Journal of Digital Ecosystem for Natural Sustainability (JoDENS)
Vol.1, No.2, December 2021, pp. 78~83
e-ISSN: 2798-6179
DOI https://doi.org/10.63643/jodens.v1i2.43

Implementasi Text Mining untuk Mengetahui Kata Abreviasi dalam Percakapan Media Sosial

Geovaldo Reggie Yunarfi [1*], Ricky Simdy [2], Jackson [3]
[1,2,3] Teknik Perangkat Lunak, Universitas Universal
***** *Corresponding author E-mail: Geovaldo09@uvers.ac.id*

Article Info

Article history:

Received 30-11-2021
Revised 05-12-2021
Accepted 06-12-2021

Keyword:

Abreviation, Graph, Media Social, Preprocessing, Text Mining.

ABSTRACT

Social media are technology that allow sharing or exchange of information, ideas, interests, etc., via virtual communities and networks. Social media is often use for chatting either private chat or commenting on posts, and most frequently used application in Indonesia such as Facebook, WhatsApp, Instagram, Twitter, and so on. So far, peoples are typing using abbreviations as habit instead using full word and thus cause misunderstanding for others. Descriptive qualitative method was used to collect data. Text mining is a data science technique which mine data in the form of text and look for words that can represent or analyze the content of the document, and by using network of terms that can build a graph for knowing interactions between words in document. In this study, with implementing data preprocessing in text mining process is expected to reduce word or text that are not necessary, which made analyze easier to find out abbreviations within captions or comments.

I. PENDAHULUAN

Media sosial merupakan sebuah media daring yang digunakan satu sama lain dimana para pengguna bisa dengan mudah berpartisipasi, berinteraksi, berbagi dan berbagai bentuk aktifitas dalam dunia virtual tanpa dibatasi oleh ruang dan waktu. Sebuah kelompok aplikasi berbasis internet yang dibangun di atas dasar ideologi dan teknologi Web 2.0 dan memungkinkan penciptaan dan pertukaran *user-generated content* [1].

Penggunaan teknologi media sosial sudah menjadi aktivitas kehidupan sehari-hari serta telah mengubah gaya hidup bahkan pola pikir. Dampak yang diperoleh dari keberadaan teknologi media sosial adalah nilai-nilai budaya masyarakat yang semakin memudar, mulai dari gaya hidup, pola pikir bahkan cara menyampaikan pesan. Dalam percakapan media sosial, seringkali ditemukan kata-kata yang tidak formal maupun singkatan dari sebuah kata (abreviasi) sehingga kata tersebut menyebabkan kesalahpahaman terhadap hal yang sebenarnya ingin disampaikan. Abreviasi merupakan proses pemotongan satu atau beberapa bagian leksem atau kombinasi leksem sehinga menjadi bentuk baru yang berstatus kata [2].

Text mining merupakan proses mengeksplorasi dan menganalisis data dalam bentuk teks dengan tujuan mengidentifikasi konsep, pola, kata kunci dan bentuk atribut lainnya. *Text mining* merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [3]. Tahapan *preprocessing* dalam proses *text mining* secara umum yaitu, tahap pertama *tokenizing*, tahap kedua *filtering*, tahap ketiga *stemming*, tahap keempat *tagging*, dan tahap terakhir *analyzing* [4].

Penelitian ini diambil karena pernah terjadinya kesalahpahaman dalam percakapan media sosial pada komentar sebuah postingan dan menjadi popular, dimana seseorang menggunakan kata abreviasi "*LOL*" yang artinya tertawa terbahak-bahak ataupun "*Laughing out Loud*" dalam bahasa inggris dan pihak pembaca berfikir bahwa "*LOL*" merupakan kata "tolol" yang artinya bodoh. Penggunaan kata abreviasi tidak hanya menyebabkan kesalahpahaman bagi pembaca namun juga membuat percakapan menjadi canggung sehingga percakapan yang dibahas berbeda dengan pemikiran pembaca.

Oleh karena itu, penelitian ini akan mencoba menggunakan *text mining*, supaya dapat mengetahui bahwa abreviasi apa yang sering digunakan dalam percakapan media sosial dan mempelajari arti dari belakang abreviasi tersebut.

