



Detecting Alter Ego Accounts using Social Media Mining

Deyana Kusuma Wardani¹, Iwan Syarif², Tessy Badriyah³

^{1,2,3}Department of Informatics and Computer Engineering, Electronic Engineering, Polytechnic Institute of Surabaya

¹deyanakusuma@pasca.student.pens.ac.id, ²iwanarif@pens.ac.id, ³tessy@pens.ac.id

Abstract

Alter ego is a condition of someone who creates a new character with a conscious state. Original character role play is a game to create new imaginary characters that is used as research material for identification alter ego accounts. The negative effects of playing alter ego are stress, depression, and multiple personalities. Current research only focuses on the phenomenon and impacts of a role-playing game. We propose a new method to detect accounts of alter ego players in social media, especially Twitter. We develop an application to analyze the characteristics of alter ego accounts. Psychologists can use this application to discover the characteristics of alter ego accounts that are useful for analyzing personality so that the results can be used to appropriately handle alter ego players. Most user profiles, tweets, and platforms are used to detect account Twitter. This research proposes a new method using bio features as input data. We crawled and collected 565 bios from Twitter for one month. We observe the data to search for unique words and collect them into a classification dictionary. In this research, we use the cosine similarity method because this method is popular for detecting text and has a good performance in many cases. This research could identify alter ego accounts and other types of Twitter accounts. From the detection results of alter ego accounts, it is possible to analyze the characteristics of Twitter accounts. We use a sampling technique that takes 30% of the data as testing data. According to the results of the experiment cosine similarity obtained an accuracy of 0.95.

Keywords: *alter ego; social media mining; twitter; preprocessing data; cosine similarity*

1. Introduction

Alter ego is a condition when someone creates a new character and gets to know that character. Multiple personalities differ from alter egos because they do not recognize other character traits within themselves. The purpose of creating an alter-ego character is to show another personality they cannot express using his real identity. There are examples of alter ego cases on social media [1] - [4]. One example is the original character role play community on social media Twitter.

Original Character Role Play (OCRP) is a type of role play. The definition of role play is an activity for fans to act as public figures on social media [5]. Generally, players create an account on Twitter that contains information about the public figures they play and interact with each other.

How to play an original character role play is different from a role play. Players create a new character starting from the name, background, and personality and using the visualization from public figures. Most players

stated that the characters they created had different personalities from their real personalities, so the original character role play was one of the alter ego cases on social media.

From time to time, if the players frequently use alter ego characters, they may be badly affected in their real life. The negative impact is stress due to differences in life between the real and virtual. A fatal effect is DID (Dissociative Identity Disorder), or what is known as multiple personalities because they are too comfortable with the characters they create. Research using the same case study only uses qualitative methods or questionnaires and focuses on the phenomenon of playing role play [5] - [7]. Recent research [7] is still discussing the phenomenon of role play. Thus, there is no research about original character role play.

This research proposes a new method to detect alter ego accounts on social media. It aims to know the characteristics and differences between alter ego accounts and others. The benefit of this research is to

help psychologists find alter ego accounts on social media, especially Twitter.

Previous studies [8] - [12] identify Twitter accounts using user profiles, tweets, and platforms. With these three features, we can determine the activity difference between real users, bots, and spammers. In this study, we proposed an innovation using a different dataset from previous studies [8] - [12], namely utilizing the bio feature. We realize that the bio feature contains unique keywords that represent a Twitter account and accumulate it. Users describe themselves using the bio feature [13]. Our contribution to this research is identifying Twitter accounts using the bio feature and cosine similarity method. In previous studies [14], [15], the cosine similarity method has high accuracy in detecting text similarity. We use this method to identify text on bio features from Twitter with a classification dictionary.

2. Research Methods

This study proposes a new approach to identifying Twitter accounts. Figure 1 shows the system design. The first step is collecting the user's bio from Twitter. The second step is data preprocessing. The results are used as data for detecting Twitter accounts using the cosine similarity method.

