Twitter Scrapping for Profiling Education Staff

by Herlawati Herlawati

Submission date: 17-Aug-2024 09:28AM (UTC+0700)

Submission ID: 2433252402

File name: IEEE-twitter_scraping-herlawati2020.pdf (1.13M)

Word count: 2927

Character count: 15941

Twitter Scrapping for Profiling Education Staff

1st Herlawati Herlawati *Informatics Engineering*
Universitas Bhayangkara Jakarta Raya
Bekasi, Indonesia 17121
herlawati@ubharajaya.ac.id

2nd Rahmadya Trias Handayanto
Computer Engineering
Universitas Islam 45
Bekasi, Indonesia 17113
rahmadya.trias@gmail.com

3rd Inna Ekawati
Computer Engineering
Universitas Islam 45
Bekasi, Indonesia 17113
inna.ekawati@gmail.com

4th Kardinah Indrianna Meutia *Management*
Universitas Bhayangkara Jakarta Raya
Bekasi, Indonesia 17121
kardinah.indrianna@dsn.ubharajaya.ac.id

5th Jelita Asian
School of Computer Science Nusa Putra University
Sukabumi, Indonesia 43152
jelitayang@gmail.com

6th Umar Aditiawarman
School of Computer Science Nusa Putra University
Sukabumi, Indonesia 43152
oemar99@gmail.com

***Abstract* — Social media (Facebook, Instagram, Twitter, etc.) have been widely used. They have many advantages, especially for business. However, such media sometimes invite negative effects, e.g. decreasing employee performance, conflict in a relationship, crime, etc. Therefore, this study proposes a method to scrap one of the social media, i.e. Twitter for profiling. Gephi application is used for network analysis after scrapping the network using Twecoll, a Python-based scrapping application. A web-based application is also created including the Apache-based server and Python-based script. The result shows that the scrapped account has several groups/communities including the weight of each connection. In addition, the result can be used for group profiling and additional analysis to complete the sentiment analysis based on tweets.**

Keywords—Gephi, Social Media, Twitter, Web-based Profiling, Twecoll

I. Introduction

Social media has become a popular application in the current global condition, especially when the meeting is difficult because of geographic location or a disaster, e.g. pandemic COVID-19. It can support e-learning, communication, e-commerce, and other online facilities. However, several negative effects have been reported, e.g. mental health problems, fraud, social conflict, etc. Several mental health problems from social media use have been studied ranged from loneliness, decreased empathy, to suicidality[1]–[3].

Employee productivity is a factor in the sustainability of a business. It relates to the activity of employees at works. For education, the teacher is the main business actor. The performance of the educational institution depends on the performance of the teacher. If there is no controlling activity from management to the social media activity of a teacher, some problems might arise regarding social conflict and sometimes becomes a criminal sentence. To overcome these problems, this study proposed a system to scrap an educational staff's Twitter account using network analysis.

Network analysis has been widely studied. This can be used to understand the relations among the nodes. Research on this topic is wide from a social, computer network, to medical. For example, a previous study on complementary and alternative medicine (CAM) in Iran [4]. The proteinprotein network analysis is done with other medical systems, e.g. proteomic tools, mitochondrial lysate isolated, etc. The network analysis used Gephi software similar to another study on immunological profiling in chronic rhinosinusitis with nasal polyps that creates a network of serum specimen for analyzing the growth factors VEGF and GM-CSF [5]. Geographic Information System (GIS) can also use network analysis for understanding land model within the earth system modeling framework [6]. A Community Land Model (CLM) is generated to understand the individual ecosystem function.

Another study uses network analysis for human mobility through the use of mobile access which is composed of a web analytics platform, and smartphone application [7]. This system generates a spider graph that is also integrated with the Wi-Fi network.

Nowadays, social media has become an important need for people around the world. Sentiment analysis of this online media has been widely studied[8]–[10]. This analysis only uses a particular hashtag and needs more analysis for the relation among account through the network. Many studies have been conducted on network analysis, e.g. group and community profiling in educational social networks [11], [12], pattern analysis, and comprehensive graph [13]–[15].

