PEMBANGUNAN MODEL PREDIKSI KEPRIBADIAN BERDASARKAN *TWEET* DAN KATEGORI KEPRIBADIAN *BIG FIVE* DENGAN METODE *AGGLOMERATIVE HIERARCHICAL CLUSTERING*

Axel Haikal Yusup[1] , Warih Maharani[2]
[1,2]Universitas Telkom, Bandung
[1]axelhaikalyusup@students.telkomuniversity.ac.id,[2]wmaharani@telkomuniversity.ac.id

Abstrak

Media sosial merupakan forum tempat berinteraksi dan berbagi informasi bagi para penggunanya, melalui komunitas dan jejaring sosial. Banyaknya pengguna menjadikan media sosial menjadi sumber data untuk mengekstrak dan membuat informasi baru. Informasi yang diunggah di media sosial dapat digunakan oleh berbagai pihak untuk tujuan tertentu, seperti memprediksi harga saham, mendeteksi spam, atau mengukur kepribadian seseorang. Penelitian ini bertujuan untuk membangun model prediksi kepribadian dengan pendekatan *unsupervised learning* dengan memanfaatkan data *tweet* berbahasa Indonesia dalam jumlah terbatas. Metode *Agglomerative Hierarchical Clustering* dipilih untuk membangun model prediksi kepribadian seseorang berdasarkan konten di media sosial. Model pada penelitian ini memiliki akurasi 20.1% dengan rata-rata *silhouette score* -0.23. Keunikan kata yang tinggi dari setiap tweet yang diproses menjadi tantangan bagi model ini untuk menghasilkan performa yang optimal. Model ini dapat menangani data dalam jumlah besar dalam waktu singkat tetapi belum memberikan performa yang lebih optimal dibandingkan kasus serupa yang diselesaikan dengan *supervised learning*.

Kata kunci: media sosial, kepribadian, prediksi, metode, *tweet*

Abstract

Social media is a forum where users interact and share information through communities and social networks. The large number of users makes social media a source of data to extract and create new information. Information uploaded on social media can be used by various parties for specific purposes, such as predicting stock prices, detecting spam, or measuring a person's personality. This Research aims to build a personality prediction model with an unsupervised learning approach by utilizing a limited number of Indonesian-language tweets. The Agglomerative Hierarchical Clustering method was chosen to build a predictive model of a person's personality based on content on social media. The model in this research has an accuracy of 20.1% with an average silhouette score of -0.23. The high word uniqueness of each processed tweet is a challenge for this model to produce optimal performance. This model can handle large amounts of data in a short time but has not yet provided a more optimal performance than similar cases that were solved by supervised learning.

Keywords: social media, personality, prediction, methods, tweet

I. PENDAHULUAN

Media sosial adalah aplikasi komunikasi yang dimediasi oleh komputer yang memungkinkan penggunanya untuk membuat dan berbagi informasi melalui komunitas dan jaringan virtual. Kehadiran media sosial untuk masyarakat secara global bukanlah hal baru, data dan tren yang dihadirkan oleh *Hootsuite* menjelaskan bahwa jumlah pengguna aktif media sosial saat ini telah mencapai ± 3800 miliar pengguna pada tahun 2020. Pertumbuhan jumlah informasi di media sosial yang pesat sangat bergantung pada antusiasme penggunanya dalam kurun waktu tertentu. Informasi yang diunggah di media sosial dapat digunakan oleh berbagai pihak untuk tujuan tertentu, seperti memprediksi harga saham, memfilter suatu informasi, mendeteksi spam, atau mengukur kepribadian seseorang.

Kepribadian merupakan sesuatu yang melekat pada diri seseorang yang terbentuk dari karakteristik, pola pikir, perasaan, dan perilaku [1].

Ada lima ciri kategori kepribadian dari *Big Five Personality Traits*, yaitu *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism* [3]. Identifikasi kepribadian berdasarkan *Big Five Personality Traits* membutuhkan waktu yang lama dan kurang efisien karena diperlukan responden untuk mengisi kuesioner dengan total sekitar 20 sampai 360 pertanyaan [4]. Penelitian tentang identifikasi kepribadian berkembang dan menghasilkan metode baru seperti pengujian kepribadian otomatis, dan analisis media sosial [5]. Pengukuran kepribadian otomatis dengan memanfaatkan fitur linguistik dari sebuah teks di media sosial juga dapat menjadi cara lain untuk mengurangi biaya yang diperlukan jika dibandingkan dengan pengukuran manual [6].