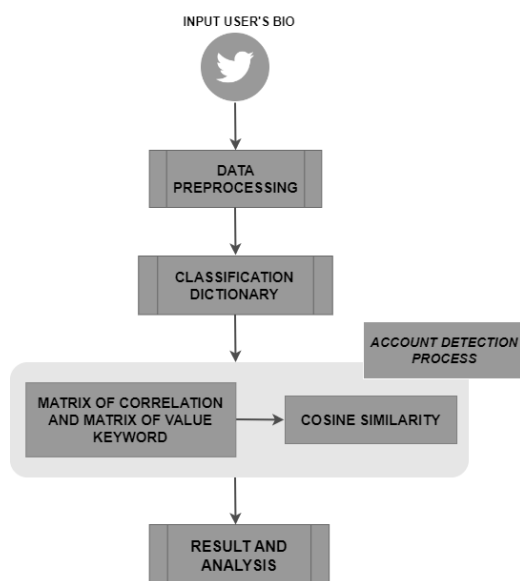


Figure 1. Design System in This Research

2.1 Data Collection

In this study, this data is collected from bio features on Twitter. It consists of document text which contains unique keywords that represent the type of Twitter account. We collected a dataset of 565 bios from various users on Twitter, also used for testing using 30% of the total data.

2.2 Data Preprocessing

The data collected from Twitter should be clean and structured adequately, so data cleaning is needed. The method we use is data preprocessing. Data preprocessing is the initial stage in cleansing data and preparing the data for further processing [16]. The initial stage starts with case folding to tokenizing. Figure 2 shows the steps of data preprocessing in this study.

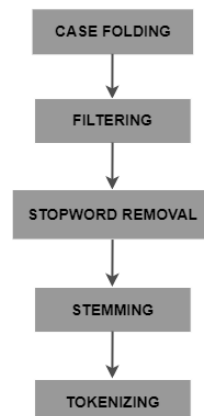


Figure 2. Stages of Data Preprocessing

The first step is case folding. At this stage, all the letters in the dataset are changed to lower case [17], and other than letters, the system will read as delimiters, as shown in Table 1.

Table 1. Case Folding

Sentence		Case Folding
KARAKTER	FIKSI	karakter fiksi namanya gema
Namanya Gema	Aru Sagara,	aru sagara, cita-citanya jadi
cita-citanya jadi	orang kaya	orang kaya yang bersahaja.
yang bersahaja.	Masih betah	masih betah jadi tukang sapu di
jadi tukang sapu di	@80820 &	@80820 & @z_idleidol
@Z_IDLEIDOL		

The second stage is filtering. The data will be filtered by removing punctuation marks, numbers, symbols, URL links, and usernames [16]. An overview of the second stage in this study is shown in Table 2.

Table 2. Filtering

Sentence		Filtering
KARAKTER	FIKSI	karakter fiksi namanya gema
Namanya Gema	Aru Sagara,	aru sagara cita citanya jadi
cita-citanya jadi	orang kaya	orang kaya yang bersahaja
yang bersahaja.	Masih betah	masih betah jadi tukang sapu di
jadi tukang sapu di	@80820 &	
@Z_IDLEIDOL		

The third stage is stopword removal. A group of common words often appear in a sentence but cannot give the sentence's meaning —a group of common words that become noise that must be eliminated. We use the Sastrawi library [18] to find common words. At this stage, the system removes common words in the sentences shown in Table 3.

Table 3. Stopword Removal

Sentence	Stopword Removal
KARAKTER FIKSI Namanya Gema Aru Sagara, cita-citanya jadi orang kaya yang bersahaja. Masih betah jadi tukang sapu di @80820 & @Z_IDLEIDOL	karakter fiksi namanya gema aru sagara cita citanya jadi orang kaya bersahaja betah jadi tukang sapu

After the stopwords removal stage, the fourth stage is stemming. Stemming is returning words to their essential form [16], then removing sentence affixes. We also use the Sastrawi library to get a list of essential Indonesian words. An overview of the stemming process steps is shown in Table 4.

Table 4. Stemming

Sentence	Stemming
KARAKTER FIKSI Namanya Gema Aru Sagara, cita-citanya jadi orang kaya yang bersahaja. Masih betah jadi tukang sapu di @80820 & @Z_IDLEIDOL	karakter fiksi nama gema aru sagara cita cita jadi orang kaya sahaja betah jadi tukang sapu

The final stage in data preprocessing is tokenizing. The process at this stage is cutting the sentence into a word [17] and turning it into a token. This process aims to get tokens that become entities in preparing the document matrix. Table 5 is an overview of the tokenizing stages.