The use of social media profiling for employee monitoring is still rare. Therefore, social media scrapping is used in this study for employee profiling. Twitter is chosen since this social media has some characteristics, i.e. high speed in the viral process, easy to mention others, simple statements (tweets), and most people use the Twitter account.

The rest of the paper is organized as follows. The data and Method section discusses how to scrap the twitter network using a python-based application, Gephi implementation, and proposed system design. The results discuss some issues about the software implementation and the conclusion concludes the research.

II. Data and Method

A. Data Scrapper

Scrapping is an activity to gather information from the internet (website, social media, etc.) which uses a computer program. Data was collected using the Twecoll, a Pythonbased script (https://github.com/jdevoo/twecoll) which is supported only for Python 2 and for Python 3, another application, i.e. Nucoll, can be used (https://github.com/jdevoo/nucoll) [16], [17]. Because of the privacy of the Twitter account, this paper used the author's account for testing, i.e. @mrsherlawati. The case study is in Universitas Bhayangkara Jakarta Raya.

To scrap the Twitter network, there are at least three steps that are needed, namely: initialization, fetching, and network generation through init, fetch, and edge list functions, respectively. The result is a GML file which contains the JSON-based network for network analysis. Twecoll needs Twitter API from Twitter for a developer (https://developer.twitter.com/en). Four credentials are needed, i.e. API key, API Secret, Access Token, and Access Token Secret.

B. Network Analysis

Scrapping result which is a GML file is inserted to Gephi, a java-based application, to view its network. This study used Gephi version 0.92 that available to be downloaded from its official site (https://gephi.org/) [18]. Many researchers have been used Gephi, especially for network analysis.

Gephi presents both network and data table after inserting a GML file from the data scrapper. The graph contains the node and edge with the weight. A node represents a Twitter account and edges the directed graph. Weight represents the frequency of tweets (Fig 1). Edge consists of one-directional and bi-directional. The arrow direction is similar to the following direction on Twitter. If two nodes follow each other, the edge is called a friend.

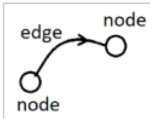


Fig. 1. Node and Edge

Not only generating a network, but Gephi also calculates some important statistics, i.e. average degree, avg. weighted degree, network diameter, graph density, HITS, Modularity, PageRank, Connected Components, Avg. clustering components, Eigenvector Centrality, Avg. Path Length, and another dynamic aspect (#nodes, #edges, Degree, and Clustering Coefficient). Not all those statistics have to use, only the important one is needed, e.g. network diameter, modularity, and eigenvector centrality. Other statistics can also be used in the future by running the process again.

To get a more user-friendly view, the network can be modified based on color, size, etc. Some layouts are available, i.e. contraction, expansion, force atlas, force atlas2, Fructhterman Reingold, label adjust, noverlap, open ord, random layout, rotate, yifan hu, and yifan hu proportional. In this study, the Fructhterman Reingold was chosen because it shows the network in a compact circle with separation based on the modularity class.

After network analysis, the GEXF-file should be exported for viewing the network online. Hence, for network analysis three files should be ready: DAT-file from initialization, GML-file from fetching, and GEXF-file from network analysis in Gephi.

C. Web-based Network Viewer

Gephi is a desktop application that difficult to share with other users. Therefore, a web-based network viewer is needed. This study used a web-based network viewer, called Gexf-Js (https://github.com/raphv/gexf-js) to view the network from Gephi online [19]. This application is under MIT license.