II. METODE

A. Pengumpulan Data

Metode pengumpulan data yang dilakukan dalam penelitian ini adalah metodologi deskriptif kualitatif. Metodologi kualitatif adalah pendekatan holistik yang melibatkan penemuan, metodologi kualitatif juga digambarkan sebagai model yang berlangsung terjadi dalam pengaturan alami yang memungkinkan peneliti untuk mengembangkan tingkat detail dari keterlibatan yang tinggi dalam pengalaman nyata [5].

Data dalam penelitian ini diambil secara acak dari *caption* dan komentar yang berasal dari hastag seperti #dramatwitter, #sinetronindonesia, #oposisikebalhukum, dan lainnya dalam media sosial *Facebook*, *Twitter*, dan *Instagram*. Data yang dikumpulkan merupakan data yang mengandung kata abreviasi yang disajikan dalam bentuk akronim, singkatan, dan penggalan.

Tabel 1. Data Observasi

Data	Media Sosial		
	Facebook	***Instagram***	***Twitter***
1	Naik angkot sm temen smp se genk.. pulang sekolah. temen gw, cwek 1 udh dr tadi pengen pipis di tahan2 terus	ampun bang itu cuma becanda doang kok, Sering bnget d ig d pnggil dek, bocil klw komen bagusan dkit	ini loh contoh awal org pd muak mau lapor polisi. lama2 kemuakan itu menjadi kebencian jika sudah benci
2	ini dri pengalaman gua dan ini kisah gua. jdi klo cwo nanya knp cwe gprnh abisin mknan klo lgi sm cwonya	Klw emng info diatas valid, knp perusahaan sebesar *Facebook* mengumumkan nama brand baru tanpa mengecek ke lembaga terkait	Di masa pandemi kaya gini emang semuanya serba susah, tak heran banyak yg menjadi buzer.
n	…	…	…

B. PreProcessing

Teknik *preprocessing* terdiri dari:

1) *Data cleaning*, diterapkan untuk menghilangkan *noise* dan memperbaiki inkonsistensi dalam data.
2) *Data integration*, menggabungkan data dari beberapa sumber ke dalam penyimpanan data yang koheren.
3) *Data reduction*, mengurangi ukuran data contohnya menggabungkan fitur yang berlebihan atau pengelompokkan.
4) *Data transformation*, dapat diterapkan dimana data berada diskalakan supaya dalam kisaran yang lebih kecil seperti 0.0 hingga 1.0 [6].

Metode *preprocessing* dalam penelitian ini digunakan untuk menghilangkan kata atau kalimat yang tidak berarti dalam penelitian seperti mengandung bahasa lain selain bahasa Indonesia, pengunaan karakter spesial dan lain sebagainya. Penggunaan teknik ini diketahui mampu mempermudah proses kerja *text mining* untuk menambang kata abreviasi dalam data. Tahapan *preprocessing* memiliki beberapa proses, yaitu *case folding*, *stopwords removing*, *tokenizing*, dan *stemming*, selanjutnya data yang sudah mengalami preprocessing akan diubah menjadi bentuk numerik dengan tahap *term weighting* [7].

C. Clustering

Clustering merupakan teknik untuk mengelompokkan data-data yang mempunyai karakteristik yang umum, tujuan utama dari metode *clustering* adalah mengelompokkan sejumlah data ke dalam suatu kelompok (*cluster*) sehingga setiap cluster akan berisi data yang mempunyai kesamaan signifikan [8].

Dalam penelitian ini digunakan *cluster dendrogram*, dimana *graph clustering* tersebut merupakan diagram berbentuk pohon, *dendrogram* mengelompokkan 2 data menjadi 1 *cluster* berdasarkan jumlah kuadrat dari gabungan data tersebut. Jarak pembagian atau pengabungan data disebut juga ketinggian (*height*).

D. Social Network Analysis

Social Network Analysis (SNA) merupakan struktur sosial melalui penggunaan jaringan dan teori *graph.* SNA mencirikan struktur jaringan dalam hal *nodes* dan *edges* yang memiliki hubungan atau interaksi dengan satu sama lain [9]. *Nodes* merupakan data yang ditampungi sedangkan *edges* merupakan garis hubungan antar relasi.

