International Journal of Advances in Data and Information Systems
Vol. 2, No. 1, April 2021, pp. 53~61
ISSN: 2721-3056, DOI: 10.25008/ijadis.v2i1.1214

Features Selection for Entity Resolution in Prostitution on Twitter

Reisa Permatasari[1] and Nur Aini Rakhmawati[2]
[1,2]Departement of Information System, Institut Teknologi Sepuluh Nopember
Jl. Raya ITS, Keputih, Sukolilo, Surabaya, 60111, Indonesia

Article Info

Article history:

Received Feb 11, 2021
Revised Mar 23, 2021
Accepted Apr 03, 2021

Keywords:

entity resolution
online prostitution
regularized logistic regression
twitter

ABSTRACT

Entity resolution is the process of determining whether two references to real-world objects refer to the same or different purposes. This study applies entity resolution on Twitter prostitution dataset based on features with the Regularized Logistic Regression training and determination of Active Learning on Dedupe and based on graphs using Neo4j and Node2Vec. This study found that maximum similarity is 1 when the number of features (personal, location and bio specifications) is complete. The minimum similarity is 0.025662627 when the amount of harmful training data. The most influencing similarity feature is the cellphone number with the lowest starting range from 0.997678459 to 0.999993523. The parameter - length of walk per source has the effect of achieving the best similarity accuracy reaching 71.4% (prediction 14 and yield 10).

Corresponding Author:

Nur Aini Rakhmawati,
Department of Information System, Institut Teknologi Sepuluh Nopember
Jl. Raya ITS, Keputih, Sukolilo, Surabaya, 60111, Indonesia
Email: nur.aini@is.its.ac.id

1. INTRODUCTION

The use of information, media, and communication technology have changed both the behavior of people and human civilization globally. Information globalization has placed Indonesia as part of the global information community, which requires the establishment of regulations on managing information and electronic transactions at the national level. In line with the development process and the era of globalization, as well as the increasing quality of life technology in Indonesia, people have experienced many changes. Electronic media has had a significant influence on society. These influences can be either positive or negative influences.

The government, on April 26, 2008 passed the enactment of the law on Information and Electronic Transactions (ITE). The Law on Information and Electronic Transactions are intended to provide many benefits, including to ensure legal certainty for people who conduct electronic transactions, encourage economic growth, prevent the occurrence of information technology-based crimes and protect the public service users by utilizing information technology. ITE Law meant to regulate information technology-based crimes (cybercrimes); one of the regulations is illegal content, which consists of decency, gambling, defamation, threatening and extortion, namely Article 27, Article 28, and Article 29 of the Law ITE [1].

ITE Law No. 19 of 2016 amended due to the rise of online crime such as online prostitution through social media networks. These crime cases caught up via Twitter. The latest and still fresh in the minds of the people of Indonesia because it became the opening news in 2019, namely news about prostitute involving famous television artists [2]. Not long after that, the authorities seemed to increase serious efforts to eradicate online prostitution on social media, because it not only involved national-level artists, artists on social media Instagram or who are often called celebrities were also involved in this immoral business practice [3]. The public starts to think of this illegal business because it can also include children [4] and female students [5] who should be far from the industry. The internet crime has become a black world business trend. This black business manager utilizes free social media as a forum to market his "hot products."

This social problem is indeed not yet something that is considered an emergency to resolved by the government, but with the loss of many victims in the present, it is possible to cause more victims in the future because the perpetrators easily hide behind the anonymous identity of social media. Many parties who want to help solve this problem, including academics, one of them is by applying Entity resolution to find the real identity of the many social media accounts that have one person or the similarity of people behind the operation. Entity resolution itself is the process of determining whether two references to real-world objects refer to the same object or different objects. An entity is a term describing a real-world object, a person, a place or an object. Research on Entity resolution and the like that uses other names is mostly done, depending on the data source and the problem.

There are some researches that related to online prostitution. Nagpal et.al implement entity resolution models to identify the source of escort advertising in human trafficking cases [6]. The source of advertising becomes the main object of search as an effort to fight human trafficking, so escort type ads must be removed from the main root because usually advertising sources have many subjects traded.Entity resolution also applied in social media in studies that detect real personal entities from two different social media accounts namely Facebook and Xing with the web crawling method, by extracting the main features of social media which usually include personal data of social media users [7]. There are also studies on two more popular social media and have different features namely Twitter and Facebook [8].