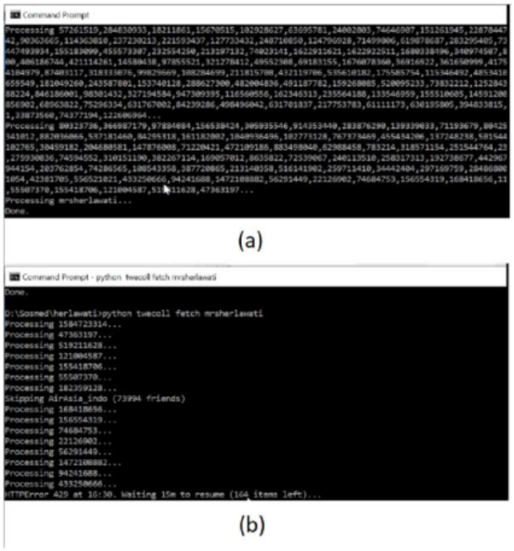
To run the Gexf-Js application, the following requirements should be prepared. First, a web server should be chosen because based on Cross-Origin Resource Sharing security AJAX is disabled to run locally. Therefore, some folder should be stored on the webserver. In this study, an Apache-based Server, Wamp Server, is used with some browser, e.g. Chrome, Mozilla, edge, etc. for accessing the webserver. For an online network viewer, only a GEXF-file for each employee profile is needed. All the processes in this study use a Windows-based platform.

III. Result And Discussion

The data scrapping step needs some hours to finish. For an active Twitter account having large numbers of followers and following, the fetching process might need a day because of the restriction from Twitter that a scrapping process must be paused for about 15 minutes after collecting about 15 accounts. For an account with a simple network, e.g. 100 nodes (followers, following, and friends), the scraping process was about 1.7 hours. Some researchers used an IoT, Raspberry Pi, with a mobile network to gather the data, instead of using a personal computer or laptop [20].

A. Data Scrapping Result

Fig 2 shows the initialization and fetching process in the data scraping step. A library is needed to run the Twecoll, i.e. Tweepy after PIP update. HTTP error 429 (Fig 2b) shows the restriction from Twitter to pause the scrapping activity for 15 minutes before continue until all items have been downloaded.



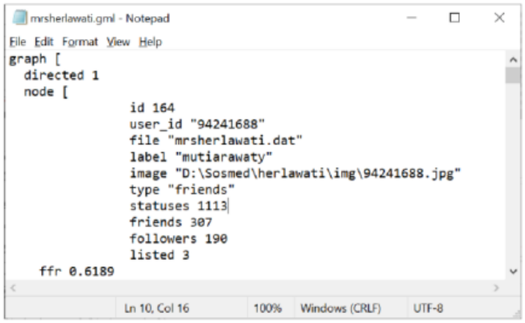
```
Processing 57261519,284830933,18211861,15670515,102928627,63695781,24002803,74646907,151261945,22878447
42,90363665,114363810,237230213,221593437,127733432,240710050,124796928,71499006,619078687,283295405,73
447493934,155183099,455573307,232554250,213197132,74023141,1622911621,1622922511,1680338496,340974507,8
00,406186744,421114261,14580438,97855521,321278412,49552308,69183155,1676078360,36916922,361650999,4175
4184979,87403117,318333076,99829669,1082284699,211815708,432119706,535610182,175585754,115346492,4853418
659549,181049260,243587801,153740128,288627300,482004836,491187782,159268085,520095233,73832212,1252843
88224,846186001,98301432,327194584,947309395,116560556,162346313,235564188,133546959,155510605,14591200
856902,68963822,75296334,631767002,84239286,498496042,631701837,217753783,61111173,630195805,394833815,
1,33873560,74377194,122606964...
Processing 80323736,366987179,97884084,156538424,305935546,914353440,283876290,139339033,71193679,80425
341012,882036066,537181460,84295318,161182002,1040936496,102773128,767373469,455434206,137248238,581544
102765,38459182,204680581,147876008,71220421,472109186,883498040,62988458,783214,318571154,251544764,23
,275930836,74594552,310151190,382267114,169057012,8635822,72539067,240113510,258317313,192738677,442967
944154,203762854,74286565,108543358,387720865,213140358,516141902,259711410,34442404,297169759,28486800
1054,42381705,556521021,433250666,94241688,1472108882,56291449,22126902,74684753,156554319,168418656,11
,55507370,155418706,121004587,519211628,47363197...
Processing mrsherlawati...
Done.
```

```
Done.

D:\Sosmed\herlawati>python twecoll fetch mrsherlawati
Processing 1584723314...
Processing 47363197...
Processing 519211628...
Processing 121004587...
Processing 155418706...
Processing 55507370...
Processing 182359128...
Skipping AirAsia_indo (73994 friends)
Processing 168418656...
Processing 156554319...
Processing 74684753...
Processing 22126902...
Processing 56291449...
Processing 1472108882...
Processing 94241688...
Processing 433250666...
HTTPError 429 at 16:30. Waiting 15m to resume (164 items left)...
```