E. Text Mining

Text mining merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [3].

Algoritma umum yang digunakan setelah proses tahapan *text mining* yaitu *Term-Frequency Inverse Document Frequency* (TF-IDF). Suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen [4], [10].
Rumus TF-IDF:

$$IDF_t = \log\left(D/df\right) \quad (1)$$

$$W_{d,t} = tf_{d,t} * IDF_t \quad (2)$$

Keterangan (1):
IDF = Inversed Document Frequency
t = kata ke-*t* dari kata kunci
D = total dokumen
df = banyak dokumen yang mengandung kata yang dicari
Keterangan (2):

Pada persamaan (1) digunakan untuk mencari nilai *IDF*
d = dokumen ke-*d*
t = kata ke-t dari kata kunci
W = bobot dokumen ke-*d* terhadap kata ke-*t*
tf = banyak kata yang dicari pada sebuah dokumen

Berikut penerapan tahapan *preprocessing* dalam mengumpulkan abreviasi pada data yang telah diperoleh.

1) Tahap Tokenizing

Tahap *tokenizing* merupakan proses untuk membagi atau memecahkan teks yang dapat berupa kalimat, paragraph atau dokumen menjadi token atau bagian tertentu seperti kumpulan data, dengan cara menghilangkan tanda baca atau mengubah huruf kapital menjadi huruf kecil (*lower case*) [10].

Tabel 2. Contoh Penerapan *Tokenizing*

Data	Hasil Token
Dia mempnyai permen yang sangat bnyk dan juga permainan unik	dia mempnyai permen yang sangat bnyk dan juga permainan unik

2) Tahap Filtering

Tahap *Filtering* merupakan tahap pengambilan kata-kata penting dari hasil token, tahap *filtering* dapat menggunakan 2 algoritma yakni: (1) *stoplist*, proses ini membuang kata yang tidak penting, dan (2) *wordlist*, proses menyimpan kata penting atau kata yang perlu digunakan.

Tabel 3. Contoh Penerapan *Filtering*

Data	Hasil Filter
dia mempnyai permen yang sangat bnyk dan juga permainan unik	mempnyai permen sangat bnyk permainan unik

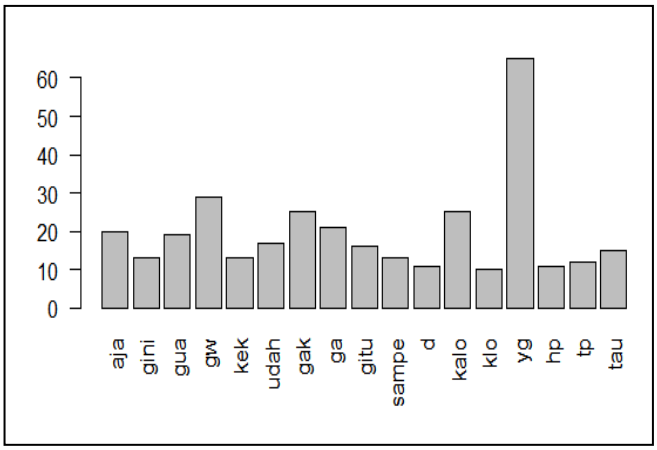
3) Tahap Analyzing

Tahap ini menggunakan algoritma SNA dengan tata letak *Fruchterman-Reingold* untuk memvisualisasi hubungan antar relasi terhadap kata abreviasi yang telah ditambang. Algoritma *Fruchterman-Reingold* merupakan tata letak yang diarahkan oleh gaya yang menggunakan analogi pegas fisik sebagai *edge* yang menarik *node* terhubung satu sama lain dan gaya tolak bersaing yang mendorong semua *node* menjauh dari satu sama lain, baik *node* tersebut terhubung ataupun tidak [11].

III. HASIL DAN PEMBAHASAN

Dalam penelitian ini, data diambil secara acak dalam *caption* maupun komentar pada media sosial *Facebook*, *Instagram*, dan *Twitter* untuk menambang kata abreviasi. Abreviasi ini sering digunakan sebagai penyingkatan sebuah kata dasar, sehingga membuat kesalahpahaman bagi pembaca.

