedule and the frequency of tweeting, or rather, contrary to the hypothesis, that there is an inverse relationship between specific schedules and tweeting, i.e., tweets also become less frequent when there is no schedule.

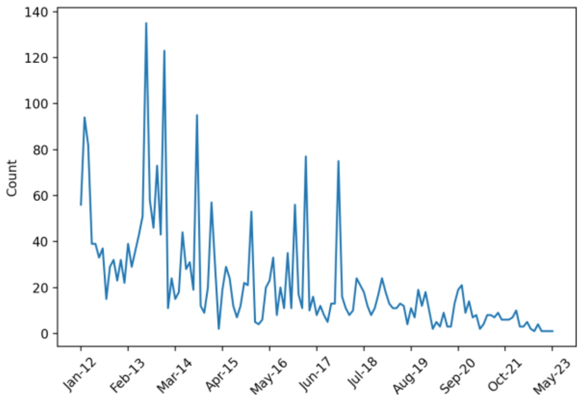

Figure 4. Monthly tweet frequency of User 3

In the case of third user, the number of tweets is large with great fluctuations from the beginning to 2017, but it remains decreased after that. Based on this, we looked at the news before and after 2017, but in both cases, there were only news introducing the works in which the actor appeared or about the movie they would appear in. From this, we could confirm that there is no relationship between the frequency of tweets and the schedule.

B. Frequency of used words

Word	Count
Supernatural(supernatural)	501
Walker(walker)	495
SPNFamily(spnfamily)	324
WalkerFamily	320
AKF(akf)	177
time	159
episode	116
WalkerIndependence	81
Sam	78
scene	72

Figure 5. Table of the top 10 most frequently used words by User 1

Looking at the top 10 frequency words of the first user, most of them are the title of the work he starred in and related words. Other words used are also drama-related words such as episode, scene. From this, we can speculate that he is a fan or related to a specific drama series, and if we investigate the commonalities between the dramas, we can derive the possibility that he is a fan of a particular actor or that actor himself. Also, by paying attention to the word AFK, which has nothing to do with drama, and investigating the commonalities between other words, it seems that we can add credibility to such a possibility from the fact that the campaign launched by the actor.

Word	Count
women	44
HeForShe(heforshe)	41
new	34
amazing	32

UK	23
Noah	21
gender, girls, new	19
Emma, interview, equality, Perks	18
world	17
book	15
excited	14

Figure 6. Table of the top 10 most frequently used words by User 2

In the case of the second user, we can infer from words such as HeForShe, women, and gender that he is enthusiastic about gender equality. Also, the title of the work he appeared in or parts of it are used quite frequently, so it is possible to speculate that he is interested in the actor or the actor himself in the process of tracing the commonalities of those words. Also, associating the work with the HeForShe campaign seems to reinforce such a belief.

Word	Count
Congrats	90
good	88
man	77
awesome	74
show	61
Great	57
fantastic	45
cast	42
crew	41
amazing	39

Figure 7. Table of the top 10 most frequently used words by User 3

In the case of the third user, it was difficult to derive meaningful speculation only from the top 10 words of frequency because positive exclamations and some common adjectives were not excluded. It is possible to speculate that he is interested in a particular show or a crew member that leads such a show through show, cast, crew, etc. in the mid and lower ranks, but it is difficult to get more information. We need to look at less frequently used words to get information related to the actor.

C. Sentiment analysis

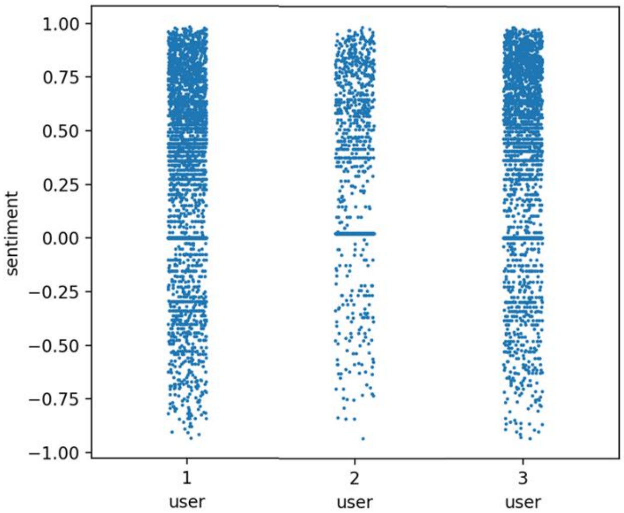

Figure 8. Graph of sentiment analysis compound value distribution for each user's tweets

User	1	2	3
Average of Compound	0.240	0.283	0.329
The number of Positive	2079	625	1824
The number of Negative	569	121	394
Total (include neutral)	3983	1152	2868
Reliability	0.78	0.83	0.82

Figure 9. Table of average sentiment analysis compound, count of positive tweets, count of negative tweets, total tweet count, and reliability for each user

As can be seen from the distribution in Figure 8 and the 'average of compound' item in Figure 9, although there were numerical differences, most posts from all three users were determined to be positive. In calculating the reliability, the average of compound values is considered. If the average of compound values is positive, the number of positive values is divided by the total count, excluding neutral values, to derive the reliability measure. Conversely, if the average of compound values is negative, the number of negative values is divided by the total count, excluding neutral values, to calculate reliability. This approach enables a balanced assessment of reliability, taking into account both the overall trend indicated by the average compound value and the distribution of positive and negative values. The reliability measure computed in this way exhibited a high score, nearing 0.8, across all three users.