Fig. 2. Initialization (a) and Fetching the network (b)

The last step in data scrapping is GML-file creation using the edge list function. This step does not need much time because it just creates the information about the scrapped node, e.g. id, user_id, file, label, image location, type relation, statuses, friends, followers, and listed (Fig 3).



```
graph [
  directed 1
  node [
               id 164
               user_id "94241688"
               file "mrsherlawati.dat"
               label "mutiarawaty"
               image "D:\Sosmed\herlawati\img\94241688.jpg"
               type "friends"
               statuses 1113
               friends 307
               followers 190
               listed 3
    ffr 0.6189
```

Fig. 3. GML-file structure

GML-files is the biggest file in data scrapping (from a hundred kilobytes to megabyte) according to the size of its network. This file is the source of the Twitter network of an account.

B. Network Analysis Result

A new project should be prepared after installing Gephi 0.9.2. After opening the GML-file from the menu, the raw network appears (Fig. 4). It only shows the original node and edge. The label of the network can also be presented. Of course, further analysis should be conducted to get the network meaningful.

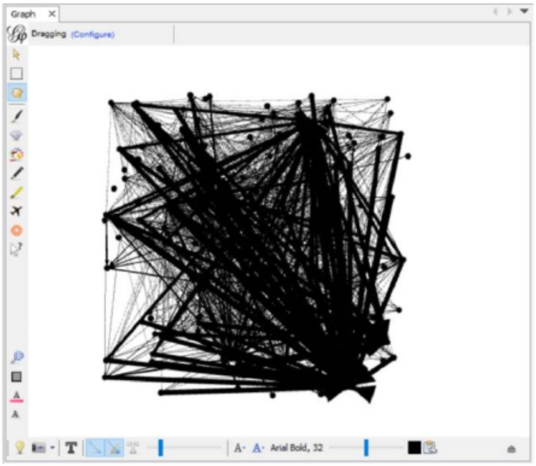
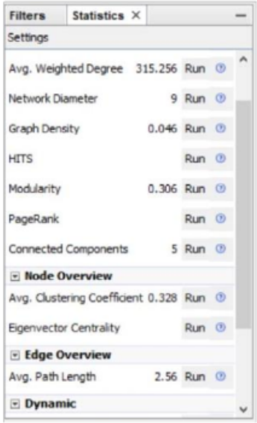


Fig. 4. Raw Network from GML-file

Some statistics manipulation to get the meaning of the networks should be done. Gephi provides statistics calculation in the filters and statistics tab (Fig 5). The important statistics parameters are betweenness centrality, modularity class, and weight.



Avg. Weighted Degree	315.256	Run
Network Diameter	9	Run
Graph Density	0.046	Run
HITS		Run
Modularity	0.306	Run
PageRank		Run
Connected Components	5	Run
Node Overview		
Avg. Clustering Coefficient	0.328	Run
Eigenvector Centrality		Run
Edge Overview		
Avg. Path Length	2.56	Run
Dynamic		

Fig. 5. Statistics Module Result

Fig 6 shows the Twitter network after appearance setting, i.e. unique, partition, and ranking. Partition was used to separate by color its modularity class (grey, red, pink, and green). Nodes were ranked by their betweenness centrality in terms of circle class. The size of the edge represents the weight. Betweenness centrality shows the influence of an account in the network, in this graph is represented by a circle. The network should be export into GEXF-file for Gexf online viewer.

The Fruchterman Reingold layout was chosen with the circle environment style and the modularity classes are separated around the edge. This layout will include in the GEXF-file, hence, when it is viewed in the online viewer, the circle layout with its properties is also presented. In addition, before exporting to GEXF-file, the fixed layout should be decided since it cannot be changed in an online viewer but must use Gephi for modification.