The main objective is to find out the facts of entities in the real world that come from various data sources both data in a system, social media and others. Some examples include focusing on the algorithms used for entity resolution; "Comparative analysis of approximate blocking techniques for entity resolution" [9], evaluating Entity resolution in real-world problems; "Evaluation of Entity resolution approaches on real-world match problems" [10], even making Entity resolution for social media "D-dupe: An interactive tool for Entity resolution in social networks" [11].

The research above inspired this research and tried to help real-world problems with data from social media. The related research that forms the basis of this research is a study conducted by Nagpal [6], where the research builds an Entity resolution model to identify the source of escort advertising in human trafficking cases. Data obtained from escort type ads on the site www.backpage.com based on ad details that include personal data on the subject of human trafficking. The source of advertising becomes the main object of search as an effort to fight human trafficking, so escort type ads must be removed from the main root because usually advertising sources have many subjects traded. So, if only solve trade issues, and then there is still another subject.

Entity resolution applied in social media in research [7] which detected real personal entities from two different social media accounts. The habit of internet users who have more than one social media account makes there is data redundancy in the search for an entity. The data source is taken from Facebook and Xing's social media using the web crawling method by extracting the main features of social media which usually include the personal data of the users of social media.

Goga [8] used two more popular social media and has different features namely Twitter and Facebook. But many real users who have both accounts, even in one type of social media can

have more than one account. Retrieval of property in feature extraction using Availability, Consistency, Non-Impersonability, and Discriminability (ACID) Framework.

Based on the explanation above, entity resolution has the opportunity to become a new solution in detecting social media accounts that commit crimes with examples of online prostitution cases. As for what makes opportunities open on this research topic because there is no research on Entity resolution based on machine learning in the field of online crime with data sourced from Twitter social media with user locations in Indonesia. It hoped that if this research can apply entity resolution in examples of cases of prostitution and online crime, then at least it is expected to be able to provide insight into how the parties concerned about the problem in Indonesia, trying to find solutions that can be immediately applied to find and arrest the perpetrators.

2. RESEARCH METHOD

The methodology our research can be explained as follows:

2.1 Data Collection

We collect data from Twitter, especially from accounts that carry out promotions, transactions and consumption of online prostitution. Gaffney and Puschmann in the section of the book "Twitter and Society" [12] entitled "Data Collection on Twitter" states [13] that instead of offering a single API, three different Twitter data interfaces are available for researchers who wish to inquire about services: Streaming API, REST API and Search API. With few exceptions, the collection of research produced to date has relied on data collection through one of these three sources [13].

a. Streaming API
b. REST API
c. Search API

Given the limited length of tweets and the style of posts produced by this limitation, information that would be valuable for search purposes is rarely available in the form of tweets on the surface. The challenge is how to extract useful semantics from tweets

There is the primary data source taken with TwitterScraper tools. Based on the explanation in the previous paragraph about data sourced from the Twitter API, Twitter has provided a REST API that can be used by developers to access and read Twitter data. They also give a Streaming API that used to access Twitter data in real-time. Most software written to access Twitter data provides libraries that function as wrappers around the Twitters Search and Streaming API's and are therefore limited by API limitations.

The Twitter Search API can only send 180 requests every 15 minutes. With a maximum number of 100 tweets per Request, this means you can mine 4 x 180 x 100 = 72,000 tweets per hour. Using TwitterScraper is not limited by this number but by internet speed/bandwidth and the number of instances TwitterScraper is willing to start. One of the more significant disadvantages of the Search API is that it can only access Tweets in the last seven days, which becomes a significant obstacle for anyone looking for past data to make a model.

2.2 Feature Extraction

Guyon and Elisseef [14] in the book section entitled "An Introduction to Feature Extraction" states that machine learning problems occur when there are tasks that are defined by a set of cases or examples with predetermined rules. These problems found in various application domains, ranging from engineering applications in robotics and pattern recognition (speech, handwriting, face recognition) to Internet applications (text categorization) and medical applications (diagnosis, prognosis, drug discovery). Several "training" examples (also called data points, samples, patterns, or observations) related to desired results, the machine learning process consists of finding relationships between patterns and outcomes using solely training examples. This shares much with human learning where students are given examples of what is right and what is not and must deduce which rules underlie decisions. To make it concrete, consider the following example: the data point or case is the clinical observation of the patient and the result is health

status: healthy or suffering from cancer. The aim is to predict unknown results for new "test" examples, for example the health status of new patients. Performance on test data is called "generalization." To do this task, one must build a prediction or predictor model, which is usually a function with adjustable parameters called "machine learning." Training examples used to choose the optimal parameter set.