Table 5. Tokenizing

Sentence	Tokenizing
KARAKTER FIKSI Namanya Gema Aru Sagara, cita-citanya jadi orang kaya yang bersahaja. Masih betah jadi tukang sapu di @80820 & @Z_IDLEIDOL	[['karakter', 'fiksi', 'nama', 'gema', 'aru', 'sagara', 'cita', 'cita', 'jadi', 'orang', 'kaya', 'sahaja', 'betah', 'jadi', 'tukang', 'sapu']]

2.3. Classification Dictionary

Twitter social media users have various goals in creating a Twitter account. From our observations, we can find the purpose of creating a Twitter account in the bio feature. Unique keywords represent the purpose of creating a Twitter account. In this study, we collected these words and formed them into a Twitter account classification dictionary. The steps for making a Twitter account classification dictionary in this study are shown in Figure 3.

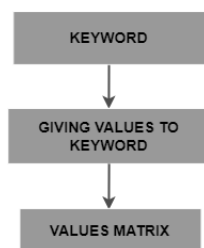


Figure 3. Stages of Create Classification Dictionary

A collection of words that can represent a Twitter account. In this study, we call them keywords. Because there are many purposes for creating a Twitter account, in this study, we differentiate into 5 (five) types of Twitter accounts: Business Account is a type of Twitter account that users use to promote the products they sell; Role Play Account is a Twitter account that users use to follow or act like a public figure [5]; Original Character Role Play Account is the same type of Twitter account as role play, but users can create a new character by using the appearance of a public figure; Fan Account or fans is a type of Twitter account used by fans to receive information about the activities of public figures they like; and Promotor Giveaway Account is a Twitter account that broadcasts advertisements or promotes products [10] by giving away a prize.

Those type of Twitter accounts have unique keywords that match the purpose of the Twitter account created. Unique keywords from the five types of Twitter accounts are collected into a classification dictionary with total keywords are 14. From our observations, if accounts do not have a unique keyword, the system would labelled them as personal accounts. Table 6 shows a collection of keywords used as a classification dictionary in this study.

Table 6. Classification Dictionary

Business	Role Play	Original Character Role Play	Fans	Promotor Giveaway
business	roleplayer	fictional	fa	social
ba	rp	illusoire	fans	promoter
testi	from	lakon	fangirl	crypto
proofs	known	ocrp	fanboy	promotions
since	parody	fiksi	mulfand	sponsor
payment	portrayed	fiksional	dedicated	dm
via	singer	oc	kdrama	media
open	roleplay	character	bias	project
selling	rapper	illusive	mostly	winner
order	faceclaim	ilusi	stan	marketing
testimoni	leader	karakter	fangirling	inquiries
trusted	maknae	nirnyata	enthusiast	nft
bisnis	unreal	fiction	addicted	grow
respon	visual	written	enthu	investor

Each unique keyword in the classification dictionary is given a different value using the first equation, namely the average formula as seen in formula 1 [19]. First, each keyword is calculated by the number of occurrence frequencies in each dataset. Then all frequency values in one category are summed up. Then the sum of the frequency values for each keyword is divided by the total frequency according to the category.

$$\bar{x} = \frac{\sum x}{n} \quad (1)$$

$\bar{x}$ is an average value, x is frequency value of each class, and n is total frequency value of all classes in one category.

After each keyword is assigned a value, the next step is manually creating a value matrix. The matrix row contains keywords, and the column contains the type of Twitter account. The content of the matrix is the value of each keyword and a value of 0 if the keyword does not exist in the Twitter account type. Table 7 shows an overview of the value matrix made in this study.

Table 7. Overview of Value Matrix

Twitter Account Type	Keyword				
	testi	parody	lakon	fan	social
Business	0.1	0	0	0	0
Role Play	0	0.08	0	0	0
Original	0	0	0.05	0	0
Character Role Play					
Fans	0	0	0	0.16	0
Promotor	0	0	0	0	0.07
Giveaway					

The method that we use is cosine similarity. Cosine similarity is a method for calculating the similarity between two vectors by finding the cosine value of the angle between the two vectors. The first step is to make a correlation matrix. The technique used to create a correlation matrix is if the keywords contained in the data from the preprocessing results are given a value of 1 and otherwise provided a value of 0. Table 8 is an illustration of the correlation matrix that was built.