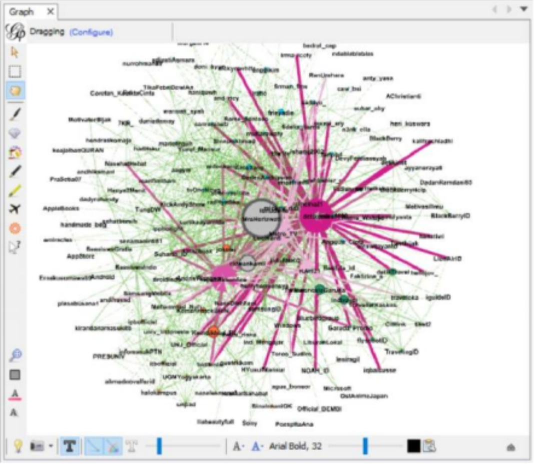
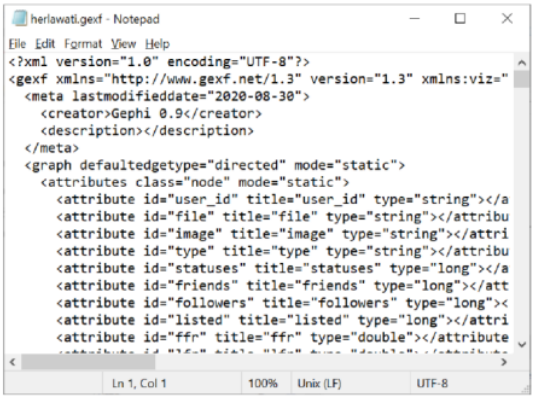


Fig. 6. Twitter Network Result

C. Web-based Network Viewer

After installing the web server, the Gexf-Js application was implemented by inserting its folder in htdocs or www folder in the apache-based server. This folder was modified by inserting the GEXF-file, the network-analysis result file from Gephi (Fig 7).

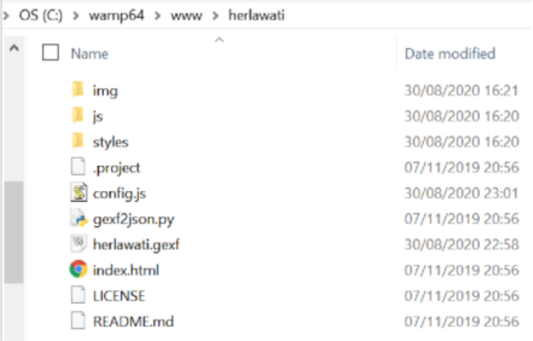


```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.3" version="1.3" xmlns:viz="
  <meta lastmodifieddate="2020-08-30">
    <creator>Gephi 0.9</creator>
    <description></description>
  </meta>
  <graph defaultedgetype="directed" mode="static">
    <attributes class="node" mode="static">
      <attribute id="user_id" title="user_id" type="string"></a
      <attribute id="file" title="file" type="string"></attribu
      <attribute id="image" title="image" type="string"></attri
      <attribute id="type" title="type" type="string"></attribu
      <attribute id="statuses" title="statuses" type="long"></a
      <attribute id="friends" title="friends" type="long"></att
      <attribute id="followers" title="followers" type="long"><
      <attribute id="listed" title="listed" type="long"></attri
      <attribute id="ffr" title="ffr" type="double"></attribute
```

Fig. 7. GEXF-File Result from Gephi

GEXF-file is a JavaScript Object Notation (JASON), a kind of semantic to represent semi-structured data. After inserting this file and edit the Config.js file by renaming the GEXF-file, the GEXF online viewer shows the Twitter network. Fig 8 shows the GEXF online viewer.

Fig 8 shows the file needed in web-server. A gexf2json.py is a python file for creating a web-based network. An index.html will serve the layout with styles.