Features are synonyms of input variables or attributes. Finding a good data representation is very specific to the domain and is related to available measurements. There are four aspects to feature extraction:

a. feature construction;
b. creating a subset of features (or search strategies);
c. estimated evaluation criteria (or valuation methods);
   definition of evaluation criteria (e.g., relevance index or predictive power);

The last three aspects relevant to feature selection and schematics summarized in the figure below. Filters and wrappers are mostly different from evaluation criteria. Cleaners use approaches that do not involve any machine learning, for example relevance indexes based on correlation coefficients or test statistics, while wrappers use machine learning performance that trained using a subset of features provided.

This research will use three main types of features as well, with adjustments that the data source used is only one type of social media, namely Twitter. The main types of feature adjustments are:

a. Name-based: username and profile name
b. user-based information: information in the bio column consisting of text, location, date of birth, join since, link
c. based on social networking topology: followers and following data flow based: tweet content in the form of text, hashtags, and designations

Cluster extraction and identification of records related to online prostitution use researchers' observations as Twitter users. Initial considerations made to determine sub-features, for example in the bio column in the form of text; prostitution twitter accounts usually include TB (Height), BB (Weight), bra size and telephone numbers that can able to contact

Data verification finished after if the account posting contains hashtags that are used for data collection and through the initial sorting of data by the author. The amount of data taken from the period of January - June 2019 during the study.

2.3 Preparation of Training Data and Test Data

To get compared results, then in the research must build training and testing dataset, researchers conducted the following steps:

a. Reviewing data collection from Twitter with scrapping techniques with particular keywords from features in the Twitter account.
b. The data that has been collected was re-selected to get accuracy whether included in the group of online prostitution accounts.
c. Some data included in the online prostitution account group set aside to build the initial dataset.

Apply the M model to match the new pair. Data that processed at the preprocessing stage divided into two types of data namely training data and test data. Alternately, some groups used as training data and test data.

2.4 Entity resolution process

In this stage, we compare the results to conclude which data set produces the best entity. Dedupe assisted the determination of entity resolution. Dedupe is a library that uses machine learning to quickly deduplicate and resolve objects on structured data [15]. Examples of problems that can be helped by dedupe:

a. Remove duplicate entries from the name and address spreadsheet

b. link the list with customer information to another with order history, even without a unique customer-id

take a database of campaign contributions and find out which ones are made by the same person, also if the name entered a little differently for each note

Dedupe retrieves human training data and makes the best rules for datasets to quickly and automatically find similar records, even with extensive databases. Dedupe.io is a full-service web service supported by dedupe to remove duplicates and find matches in messy data. And provides an easy-to-use interface and includes cluster review and automation, as well as advanced recording links, continuous matching, and API integration. Dedupe features:

a. machine-learning - read data labeled human automatically create optimal weights and blocking rules
b. run on laptops - make smart comparisons, so there is no need for a powerful server to run it built as a library - so that it integrated into applications or import scripts
c. extensible - supports adding particular data types, string comparators, and blocking rules
d. open-source - anyone can use, modify or add it

2.5 Experiment

In this section, we will explain the details of the process of experiment carried out following each of the trial scenarios previously described.

Data Pre-processing

Preprocessing implemented to avoid incomplete data, data disturbances, and inconsistent data. The pre-processing text stages in this study are:

a. Retrieve only username data by deleting rows containing more than one same username
b. Re-scrapping the last five tweets from the account list, to get text content that states availability
c. Remove the symbols and special characters included in the tweet text
d. Removing Special Twitter Characters. This process is deleting special Twitter characters such as hashtags, user names (@username), and special characters (for example, RT, which indicates that the user is retweeting something).

Remove Symbols. This step removes the symbols and punctuation in the tweet.
After pre-processing, 870 Twitter account tweets obtained.