Table 8. Overview of Correlation Matrix

Twitter Account Type	Keyword				
	testi	parody	lakon	fan	social
Business	1	0	0	0	0
Role Play	0	0	0	0	0
Original	0	0	1	0	0
Character Role Play					
Fans	0	0	0	0	0
Promotor	0	0	0	0	0
Giveaway					

The value matrix created in the previous process (Table 7) is merged with the correlation matrix formed (Table 8). Table 9 shows an overview of combining the value and correlation matrices.

Table 9. Overview of Result Combining Value Matrix and Correlation Matrix

Twitter Account Type	Keyword				
	testi	parody	lakon	fan	social
Business	0.1	0	0	0	0
Role Play	0	0.08	0	0	0
Original	0	0	0.05	0	0
Character Role Play					
Fans	0	0	0	0.16	0
Promotor	0	0	0	0	0.07
Giveaway					
Input	1	0	1	0	0

Formula 2 is calculated, namely the cosine similarity formula [20] between the dataset and the built classification dictionary. For each type of Twitter account, there is a similarity value from the results of the cosine similarity calculation. Then, look for a bigger value to label the Twitter account. However, if the similarity value is NAN, it is labeled a personal account.

$$similarity = \frac{\bar{x} \cdot \bar{y}}{||\bar{x}|| \cdot ||\bar{y}||} \quad (2)$$

Similarity is a level of similarity of keywords and data, x is a data, and n as a keyword.

3. Results and Discussions

3.1 Results and Analysis

There are five different types of Twitter accounts from the method that we used: business accounts, role play accounts, original character role play accounts, fans accounts, and promotor giveaway accounts. However, if the account is not entered into one of the five types of Twitter accounts, our system will detect it as a personal account. Table 10 shows some of the results and marks the keywords we collected in the classification dictionary.

Table 10. Some of The Detection Results Twitter Account

Results of Detection	Bio
Business	Business Account ? Menyediakan keperluan twitter terutama verif. DM atau WA untuk info lebih lanjut. (Testi : http://bit.ly/NikiProof)
Business	Business ~ selling account, premium apps, open convert and other stuff aka ba serabutan
Business	business realm ? here's providing you ?? second account @naiaaraagiza, open : 07.00 - 23.00 jangan takut order ^^ : akun ketiga @makimacsaw
Role Play	PORTRAYED ?? Born in 1993 under the name Min Yoongi from Daegu. My crew's been here to witness how I spit fire through my rap.
Role Play	# ROLEPLAYER ? ? South Korean rapper and singer currently under MYSTIC Story. She is a member of the girl group Billie. Moon Sua (???)
Role Play	RP . Citation on every hue that chronicle about elysium, she's in the cynosure because of pure demeanor and ephemeral like a fairy, Ahn Yujin.
Original Character Role Play	ILLUSIVE . In the midst of it's chaos, her conception of the world remains serene. She's a xyst of lights and blue, the progeny of the clouds hue.
Original Character Role Play	FIKSIONAL hasil guratan LALUNA. Ia yang mencoba memahami tempatnya berlabuh, kini terdampar di keruhnya satu sisi dunia. Works related: @MISDEMEANOR
Original Character Role Play	lakon fiksi , si penyesak bumi sejak warsa dua ribu. Bunda beri ia asma Nanggala Eshan Pradipta, dan kerap dipanggil Eshan. Namun bukan teman Fizi, atau harimau.

Results of Detection	Bio
Fans	Fan Account khusus gg RV, SNSD, aëspa, WJSN, Rocket Punch, LiuYuxin, Choi Yujin & Shen Xiaoting, IVE DRIPPIN for bg — Read my pinned??
Fans	24/7 fan account for @JYPELOUD @TNX_Official but sometimes for @NMIXX_official @XH_official @Stray_Kids and others. cek carrd before you follow
Fans	?? Fangirl ? akun suka suka ?? menulis dan membaca ?? sayang kamu banyak banyak ? Halu terus dan melampauinya
Promotor Giveaway	Finding Crypto & NFT GEMS???? • Promotor/Shiller • DM me for promotion?? • Check #WonByRich & LIKES on profile for vouch?? • Back-up: @DutchCryptoRich • #DYOR
Promotor Giveaway	crypto shiller • Social media & crypto promoter • DM for promotion ?? Proof of winners: #Gemwinners #BTC #TRX #BNB #ETH #NFT #solana #gemhostfamily #gemeut
Promotor Giveaway	social media & crypto project promoter ?? #cutetesty vouch @winnerchita #nfts crypto paypal gcash ewallet bank reborn worldwide ??
Personal	don't tell somethin' that you don't know exactly, coz none will ask you to tell somethin' that you don't know exactly = JANGAN SOK TAHU????
Personal	jadilah bunga d tepi jurang indah untuk di pandang tapi sukar untuk di dapatkan.
Personal	retweet bukan berarti suka, follow akun beneran, bukan times atau relawan,hanya simpatisan