IV. Discussion

A. Interpretation of Hypotheses Results

1) Date of tweets

Contrary to the hypothesis, the frequency of tweets by the three celebrities, especially actors, had no relevance to their schedules. For the second user, unlike other users, there was a unique point that the time of active activity was the time of participating in the campaign, but this is still unrelated to the hypothesis.

2) Frequency of used words

Looking at the words with high frequency can help to find items that are of interest to the user or directly related (in the case of an actor, the work in which the actor appears). Except for the third user, finding commonalities among frequently used words has helped to increase the accuracy and credibility of speculation. However, as seen in the case of the third user, there are cases where looking at only high-frequency words does not provide any information. This point was highlighted in this case because common adjectives were not removed in the filtering process. From this, We can confirm that deciding which types of words (such as common noun, common adjectives, or domain-specific jargon) to include or exclude during the data preprocessing stage can significantly affect the acquisition of meaningful information in word frequency analysis.

As seen in the case of User 1 and User 2, if commonalities are found among the words, we can consider a scenario where this information is used as part of the evidence for re-identification. Through the process of matching the common points of these words with a specific person's interest data or job data, we can obtain clues for re-identification or improve its accuracy.

Particularly in the case of User 2, the analysis of news articles from specific time periods revealed engagement in a campaign. The significant frequency of this specific campaign (HeForShe) in the user's tweets suggests that such data, in conjunction with an individual's schedule data, could potentially enhance the precision of re-identification processes.

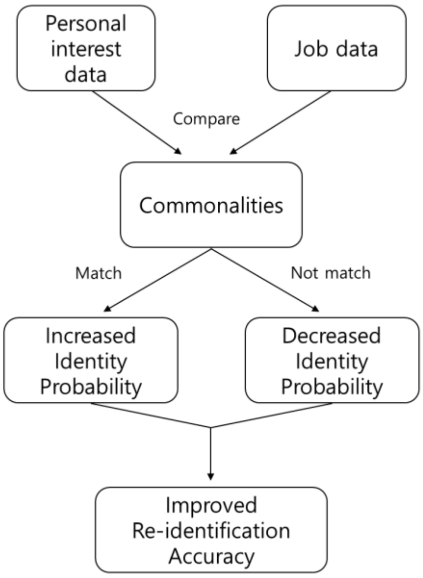
Personal interest data
Job data
Compare
Commonalities
Match
Not match
Increased Identity Probability
Decreased Identity Probability
Improved Re-identification Accuracy

Figure 10. Flowchart of the impact of comparing job data and personal interest data with word commonalities on re-identification accuracy

3) Sentiment anlysis

This finding contradicts the hypothesis, confirming that diseases such as depression do not have an absolute influence on sentiment analysis results. Consequently, sentiment analysis results from the Vader model seem unlikely to be useful for re-identification.

B. Limitations and Constraints of the Study

1) Characteristics of the Sample

In an attempt to recreate possible scenarios for re-identification, we used the SNS accounts of well-known figures, specifically actors, due to the accessibility of related data. However, maintaining a public image is crucial for these individuals, as the social media presence of celebrities plays a significant role in their image creation (Bhatti, 2015). This may have significantly influenced the discrepancy between our hypotheses set based on general users who remain anonymous and the actual results obtained. In particular, it is likely that the management of external image influenced the majority of posts to be measured as positive in sentiment analysis despite the user being openly diagnosed with depression.

2) Incomplete Data Collection

The Python library 'snscrape' is renowned for its rapid processing speed and unlimited data collection capacity. However, some issues arose during the data scraping process.

Firstly, some erroneous characters were found within the text. Even though posts written in English were scraped, characters from other languages (particularly Korean) were mixed in. Although preprocessing was conducted, the data lost during the scraping process was not recovered.

Secondly, not all posts were scraped. The number of posts scraped was lower than the total posts available on Twitter, and the number of posts with more than 50% of the data preserved after scraping was even fewer. This means that many posts were partially altered, lost, or entirely omitted during the scraping process. This likely introduced considerable error, especially in the analysis of User 2's data, who had notably fewer posts, particularly in measuring word frequency.