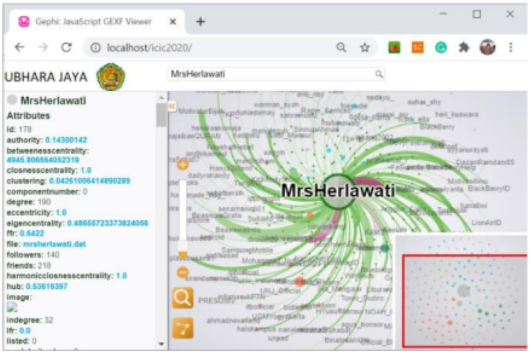


Name	Date modified
img	30/08/2020 16:21
js	30/08/2020 16:20
styles	30/08/2020 16:20
.project	07/11/2019 20:56
config.js	30/08/2020 23:01
gexf2json.py	07/11/2019 20:56
herlawati.gexf	30/08/2020 22:58
index.html	07/11/2019 20:56
LICENSE	07/11/2019 20:56
README.md	07/11/2019 20:56

Fig. 8. Files for GEXF viewer

Since the proposed system is confidential, in implementation, the GEXF viewer only runs in the intranet (local network) with a strict user and password.

Fig 9 shows the GEXF viewer from Gephi. It shows the nodes and edges that can be chosen easily. The searching box is also included in this viewer. Navigation is shown in the lower right comer. Attributes are shown on the left side. Drag and scrolling mouse can be used to navigate the network. Double-clicking the mouse will show the edges (inbound and outbound links).

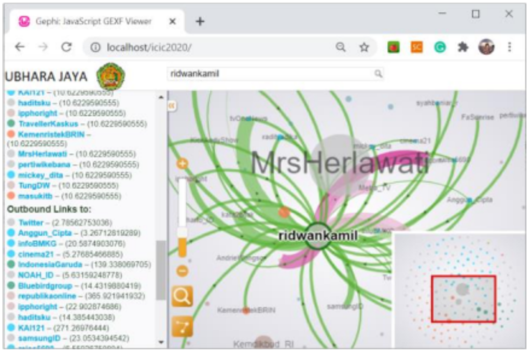


MrsHerlawati
Attributes
id: 178
authority: 0.14300142
betweenesscentrality: 4945.806564052319
closnesscentrality: 1.0
clustering: 0.04261006414890289
componentnumber: 0
degree: 190
eccentricity: 1.0
eigencentrality: 0.48655723373824066
ffr: 0.6422
file: mrsherlawati.dat
followers: 140
friends: 218
harmonicclosnesscentrality: 1.0
hub: 0.53619397
image:
indegree: 32
lfr: 0.0
listed: 0

Fig. 9. GEXF Viewer of Twitter Network

The system can also analyze the characteristic of followers, following accounts, or friends by checking the influencer in the network by selecting the big circle as shown in Fig 10. A big circle represents the high number of betweenness centrality which shows the influence of a node on the network. Big edge usually from a website or the influential people in the network.

Large numbers of nodes show that many relations have been created. They also show that the account has used Twitter frequently. If the tweets are positive, it can add some benefits e.g. organization advertisement, education information, etc. but if negative or neutral, it has a risk. It might decrease work performance as a distraction.



haditsku – (10.6229590555)
ipphoright – (10.6229590555)
TravellerKaskus – (10.6229590555)
KemenristekBRIN – (10.6229590555)
MrsHerlawati – (10.6229590555)
pertiwikebana – (10.6229590555)
mickey_dita – (10.6229590555)
TungDW – (10.6229590555)
masukitb – (10.6229590555)
Outbound Links to:
Twitter – (2.78562753036)
Anggun_Cipta – (3.26712819289)
infoBMKG – (20.5874903076)
cinema21 – (5.27685466885)
IndonesiaGaruda – (139.338069705)
NOAH_ID – (5.63159248778)
Bluebirdgroup – (14.4319880419)
republikaonline – (365.921941932)
ipphoright – (22.902874686)
haditsku – (14.385443038)
KAI121 – (271.26976444)
samsungID – (23.0534394542)

Fig. 10. An influencer in the network

The characteristic of the modularity class should be checked further by analyzing the relation. The modularity class can be seen easily in the navigation box in the lowerright of the GEXF viewer. For example, the blue modularity class in the upper region consists of friends, who are lecturers, in university. Orange modularity class contains some institutional Twitter account, whereas the dark green modularity class consists of a travel account. Another modularity class is a public figure (Fig 11).