Table 10 shows differences in unique keywords, it shows with marked blue for the five types of Twitter accounts. The Twitter account detected as a business account has keywords: business, *testi*, BA, open, and order. These five keywords relate to everything about the business, starting from words that have the meaning of acknowledgment [21] from consumers of products (words *testi*), information about business accounts, shop operating hours, and products [21].

The Twitter accounts detected as role play accounts (Table 10) have keywords: portrayed, from, roleplayer, rapper, singer, and RP. The six keywords explain the purpose of playing the role of a public figure with brief information related to their work, country, and origin of them.

The original character role play account is almost identical to the role play account type. However, players from original character role plays create new characters and only borrow visualizations from public figures, as shown in Table 10. Twitter accounts detected as original character role play account types have keywords: *illusive*, *fiksional*, *lakon*, and *fiksi*.

The four keywords can be interpreted as the purpose of creating a Twitter account. Everything on the Twitter account is just the players' imagination, from the characters to the activities carried out on that account. Illusive keywords are interpreted as characters that

players make, *fiksional* keyword is interpreted that the characters on the Twitter account are not real, *lakon* keyword has the meaning as the main role [21], and *fiksi* keyword is almost the same as *fiksional* keyword.

The next type of Twitter account is a fan account. In table 10, the keywords used are fan and fangirl. Both refer to the meaning of fans. The keyword “fan” refers to supporter, and “fangirl” is a combination of two words: fan (supporter) and girl (woman).

The fifth type of Twitter account in Table 10 is the promoter giveaway account; this account has the keywords crypto, NFT, DM, social, media, promoter, and project. These include a brief explanation of the products that the user can promote and the kind of work that the user does as a product promoter by giving gifts to attract consumers, and they can contact them if they want to collaborate by message via Direct Message on Twitter.

In table 10, 3 (three) data are detected as types of personal accounts. The difference between these three data and others is there no keywords representing five types of Twitter account. Most personal accounts contain a quote and a short explanation about the user. So if a Twitter account doesn't have keywords in the classification dictionary, the system will classify that Twitter accounts as a personal account.

We can see the difference between the six types of Twitter accounts detected by the system. First, the business account, their bio is about their business. In the role play Twitter account, the user's bio related to how to play the role play and information about the public figure they played. Fan Twitter accounts have information associated with a public figure they like. Then, the promoter giveaway account has a bio that promotes a product to attract consumers by giving a prize. Last, a personal account does not have unique keywords like the others because the goals of creating personal accounts are only used to interact with their friends on Twitter.

The bio of the original character role play account explains that the characters or people in the account are only imagination and made up by the player. So we can see the difference between this type of Twitter account and the others. If we meet them in real life, we only see the visualizations of the public figure used by these characters.

3.2 Performance Measurement

Accuracy calculations are used to see the performance of the method. The amount of data we collected was 565 data, and we managed to detect the type of Twitter account from each data. Table 11 shows the total number of detected Twitter accounts for each Twitter account type. Several accounts may be misinterpreted because some unique keywords are not in the

dictionary, and there are two keywords with two different types of accounts. In comparison, the weight of the keywords for other types is greater than for those account types.

Table 11. Total Twitter Accounts Successfully Detected

Twitter Account Type	Number of Accounts
Business	91
Role Play	105
Original Character Role Play	94
Fans	90
Promotor Giveaway	82
Personal	104

Accuracy with sampling technique that used 30% of the total data, so it uses 169 data for accurate calculations. We collected 28 data from each type of Twitter account, except for the personal account type; we took 29 data, a total of 169 datasets. Then we labeled the 169 data types of Twitter accounts manually. We compare the results of this manual label with the results of the system. Next, we converted the type of Twitter account into numeric, namely business accounts were changed to 1 (one), role play accounts were changed to 2 (two), original character role play accounts were changed to 3 (three), fan accounts were changed to 4 (four), promotor giveaway accounts changed to 5 (five), and personal accounts changed to 6 (six).