A. *Bar Graph*



Gambar 1. *Bar Graph* Hasil *Text Mining*

Pada *bar graph* diatas terdapat berbagai abreviasi yang sering digunakan dalam percakapan media sosial antara lain: “yg”, “gw”, “gak”, “kalo”, “aja" dan lainnya. Dapat diperhatikan dalam bar graph tersebut, bahwa penggunaan kata abreviasi yang paling banyak digunakan dalam percakapan baik media sosial *Instagram*, *Facebook*, maupun *Twitter* menggunakan kata abreviasi “yg” dengan frekuensi pemakaian 65 dibanding menggunakan kata dasar “yang” sebagai kebiasaan sehari-hari. Kemudian bisa diketahui juga bahwa penggunaan abreviasi paling sedikit yaitu “klo” dengan frekuensi 10, karena kata abreviasi “klo” dan “kalo” mempunyai kata dasar yang sama yaitu “kalau” sehingga kata dasar tersebut terpecah menjadi berbagai kata abreviasi, begitu juga kasus terhadap abreviasi “gak” dan “ga” yang mempunyai arti “tidak”. Adapun penggunaan abreviasi “d” dengan frekuensi 11 yang merupakan singkatan dari kata “di”, hal ini digunakan untuk mempersingkat pengetikan dalam percakapan.

Dalam *bar graph*, terdapat kata “gua” dimana kata tersebut dapat berupa sebuah pengertian dari “sebuah lubang alami di tanah yang cukup besar dan dalam” ataupun mengartikan sebagai kata “saya”. Namun pengunaan kata tersebut telah menjadi kebiasaan sehari-hari sebagai abreviasi dari kata “saya”.

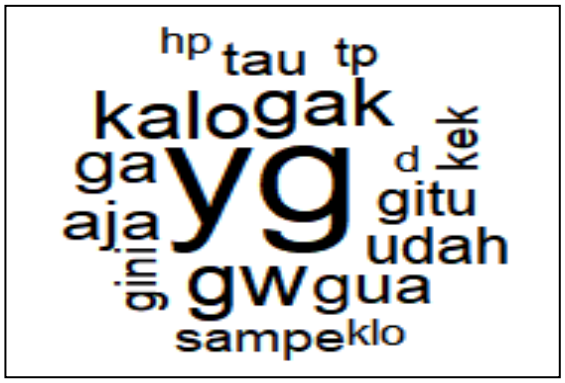
Tabel 4. Kata dasar pada abreviasi

Abreviasi	Kata dasar
aja	saja
gini	begini
gua gw	saya
kek	kayak
udah	sudah
gak ga	tidak
gitu	begitu
sampe	sampai
d	di
kalo klo	kalau
yg	yang
tp	tapi
hp	*handphone*
tau	tahu

a) abreviasi diambil dari hasil tambang dengan frekuensi diatas 9

B. *Word Cloud*

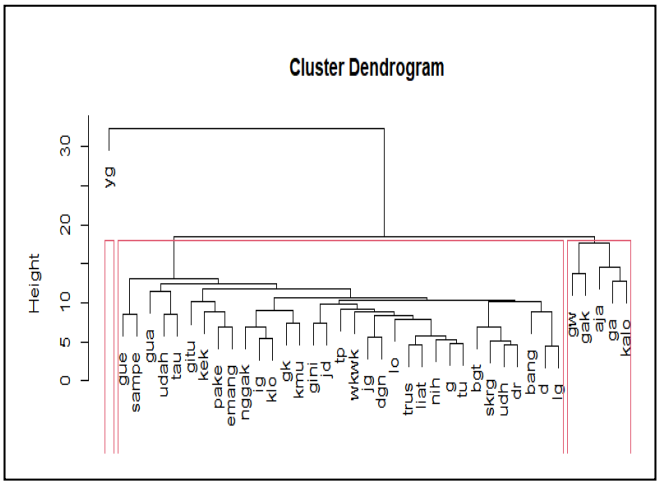
Graph word cloud mengumpulkan data-data dari hasil filter abreviasi yang akan digunakan dan menampilkan data dengan berbagai ukuran, mulai dari yang paling besar hingga yang paling kecil. Selain ukuran, posisinya juga memiliki beberapa variasi seperti tata letak yang secara horizontal, vertical, dan ada juga yang miring.