The Twitter network is a dynamic network that might change over time. Therefore, the update should be done for future analysis. In addition, the scrapping activity should choose the right account for the targeted employee.

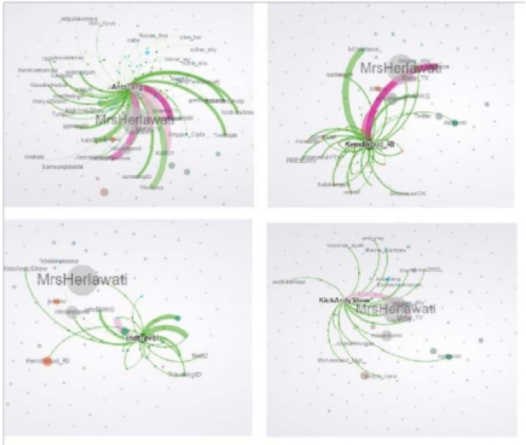
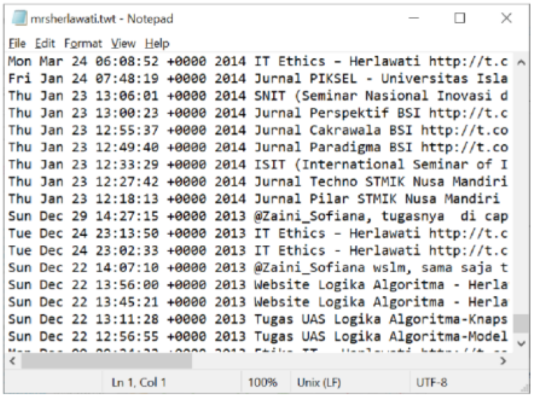


Fig. 11. Follower, Following, and friends Analysis

The Twitter network needs more analysis by the human resource department or other staff who have responsibility for human resources for promoting an employee to a particular position.

After creating the Twitter network, the targeted Twitter account can also be profiled by analyzing its tweets. Data scrapper also provide with tweets scrapping using tweets function. Fig 12 shows the result of the tweet scrapping of a

particular Twitter account.

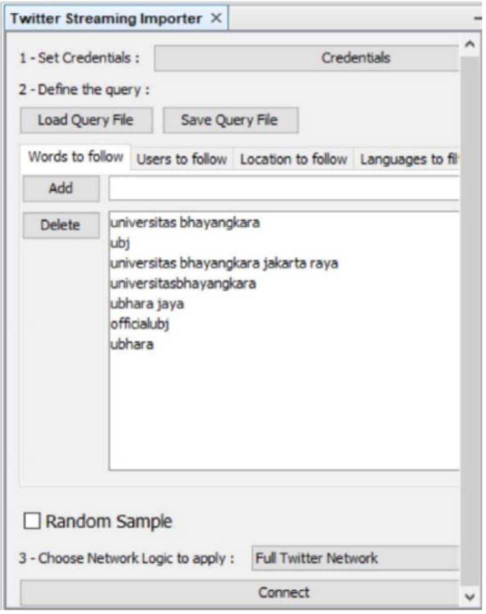


```
Mon Mar 24 06:08:52 +0000 2014 IT Ethics - Herlawati http://t.c
Fri Jan 24 07:48:19 +0000 2014 Jurnal PIKSEL - Universitas Isla
Thu Jan 23 13:06:01 +0000 2014 SNIT (Seminar Nasional Inovasi d
Thu Jan 23 13:00:23 +0000 2014 Jurnal Perspektif BSI http://t.c
Thu Jan 23 12:55:37 +0000 2014 Jurnal Cakrawala BSI http://t.co
Thu Jan 23 12:49:40 +0000 2014 Jurnal Paradigma BSI http://t.co
Thu Jan 23 12:33:29 +0000 2014 ISIT (International Seminar of I
Thu Jan 23 12:27:42 +0000 2014 Jurnal Techno STMIK Nusa Mandiri
Thu Jan 23 12:18:13 +0000 2014 Jurnal Pilar STMIK Nusa Mandiri
Sun Dec 29 14:27:15 +0000 2013 @Zaini_Sofiana, tugasnya  di cap
Tue Dec 24 23:13:50 +0000 2013 IT Ethics - Herlawati http://t.c
Tue Dec 24 23:02:33 +0000 2013 IT Ethics - Herlawati http://t.c
Sun Dec 22 14:07:10 +0000 2013 @Zaini_Sofiana wslm, sama saja t
Sun Dec 22 13:56:00 +0000 2013 Website Logika Algoritma - Herla
Sun Dec 22 13:45:21 +0000 2013 Website Logika Algoritma - Herla
Sun Dec 22 13:11:28 +0000 2013 Tugas UAS Logika Algoritma-Knaps
Sun Dec 22 12:56:55 +0000 2013 Tugas UAS Logika Algoritma-Model
```