Scenario

A trial scenario made to achieve the research objectives. The trials in this study, in general, are the entity resolution process trials. Besides, the trial process to find out the effect of the number of features and parameters on similarity between online prostitution accounts using entity resolution. A more detailed explanation in each section.
There are 2 test scenarios,

a. Based on bio features: to see if there are differences in similarity results if the number of features is determined. To find out how much influence each feature has on the equality of online prostitution accounts.
b. Based on the tweet feature: to see differences in similarity results when the parameters vary.

Both are to find out how much influence each scenario has on the equality of online prostitution accounts. Whereas the data processed in Node2Vec, parameter configuration is done manually by testing four times. The parameter that changed in each test is the -l parameter.

Feature Extraction

In this process, bio information account retrieval is usually in the form of personal measurements such as TB (height), BB (weight), Cup (bra size), Age, Services offered ranging from massage, BJ (blowjob), ML (making love), etc., Use of contraception, mobile number. But only the most prostitution accounts will be used, namely: location, cellphone number, weight, height, bra size.

Next, the retrieval of tweet information is usually in the form of text and hashtags. Extraction from 870 individual account lists by taking the top 10 hashtags from the last five tweets of each account.

Data Training and Testing

Training and testing of data from personal information on bio carried out at Dedupe by entering all the data totaling 870 data rows. While the tweet text data does not go through the training and testing process. The data for the Dedupe process compiled in CSV file format.
Dedupe suggested a minimum of 20 training data consisting of 10 training data with positive values and ten training data with negative value.

Entity Resolution Process

The entity resolution uses the Dedupe tool by taking a single data comparing it with other data. An example of the process in the table below where data with id 212 compared with unpaired data and a match obtained with id data 234. Then it is entered into cluster ID 3. Each scenario executed in the same way.

Furthermore, for testing on tweet text data, the tools are Neo4j and Node2Vec. The initial process is creating labels on Neo4j, which is the process of representing each entity into a separate node. There are 768 data cells from 191 rows of usernames, locations, hashtags filtered from the extraction stage of the top 10 hashtags from the last five tweets of each account used in the previous process.

3. RESULTS AND DISCUSSION

All scenarios processed through the same stages but there are differences in the amount of positive training data and negative training data because the data compared is random data, so training is stopped until at least positive training data appears. Tabel 1 compares processes between the scenarios.

Table 1. Comparison of Dedupe outputs for each scenario

Scenario	Training Data	Duplicate	Similarity Confidence Score (SCS)	
	Positive	Negative	Set	Total
1	10/10	4/10	60	338
2	6/10	120/10	1	204
3	7/10	21/10	25	8
4	2/10	19/10	14	31

The results obtained are in the form of vectors. Based on research conducted by Grover and Leskovec [16], the number of dimensions used in this study is 16. Table 2 shows the result of a test to analyze the similarities that might arise. The keywords are taken from one sample in the dataset.

Table 2. The results of the trial search for the similarity of the username by keyword with the keyword: "luna_sby"

Parameter : .-d:16 -q:2 -l:80	
102	BebbyNday
45	AdiraAz39274945
115	Izanami92
61	salma1150
110	yuichen9
Parameter : .-d:16 -q:2 -l:40	
84	NadiaCinta18
101	Vio87612909
30	Devi08756597
127	Sella29824444
174	openbojakarta01
157	Hiperzcikarang
168	Aldira27792806
12	AtikaDomher
Parameter : .-d:16 -q:2 -l:20	
113	Enjel62255952
42	booking_AMEL
31	lestariaja889
162	NonVia5
3	Tyra_Agatha

Parameter : .-d:16 -q:2 -l:10	
40	ElsaVanessa5
154	AgnesTa98053066
116	indah_malng
125	Audrey47371112
19	AngelWpYK
168	Aldira27792806
152	AngelinaVrn5

The pilot phase has been carried out and produced various values. The Dedupe produces output for each scenario with the same positive training data values, but negative training data are different. That happens because the process of labeling inputs forced to stop after the minimum positive training data reaches a minimum of 10. From the previous table we can analyze some of the relationships or factors that influence the similarity using entity resolution.