In calculating the method's accuracy, we utilize a library from Python: scikit-learn. The scikit-learn library is an open library from Python that can be used for machine learning, with many algorithms for machine learning [22]. We utilize two objects from scikit-learn, namely the confusion_matrix, to evaluate accuracy by calculating the confusion matrix. The results of the confusion matrix in this study are shown in Figure 4.

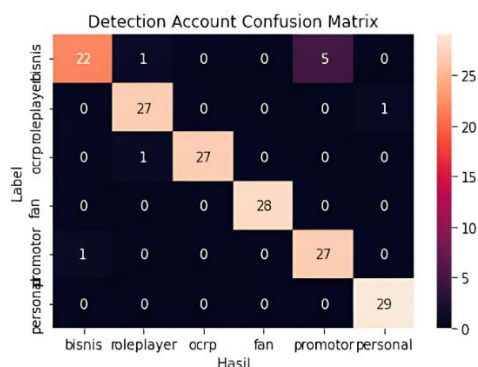


Figure 4. Confusion Matrix of Twitter Account Type Detection

Prediction class is a detection result; Actual class is a label, TP (True Positive) value is the value in the brightly colored box; TN (True Negative) value are shown in 4 (four) values after the TP value of each class. The example in business class has a TN Value of 27, 0, 1, 27.; FP Value (False Positive) is the values in the horizontal box in the actual class; for example, the business label has an FP Value of 1, 0, 0, 5, 0; FN (False Negative) value is the values in the vertical box in the

prediction class; for example, the business results have FN values of 0, 1, 0, 0, 0.

The second object we use from the scikit-learn library is classification_report to see the method's accuracy so that the accuracy value of the method used is 0.95. Table 12 shows the classification report of this research method, which contains the precision, recall, and f1-score for the six types of Twitter accounts and the accuracy of the method used.

Table 12. The classification report

Classification Report :				
	precision	recall	f1-score	support
1	0.96	0.79	0.86	28
2	0.93	0.96	0.95	28
3	1.00	0.96	0.98	28
4	1.00	1.00	1.00	28
5	0.84	0.96	0.90	28
6	0.97	1.00	0.98	29
accuracy			0.95	169
macro	0.95	0.95	0.95	169
avg				
weighted	0.95	0.95	0.95	169
avg				

4. Conclusion

This research proposes a method for detecting alter ego accounts using data from the bio feature on Twitter. In the bio, unique keywords are collected into the classification dictionary. We give value to each unique keyword using the average formula. The method used to detect Twitter accounts is cosine similarity, result can detected alter ego accounts and other types of Twitter accounts: business accounts, role play accounts, fan accounts, and promotor giveaway accounts. The system will label it as personal account if the bio on that account does not have unique keywords.

This research analyzes the difference between alter ego accounts and five other types of Twitter accounts (business accounts, roleplay accounts, fan accounts, promotor giveaway accounts, and personal accounts). The bio of the alter ego account explains that the person or character in the account isn't real and was created by a player, so we can't meet that person in real life. But we can meet in real life for a product or person from the other five types of Twitter accounts. The accuracy of this study reached 0.95. The results of this research can be used to identify personalities of the alter ego character and help psychologists to treat alter ego players appropriately.

References

- [1] A. R. Putri, "Manajemen impresi pengguna akun alter ego di twitter pada akun fanbase @AlterBase18Plus," pp. 1–26, 2021.
- [2] M. Saifulloh and A. Ernanda, "Manajemen Privasi Komunikasi Pada Remaja Pengguna Akun Alter Ego Di Twitter," *WACANA, J. Ilm. Ilmu Komun.*, vol. 17, no. 2, p. 235, 2018, doi: 10.32509/wacana.v17i2.652.
- [3] S. Maria and T. Daina, "Dramaturgi Dalam Alter Account Di Twitter Dramaturgy in Alter Account on Twitter," no. 2, pp.