Fig. 12. Tweets Scrapping Result

Tweets scrapping focuses on the detailed activity that shows the view of an employee about a topic. Hate speech, bullying, defamation, etc. can be detected earlier. This scrapping is only used when an education staff needs more intensive analysis after seeing the twitter network in GEXF viewer.

Sentiment analysis might be used to analyze the tweets after some preprocessing (text cleaning, tokenization, and other information retrieval tools). Scrapping from other social media might be needed for getting the complete analysis. Another implementation is also for searching the active account for a particular hashtag or word to follow as shown in Fig 13. It needs the TweeterStreamingImporter plugin to scrap the network based on words to follow, users to follow, location to follow, and languages to follow.



1 - Set Credentials :

2 - Define the query :

universitas bhayangkara
ubj
universitas bhayangkara jakarta raya
universitasbhayangkara
ubhara jaya
officialubj
ubhara

□ Random Sample

3 - Choose Network Logic to apply : Full Twitter Network

Fig. 13. Twitter Streaming Importer

IV. Conclusion

Education organization is a service business that needs a teacher/lecturer to deliver education. The performance of the education institution depends on the performance of their education staff. Employee profiling can be used to ensure that their education staff does not risk mental health problems from social media. This can also be used for promoting a candidate for a strategic position in the organization. Twitter network analysis is proposed in this study as part of employee profiling. The result shows the influence account in a profiled employee that might affect the mind of educated staff. This can also as an early warning before the problem arose related to electronic information and transaction act, terrorist activity, and other negative communities. Other social media, e.g. Facebook, Instagram, etc., can also be implemented using this proposed system in future research.

Acknowledgment

The authors thank Universitas Bhayangkara Jakarta Raya and Universitas Islam 45 Bekasi for funding the research and all reviewers for their insightful comments. The authors also thank Gephi and other open-source applications that are used in this study.

References

[1] C. Berryman, C. J. Ferguson, and C. Negy, “Social Media Use and Mental Health among Young Adults,” *Psychiatr. Q.*, vol. 89, no. 2, pp. 307–314, 2018.

[2] J. Gao *et al.*, “Mental health problems and social media exposure during COVID-19 outbreak,” *PLoS One*, vol. 15, no. 4, pp. 1–10, 2020.

[3] M

2

Control., vol. 18, no. 2, pp. 640–648, 2020.

[21]

Twitter Scrapping for Profiling Education Staff

ORIGINALITY REPORT

11%	10%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	ieeexplore.ieee.org Internet Source	2%
3	repository.ubharajaya.ac.id Internet Source	2%
4	dokumen.pub Internet Source	1%
5	sci-hub.ru Internet Source	1%
6	slideplayer.com Internet Source	1%
7	Jelita Asian, Moneyta Dholah Rosita, Teddy Mantoro. "Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods", Jurnal Online Informatika, 2022 Publication	1%

8	Herlawati Herlawati, Rahmadya Trias Handayanto, Prima Dina Atika, Sugiyatno Sugiyatno et al. "Semantic Segmentation of Landsat Satellite Imagery", 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022 Publication	1 %
9	hdl.handle.net Internet Source	<1 %
10	Herlawati, Rahmadya Trias Handayanto, Didik Setiyadi, Endang Retnoningsih. "Corpus Usage for Sentiment Analysis of a Hashtag Twitter", 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		