The selection of features both in terms of number and type influences the similarity value in Dedupe's output in the form of Confidence Score. The highest maximum amount that achieved is one and the lowest minimum is 0. In this study, the maximum value reached by scenario one which lists full features, the lowest maximum value achieved by scenario two that does not include cellphone number. While the smallest value but not until 0 performed by scene 2, namely 0.025662627, and the most considerable minimum value achieved by scenario three which did not include a complete bio text. So, it can be concluded that the cellphone number feature becomes the clearest separator in the process of entity resolution online prostitution cases, while the bio feature in full-text form adds to the accuracy of the prediction.

This study also found that the negative training data influenced the amount of data considered to have similarities (duplicates), of course, with varying values. The less negative training data that is input, the number of similar data predictions will be more and more. It appears in scenario one which included four negative training data resulting in the highest amount of data valued at duplicate 338 data, and in scene 3 which included 21 negative training data produced in the number of data assessed copies as low as 8 data. Negative training data also affects Similarity Confidence Score of more than 0.8, where the higher the SCS, the more similar. The less negative training data entered, the higher the SCS value. It appears in scenario two which came 120 negative training data that did not produce data that had SCS values more than 0.8, and in scenario three which included 21 negative training data resulted in the number of data that had SCS values more than 8.

Feature selection also affects the value of minimal confidence score. The scenario that does not include the cell phone number in the feature is scenario produces`low of confidence score. We analys the output of each parameter released by Node2Vec. We only observe on the hashtag "#realangels" which included in the top 10 hashtags that used the keyword username "luna_sby." From these tests, the best results obtained in condition A with the settings -d: 16 -q: 2 -l: 80 (Table 3)

Table 3. Success rate of Node2Vec

Skenario	Prediction	Result	Accuracy
A	14	10	71,4 %
B	18	10	55,55%
C	15	8	53,33%
D	17	12	70,58%

Table 4. Comparison of ER Features and ER Graph

ER Feature		ER Graph	
BebbyNday	0.97033852	luna_sby	0.42266047
AdiraAz39274945	0.96567822	inginkan_malam	1
Izanami92	0.96486527	jhesii_jilbabbo	1
salma1150	0.96073806	rere06070970	0.42266047
yuichen9	0.95355344		
Average ER Feature	0.9630347	**Average ER Graph**	0.71133023

Table 4 shows that the ER similarity value of the feature evenly distributed while the similarity value of the ER Graph occurs if there is an imbalance between maximum similarity and partial similarity. Thus, the ER Feature works better when the use of a complete feature and ER Grap using hashtag and location.

4. CONCLUSION

The application of entity resolution for online prostitution in Indonesia is based on 6015 lines of data from Twitter. The initial training data is based on the author's assumptions of accounts that meet the criteria for online prostitution accounts, compared with 870 account data. Feature selection is based on information obtained, namely, bio, which broken down into TB, BB, cup, cellphone number, then location and hashtag.

The choice of features in terms of numbers affects the similarity value in Dedupe's output in the form of Confidence Score. The highest maximum amount that achieved is one, and the lowest minimum is 0. The maximum similarity is 1 when the number of features (personal specifications, location and bio) is complete. The cell phone number significantly impacts the similarity where the result ranges from 0.997678459 to 0.999993523. At the same time, the bio feature in full-text form adds to the accuracy of predictions. The negative training data affect similarity. The similarity is at least worth 0.025662627 when more and more amount of negative training data. The less negative training data entered, the number of similar data predictions will be more numerous, and the SCS value will be higher. Feature selection also affects the amount of minimal confidence score. Scenarios that do not include the cell phone number on the feature make the average value low. Hashtag data processing in Node2Vec found that the parameter - length of walk per source gives the effect. The best similarity accuracy reaches 71.4% (14 predictions and ten results) when the parameter is -d: 16 -q: 2 -l: 80.

The data processed in this study varies from short characters, numbers, text to hashtags. Therefore, the personal data should be retrieved manually from the bio and the tweet since there might be data that do not match the conditions due to human error even though the information is correct because it takes from the same account. Then for the tweet data, it is hoped that further research will be able to process the text of the tweet by first processing it into word fragments so that new research expected to be more accurate. The explicitly stated features should not be included because similarity searches will be more interesting if using implied data.