- 140–148, 2022.
- [4] P. A. J. Retasari Dewi, “Dramaturgi Dalam Media Sosial : Second Account di Instagram Sebagai Alter Ego,” *J. Ilmu Komun.*, vol. 8, no. 3, pp. 340–347, 2018.
- [5] E. K. I. P. Nuraini, “Peran roleplayer dalam membentuk identitas virtual di jejaring sosial line,” *Univ. Muhammadiyah Surakarta*, 2021.
- [6] D. Hapsari, “Korean roleplayer dan dampaknya terhadap kepribadian di dunia nyata (studi kasus pada remaja),” *Univ. Sebel. Maret*, 2019.
- [7] A. P. Pratiwi, “Identitas virtual pada roleplayer di twitter,” vol. 1, pp. 215–227, 2023.
- [8] I. Inuwa-Dutse, B. S. Bello, and I. Korkontzelos, “Lexical analysis of automated accounts on twitter,” *Proc. Int. Conf. WWW/Internet 2018 Appl. Comput. 2018*, pp. 75–82, 2018.
- [9] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, “Twitter spam account detection based on clustering and classification methods,” *J. Supercomput.*, vol. 76, no. 7, pp. 4802–4837, 2020, doi: 10.1007/s11227-018-2641-x.
- [10] K. E. Daouadi, R. Z. Rebaï, and I. Amous, “Organization vs. individual: Twitter user classification,” *CEUR Workshop Proc.*, vol. 2279, pp. 1–8, 2018.
- [11] M. Haidermota, “Classifying Twitter User As a Bot or Not and Comparing Different Classification Algorithms,” *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 3, pp. 29–33, 2018, doi: 10.26483/ijarcs.v9i3.5949.
- [12] O. Loyola-Gonzalez, R. Monroy, J. Rodriguez, A. Lopez-Cuevas, and J. I. Mata-Sanchez, “Contrast Pattern-Based Classification for Bot Detection on Twitter,” *IEEE Access*, vol. 7, pp. 45800–45817, 2019, doi: 10.1109/ACCESS.2019.2904220.
- [13] S. R. I. Rezeki, “Penggunaan sosial media twitter dalam komunikasi organisasi (studi kasus pemerintah provinsi DKI Jakarta dalam penanganan covid-19),” *J. Islam. Law Stud.*, vol. 04, no. 02, pp. 63–78, 2020.
- [14] D. N. Lindang, A. Y. Muniar, A. Halid, and A. Amiruddin, “Sistem Penentuan Kemiripan Antar Skripsi Menggunakan Metode Cosine Similarity Pada Perpustakaan,” pp. 321–324, 2022.
- [15] M. M. N-gram, A. Pradana, H. Sujaini, and H. Sasty, “Sistem Rekomendasi Artikel sebagai Acuan Studi Literatur Article Recommendation System as a Reference for Literature Studies Using the N-Gram Method,” vol. 01, no. 1, pp. 69–84, 2023, doi: 10.26418/juristi.v1i1.61170.
- [16] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.
- [17] H. Ma’rifah, A. P. Wibawa, and M. I. Akbar, “Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing,” *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 2, no. 2, p. 70, 2020, doi: 10.30872/jsakti.v2i2.2681.
- [18] Hanif Amal Robbani, “Sastrawi Library,” *Python Package Index*, 2023. <https://pypi.org/project/Sastrawi/>
- [19] R. Kartika, “Pengaruh Model Problem Centered Learning terhadap Kemampuan Menulis Puisi Kelas X SMK PAB 3 Medan Eetate,” *J. Pendidik. Bhs. dan Sastra Indones.*, vol. 3, no. 1, pp. 60–67, 2018, [Online]. Available: <https://journal.unnes.ac.id/sju/index.php/jpbsi/article/view/24018>
- [20] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, “Word2vec model analysis for semantic similarities in English words,” *Procedia Comput. Sci.*, vol. 157, pp. 160–167, 2019, doi: 10.1016/j.procs.2019.08.153.
- [21] P. B. D. P. Nasional, *Kamus Bahasa Indonesia*. Pusat Bahasa, 2008. [Online]. Available: <https://perpus.unimus.ac.id/wp-content/uploads/2012/05/Kamus-Besar-Bahasa-Indonesia.pdf>
- [22] G. Hackeling, *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd., 2014. [Online]. Available: <https://www.smallake.kr/wp-content/uploads/2017/03/Mastering-Machine-Learning-with-scikit-learn.pdf>