3) Lack of Proficiency in Data Preprocessing and Analysis Techniques

We used Python to preprocess the data, analyze it, and then create graphs. However, our lack of knowledge in each area posed several challenges. In data preprocessing, we failed to clearly distinguish between areas to be removed and preserved, which led to difficulties in deriving meaningful results, especially in the frequency analysis of words for users like User 3.

Also, due to the absence of a concrete mechanism to deal with error characters from incomplete scraping, we couldn't properly handle portions with apostrophes like "couldn't", links to other sites, or parts where error characters were inserted in the middle of a word. This was reflected in the word count data set, where meaningless words like "t", "co" were counted. This likely affected the inevitable error occurrence in the execution of data analysis. Furthermore, although the VADER model was used for sentiment analysis, the aforementioned issues with data preprocessing and the absence of a delicate data analysis algorithm likely led to additional errors.

V. Conclusion

A. What we learned from this research

In this study, we examined potential issues related to data privacy and the associated attack scenarios that arise from data scraping on social media platforms. This approach sets our work apart from prior studies, which primarily focused on the ethical implications of data scraping. Our particular interest was in understanding the potential threats to data privacy resulting from scraping and subsequent analysis of data from social media. We postulated three hypotheses and related scenarios to illustrate these threats. These encompassed the correlation between the volume of tweets over time and schedule data, the relationship between frequently used words and the user's occupation or interests, and the association between sentiment analysis outcomes and health-related data.

To test each hypothesis, we scraped data from three celebrities on Twitter, a widely recognized social media platform. We preprocessed this data and conducted word frequency analysis, examined the number of tweets by date, and performed sentiment analysis. Our results revealed no consistent relationships between the number of tweets by date and schedule or between sentiment analysis and health data, leading us to reject these hypotheses. The only significant correlation we found was between frequently used words and occupation or interests. Based on this finding, we suggested that the combination of these three data types could potentially enhance the accuracy of re-identification.

We hope that the outcomes of our research will not only propose potential for re-identification but will also contribute to efforts to protect data privacy. Especially since the ethics of scraping data from social media has been a subject of controversy, there is a need for discussions aimed at balancing the benefits and vulnerabilities in terms of data privacy. Through this research, we hope to raise awareness of the potential threat that scraping and data analysis can pose to data privacy and to promote a sense of caution about the use of social media and scraping technologies. We also hope that this potential will be comprehensively investigated, leading to the development of alternative solutions for protecting data privacy.

B. Recommendations for Future Research

In light of the limitations outlined earlier, we propose two potential avenues for future research. Although many of the initial hypotheses set forth in this study were rejected due to inappropriate sample selection and subpar data preprocessing and analysis, we believe that a thorough data collection, preprocessing, and analysis process could indeed expose critical factors that could compromise data privacy.

1) Complementary Research on the Three Hypotheses of This Paper

Future research could focus on the general population using anonymous social media accounts. By fully collecting their social media data and analyzing it similar to this paper in terms of word frequency, frequency by date, and sentiment analysis, we could potentially illuminate critical points concerning data privacy.

In this context, it would be necessary to create the word frequency data not merely by excluding articles and common verbs but by formulating a complete dataset and assigning weights according to the significance of each data point. Moreover, this process would also need to identify how particular words serve as major interests during specific periods by categorizing the data based on date. This approach should avoid overly short periods, which can complicate the data, and instead work with month or quarter intervals. The same applies to sentiment analysis, which would benefit from the use of different models that can analyze a broader range of emotions for a more nuanced investigation. In combination with personal data such as schedule or accident-related data, job data, disease data, this approach could offer a precise evaluation of the possibility of re-identification.

This paper concluded the possibility of re-identification merely by suggesting potential scenarios, which left room for more rigorous evidence. Thus, future research could also investigate ways to represent this possibility mathematically.

2) Additional Research on the Use of Social Media by Celebrities

While the hypotheses concerning average people using anonymous accounts did not align perfectly with the analysis of celebrity accounts, this discrepancy could lead to further research into the unique characteristics of celebrity samples.

In this work, we found that although celebrities did not express depressive emotions on social media, the words they frequently used were indeed related to their activities and interests. Also, contrary to the our notion that their high-profile occupations would make them vulnerable to personal data leaks, the possibility of such leaks seemed relatively low.

By generalizing these findings to a broader sample, future research could confirm these patterns and investigate other unique characteristics of celebrities on social media. Such research could provide new insights into how celebrities use social media and could even suggest new scenarios for potential personal data leaks worth investigating.

References

[1] Meta Platforms. (2023, April 26). *Number of monthly active Facebook users worldwide as of 1st quarter 2023* [Statistics]. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[2] Bhatti, B. (2015). Social media and image management: an analysis of Facebook usage in celebrity public relations. *Media Watch, 6*(3), 339-352. https://doi.org/10.15655/mw/2015/v6i3/77896