REFERENCES

[1] Republik Indonesia (2016): UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 19 TAHUN 2016 TENTANG PERUBAHAN ATAS UNDANG-UNDANG NOMOR 11 TAHUN 2008 TENTANG INFORMASI DAN TRANSAKSI ELEKTRONIK, Sekretariat Negara, Jakarta, accessed 13 Maret 2019melalui situs internet: https://web.kominfo.go.id/sites/default/files/users/4761/UU 19 Tahun 2016.pdf.

[2] Amanah, F., dan Pradipha, F. C. (2019): Deretan Bisnis Vanessa Angel Sebelum Terjerat Kasus Prostitusi Online - Tribunnews.com, accessed 11 Maret 2019, from: http://www.tribunnews.com/section/2019/01/07/deretan-bisnis-vanessa-angel-sebelum-terjerat-kasus-prostitusi-online.

[3] Wibowo, K. S., dan Hantoro, J. (2019): Kasus Prostitusi Online, Polisi Jawa Timur Periksa Selebram - Nasional Tempo.co, , accessed 11 Maret 2019, from situs internet: https://nasional.tempo.co/read/1176078/kasus-prostitusi-online-polisi-jawa-timur-periksa-selebram.

[4] Riyadi, E. (2019): Prostitusi Online Melibatkan Anak-anak di Blitar Diungkap , accessed 11 Maret 2019, from situs internet: https://news.detik.com/berita-jawa-timur/d-4458443/prostitusi-online-melibatkan-anak-anak-di-blitar-diungkap?_ga=2.179923300.1929368686.1552154784-1328920578.1552154784.

[5] Rosadi, S. (2019): Polisi Bongkar Prostitusi Online di Tarakan, Kencani Mahasiswi Bayar Rp 1,75 juta | merdeka.com, , accessed 11 Maret 2019, from : https://www.merdeka.com/peristiwa/polisi-bongkar-prostitusi-online-di-tarakan-kencani-mahasiswi-bayar-rp-175-juta.html.

[6] Nagpal, C., Miller, K., Boecking, B., dan Dubrawski, A. (2018): An Entity Resolution Approach to Isolate Instances of Human Trafficking Online, 77–84. https://doi.org/10.18653/v1/w17-4411

[7] Peled, O., Fire, M., Rokach, L., dan Elovici, Y. (2016): Matching entities across online social networks, Neurocomputing. https://doi.org/10.1016/j.neucom.2016.03.089
[8] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., dan Gummadi, K. P. (2015): On the Reliability of Profile Matching Across Large Online Social Networks.
[9] Papadakis, G., Svirsky, J., Gal, A., dan Palpanas, T. (2016): Comparative analysis of approximate blocking techniques for entity resolution, Proceedings of the VLDB Endowment, 9(9), 684–695. https://doi.org/10.14778/2947618.2947624
[10] Köpcke, H., Thor, A., dan Rahm, E. (2010): Evaluation of entity resolution approaches on real-world match problems, Proceedings of the VLDB Endowment, 3(1–2), 484–493. https://doi.org/10.14778/1920841.1920904
[11] Bilgic, M., Licamele, L., Getoor, L., dan Shneiderman, B. (2006): D-dupe: An interactive tool for entity resolution in social networks, IEEE Symposium on Visual Analytics Science and Technology 2006, VAST 2006 - Proceedings, 43–50. https://doi.org/10.1109/VAST.2006.261429
[12] Weller, K., Bruns, A., Burgess, J., Mahrt, M., dan Puschmann, C. (Ed.) (2014): Twitter and Society, Peter Lang US. https://doi.org/10.3726/978-1-4539-1170-9
[13] Gaffney, D. F., dan Puschmann, C. (2014): Data collection on Twitter, pp.55-67 on K. Weller, A. Bruns, J. Burgess, M. Mahrt, dan C. Puschmann, ed., Twitter and Society, Peter Lang US, accessed 13 Maret 2019 melalui situs internet: https://www.researchgate.net/publication/276974275_Data_collection_on_Twitter.
[14] Guyon, I., dan Elisseeff, A. (2008): An Introduction to Feature Extraction, 1–25 on Feature Extraction, Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_1
[15] Gregg, F., dan Eder, D. (2015): Dedupe, accessed 8 Juli 2019 https://github.com/dedupeio/dedupe.
[16] Grover, A., dan Leskovec, J. (2016): Node2Vec: Scalable Feature Learning for Networks, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). https://doi.org/10.1145/2939672.2939754