Salah satu cara untuk melakukan Pengujian Kepribadian Otomatis atau pengenalan kepribadian otomatis adalah dengan menggunakan *machine learning*. Pengukuran

kepribadian pengguna *Twitter* dengan *machine learning* telah berhasil dilakukan dengan referensi [7] dan referensi [8] dimana kedua penelitian tersebut menggunakan *supervised learning* sebagai pendekatan untuk menyelesaikan masalah, namun performansi akurasinya bergantung pada jumlah data yang digunakan. Cara lain untuk mengatasi masalah tersebut dilakukan dengan referensi [9] yang menggunakan *Twitter* untuk mengukur kepribadian berdasarkan *Big Five Personality Traits* dengan menggunakan pendekatan ontologi.

Masalah umum yang terjadi dengan pendekatan *supervised learning* adalah jumlah data berlabel sebelumnya yang memiliki pengaruh besar pada kinerja model. Sebaliknya, model dengan pendekatan *unsupervised learning* dapat mendefinisikan kelompok data tanpa perlu dilatih dengan data berlabel sebelumnya. Oleh karena itu, dengan keterbatasan ketersediaan data, penelitian ini akan membangun model prediksi dengan pendekatan *unsupervised learning* dengan memanfaatkan data *tweet* berbahasa Indonesia dalam jumlah terbatas yang diperoleh dari beberapa akun yang bersedia dijadikan data dalam penelitian ini melalui pengisian kuesioner dan formulir kesediaan untuk mengetahui perbandingan hasil kinerja dan tantangan untuk dijadikan bahan perbaikan bagi penelitian di masa mendatang.

II. KAJIAN TEORI

A. Pengujian Kepribadian Otomatis

Perkembangan kecerdasan buatan dan teknologi *big data* memberikan cara baru untuk mengidentifikasi kepribadian seseorang melalui penilaian yang dilakukan oleh komputer. Dalam beberapa kasus, penilaian komputer dapat memberikan hasil yang hampir akurat atau bahkan melebihi akurasi rata-rata penilaian yang diberikan oleh manusia [5]. Penelitian sebelumnya yang berhasil mengidentifikasi kepribadian seseorang dengan pembelajaran mesin dan media sosial juga telah menghasilkan performa yang bervariasi tergantung jenis kasus dan setiap metode yang digunakan.

Pengujian kepribadian otomatis menggunakan pembelajaran mesin telah dilakukan oleh beberapa penelitian sebelumnya dengan memanfaatkan berbagai fitur data yang disajikan oleh *Twitter*. Beberapa fitur yang umum digunakan adalah metadata fitur sosial seperti *Followers*, *Following*, *Retweet*, *Likes*, dan fitur linguistik yang terdapat di setiap *Tweet* [10].

Penggunaan *tweet* bahasa Indonesia dan pembelajaran mesin pada referensi [11] dan referensi [10] menunjukkan bahwa kepribadian seseorang dapat diklasifikasikan menggunakan pendekatan *supervised learning* dengan *J48, k-NN*, *Naïve Bayes*, *Random Forest, SMO, Super Learner*, algoritma *XGBoost*, dan *Stochastic Gradient Descent.* Setiap algoritma memiliki kompleksitas tersendiri dalam mengevaluasi kinerjanya. Namun secara umum pengukuran dengan *supervised learning* dapat mengklasifikasikan dengan cukup akurat jika dibandingkan dengan penilaian oleh manusia.

B. *Big Five Inventory* (BFI-10)

BFI adalah alat untuk mengukur kecenderungan seseorang terhadap lima kategori kepribadian *Big Five* yang dibuat pada tahun 1980-an. *BFI* merupakan kuesioner yang berisi 44 pertanyaan tentang kecenderungan seseorang dalam berperilaku. Pada waktunya, mengisi 44 pertanyaan dengan rata-rata waktu 5 menit adalah sesuatu yang singkat dan masuk akal. Seiring berjalannya waktu, sesuatu yang dianggap singkat kini dianggap sesuatu yang panjang dan rumit sehingga 44 pertanyaan dari *BFI* kini disingkat menjadi 10 pertanyaan atau yang sekarang dikenal dengan *BFI-10* untuk menilai hal yang sama namun dalam satu menit atau bahkan kurang dari satu menit [12].

TABEL 1
KUESIONER *BFI-10*
DALAM BAHASA INGGRIS

I see myself as someone who…	*Likert-scale*				
… is reserved	(1)	(2)	(3)	(4)	(5)
… is generally trusting	(1)	(2)	(3)	(4)	(5)
… tends to be lazy	(1)	(2)	(3)	(4)	(5)
… is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
… has few artistic interests	(1)	(2)	(3)	(4)	(5)
… is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
… tends to find fault with others	(1)	(2)	(3)	(4)	(5)
… does a thorough job	(1)	(2)	(3)	(4)	(5)
… gets nervous easily	(1)	(2)	(3)	(4)	(5)
… has an active imagination	(1)	(2)	(3)	(4)	(5)

Tabel 1 menunjukan kuesioner BFI-10 bahasa inggris yang akan diterjemahkan ke Bahasa Indonesia pada Tabel 2 yang akan digunakan dalam penelitian ini. Kuesioner dibuat dalam bentuk skala likert dari 1 sampai 5 [13]. Skala Likert pada kuesioner ini mendefinisikan bahwa (1) berarti Sangat Tidak Setuju, (2) berarti Tidak Setuju, (3) berarti Netral, (4) berarti Setuju, dan (5) berarti Sangat Setuju.