Gambar 2. *Wordcloud* Hasil *Text Mining*

Pada Gambar 2, dari data yang telah dikumpulkan dijadikan dalam bentuk *word cloud*, yang terdapat beberapa kata abreviasi dari hasil tambang. Dari *graph* tersebut diketahui penggunaan abreviasi "yg" merupakan kata abreviasi yang paling besar di antara kata-kata lainnya, kemudian disusul oleh kata "gw", "kalo", "gak", dan seterusnya. Dapat disimpulkan bahwa dalam percakapan media sosial *Instagram*, *Facebook*, dan *Twitter* banyak menggunakan abreviasi "yg" sebagai kata hubung yang menjadi kebiasaan sehari-hari dalam percakapan dan telah mengerti apa yang diartikan abreviasi tersebut. Namun, tidak pada abreviasi lainnya seperti "tp" dan "d" yang masih tidak mencolok dalam percakapan sehingga dapat terjadi kesalahpahaman dari abreviasi tersebut, sebagai contoh: "dia ingin membeli saham perusahaan Abc untuk tp, namun ia merasa akan kehilangan hartanya.", dimana abreviasi "tp" dapat berupa "tapi", "*take profit*", "tugas pembantuan".

C. *Cluster Dendrogram*

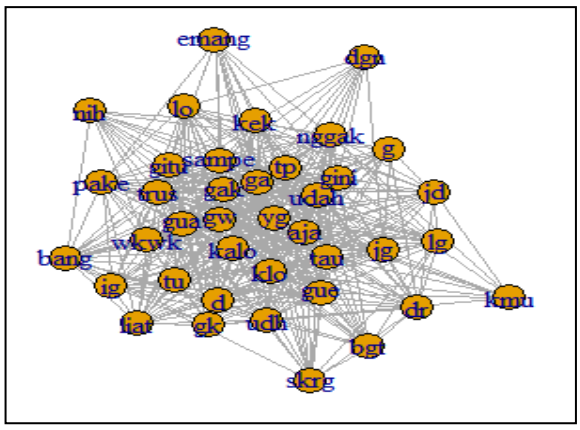


Gambar 3. Pengelompokkan Data dengan *Dendrogram*

Pada Gambar 3 dimulai dari *height* terpendek yaitu pada *cluster* "d" dan "lg" dengan *height* 4.44 dan *height* tertinggi yaitu cluster "yg" dan gabungan dari cluster terdekat dengan *height* 32.35. Pada setiap akhiran titik data bergabung dengan data yang mempunyai kesamaan nilai kuadrat yang signifikan sehingga terbentuk 1 *cluster*, proses pembentukan *cluster* ini dilakukan terus menerus hingga mencapai 1 *cluster* pada akhir *height*.

Dendrogram diatas terdapat 3 *cluster*, dimana *cluster* tersebut merupakan gabungan dari beberapa *cluster* lainnya, pada *cluster* ketiga hanya terdapat 1 data yang artinya data tersebut mempunyai nilai jumlah kuadrat yang jauh dari *cluster* lainnya.

D. *Relasi antar Abreviasi*



Gambar 4. *Network of Terms* Kata Abreviasi

Gambar diatas memvisualisasikan relasi antar abreviasi menggunakan tata letak *fruchterman-reingold*, abreviasi yang paling banyak digunakan dalam percakapan media sosial dengan memiliki relasi antar abreviasi lainnya antara lain: "yg", "kalo", "aja", "gw", "klo" dengan nilai derajat pada graph 38, 35, 34, 34, 34. Relasi antar abreviasi "yg" mempunyai hubungan dengan seluruh abreviasi yang ada pada Gambar 4.

Tabel 5. Derajat Abreviasi pada *Layout Fruchterman-Reingold*

Abreviasi	Degree
yg	38
kalo	35
aja	34
gw	34
klo	34
ga	33
tp	33
gini	32
tau	32
gua	31
gak	31
kek	28
udah	28
gitu	28
trus	28
udh	28
gk	26
nggak	26
wkwk	25
pake	24
sampe	24
d	24
liat	24
jg	24
tu	24
gue	22
bang	22
ig	22
lg	21
g	21
dr	21
jd	20
lo	20
bgt	17
nih	15
skrg	15
emang	12
dgn	12
kmu	10

Diketahui bahwa abreviasi "yg" mempunyai relasi paling banyak diantara abreviasi lainnya, hal ini bisa diketahui dengan dua cara, yaitu:

1) *Node* yang berada pada titik tengah tata letak *fruchterman-reingold* pada Gambar 4,
2) Mencari nilai derajat *graph* tertinggi dalam Table 5.

Terdapat juga beberapa abreviasi yang berada diluar jangkauan, antara lain: "emang", "dgn", "kmu", abreviasi ini merupakan abreviasi yang memiliki relasi antar abreviasi lainnya dalam percakapan media sosial, dengan nilai derajat dibawah 15.

Relasi antar abreviasi tersebut merupakan penggunaan kata-kata abreviasi pada satu dokumen, sehingga dokumen tersebut terdapat lebih dari 1 kata abreviasi yang berhubung satu sama lain.

IV. KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil penggunaan teknik text mining yang dilakukan untuk mengambil kata abreviasi yang sering digunakan dalam media sosial *Facebook*, *Twitter*, dan *Instagram* dapat disimpulkan bahwa penggunaannya memiliki persamaan yaitu menggunakan bentuk akronim, singkatan, kontraksi, penggalan, dan lambang huruf.

Selain itu, proses abreviasi yang ditemukan di dalamnya adalah pelesapan kata, pelesapan suku kata, pengekalan huruf, dan pengekalan suku kata. Sehingga, dari data yang digunakan diperoleh kata abreviasi "yg" bermakna "yang" menjadi kata yang paling banyak digunakan di media sosial *Facebook*, *Twitter*, dan *Instagram*.

Saran

Penulis mengharapkan adanya penggunaan metode stemming pada teknik preprocessing dengan tujuan agar penelitian memperoleh data yang lebih banyak dan mampu membandingkan penggunaan kata dasar dan kata abreviasi.

Selain itu, penulis menyarankan untuk peneliti selanjutnya untuk menggunakan variasi media yang lebih banyak dalam meneliti kata abreviasi, karena kata abreviasi memiliki bentuk yang sangat unik dan banyak digunakan di dalam media sosial saat ini.

UCAPAN TERIMA KASIH

Terima kasih banyak kepada bapak Akhmad Rezki Purnajaya, S.Kom., M.Kom. dan bapak Raymond Erz Saragih, S.Kom., S.S., M.Kom. telah membantu dan membimbing dalam pembuatan maupun penulisan penelitian ini.

DAFTAR PUSTAKA

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, Jan. 2010, doi: 10.1016/j.bushor.2009.09.003.

[2] R. C. Cenderamata, and A. N. Sofyan, "Abreviasi dalam Percakapan Sehari-hari di Media Sosial.", *Journal of Linguistics* vol. 4 no. 1, Apr 2019.

[3] R. Feldman and J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data," *Cambridge University Press*, 2007.

[4] M. Nurjannah, Hamdani, I. F. Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING", *Journal Informatika Mulawarman*, vol. 8, no. 3, Sep 2013.

[5] C. Williams, "Research Methods," *Journal of Business & Economic Research*, vol 5, no. 3, Mar 2007.

[6] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition", *The Morgan Kaufmann Series in Data Management Systems*, 2011.

[7] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *Jurnal RESTI*, vol. 1, no. 1, 2017.

[8] D. D. C. Nugraha, Z. Naimah, M. Fahmi, and N. Setiani "Klasterisasi Judul Buku dengan Menggunakan Metode K-Means", *Seminar Nasional Aplikasi Teknologi Informasi*, 2014.

[9] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, Dec. 2002, doi: 10.1177/016555150202800601.

[10] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 1, no. 2, pp. 58–64, Dec. 2019, doi: 10.28926/ilkomnika.v1i2.18.

[11] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim, "Installation, orientation, and layout," *Analyzing Social Media Networks with NodeXL*, pp. 55–66, 2020, doi: 10.1016/B978-0-12-817756-3.00004-2.