Selain pertanyaan, ada juga nilai atau bobot dari jawaban untuk setiap pertanyaan yang nantinya akan digunakan untuk perhitungan hasilnya. Masing-masing ciri *Big Five* diwakili oleh dua pertanyaan, *Extraversion* dengan pertanyaan 1 dan 6, *Agreeableness* dengan pertanyaan 2 dan 7, *Conscientiousness* dengan pertanyaan

3 dan 8, *Neuroticism* dengan pertanyaan 4 dan 9 dan *Openness* oleh pertanyaan 5 dan 10.

TABEL 2
KUESIONER *BFI-10*
DALAM BAHASA INDONESIA

Saya melihat diri saya sebagai orang yang…	Skala Likert				
… pendiam	(1)	(2)	(3)	(4)	(5)
… mudah percaya	(1)	(2)	(3)	(4)	(5)
… cenderung malas	(1)	(2)	(3)	(4)	(5)
… santai, dan dapat menangani stres dengan baik	(1)	(2)	(3)	(4)	(5)
… memiliki sedikit minat artistik	(1)	(2)	(3)	(4)	(5)
… ramah dan mudah bergaul	(1)	(2)	(3)	(4)	(5)
… cenderung mencari kesalahan orang lain	(1)	(2)	(3)	(4)	(5)
… mengerjakan pekerjaan dengan teliti	(1)	(2)	(3)	(4)	(5)
… mudah gugup	(1)	(2)	(3)	(4)	(5)
… memiliki imajinasi tinggi	(1)	(2)	(3)	(4)	(5)

Selain waktu yang lebih singkat, penggunaan *BFI- 10* pada referensi [12] juga menjaga reliabilitas dan validitas 44 pertanyaan *BFI* secara signifikan.

C. *Unsupervised Learning*

Unsupervised Learning merupakan salah satu metode yang digunakan untuk memperoleh informasi tentang *data mining*. Umumnya metode ini digunakan untuk mengolah data tanpa label dengan memanfaatkan nilai statistik dari data yang akan dikelompokkan dan menjadi informasi baru [8].

Penyelesaian kasus serupa dengan memanfaatkan pendekatan *Unsupervised Learning* telah dilakukan pada penelitian sebelumnya, dimana penelitian tersebut menggunakan *tweet* dan fitur dari *LIWC* dalam model, dimana model tersebut menghasilkan rata-rata nilai *F1* sebesar 0,578 [14].

D. Cosine Similarity

Cosine Similarity merupakan metode yang digunakan untuk mencari nilai kemiripan antara dua vektor. Perhitungan *cosine similarity* dilakukan dengan menghitung nilai sudut *cosinus* antara masing-masing pasangan vektor yang akan dibandingkan [15]. Nilai lain yang dapat diperoleh setelah mendapatkan nilai *cosine similarity* pada pasangan vektor adalah nilai jarak *cosinus* [16].

$$Cosine\ Similarity = Cos\theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} \quad (1)$$

$$Cosine\ Distance = 1 - Cos\theta \quad (2)$$

E. *Agglomerative Hierarchical Clustering* (AHC)

AHC merupakan metode *clustering* dengan pendekatan hierarki *bottom-up*. Metode ini mengklasifikasikan data menjadi beberapa *cluster* berdasarkan jarak antar *cluster*, dimana setiap data akan direpresentasikan sebagai *cluster* yang berbeda. Algoritma ini akan menghitung jarak dan membandingkan hasil perhitungan secara iteratif untuk mencari jumlah *cluster* maksimum. Jarak antar *cluster* dapat diperoleh dengan berbagai rumus perhitungan jarak yang umum digunakan, seperti *Euclidean Distance, Manhattan Distance, Cosine Similarity*, atau rumus lainnya [16].

Proses pengembangan model AHC dilanjutkan dengan membandingkan jarak antar *cluster* yang telah didapatkan. Perbandingan jarak tersebut dapat dilakukan dengan tiga pendekatan yaitu *Single Linkage*, *Complete Linkage*, dan *Average Linkage* [16]. Penelitian ini akan memanfaatkan pendekatan *Single Linkage* dimana model AHC akan dibangun berdasarkan jarak *cluster* terdekat untuk setiap iterasi.

F. *Silhouette Score*

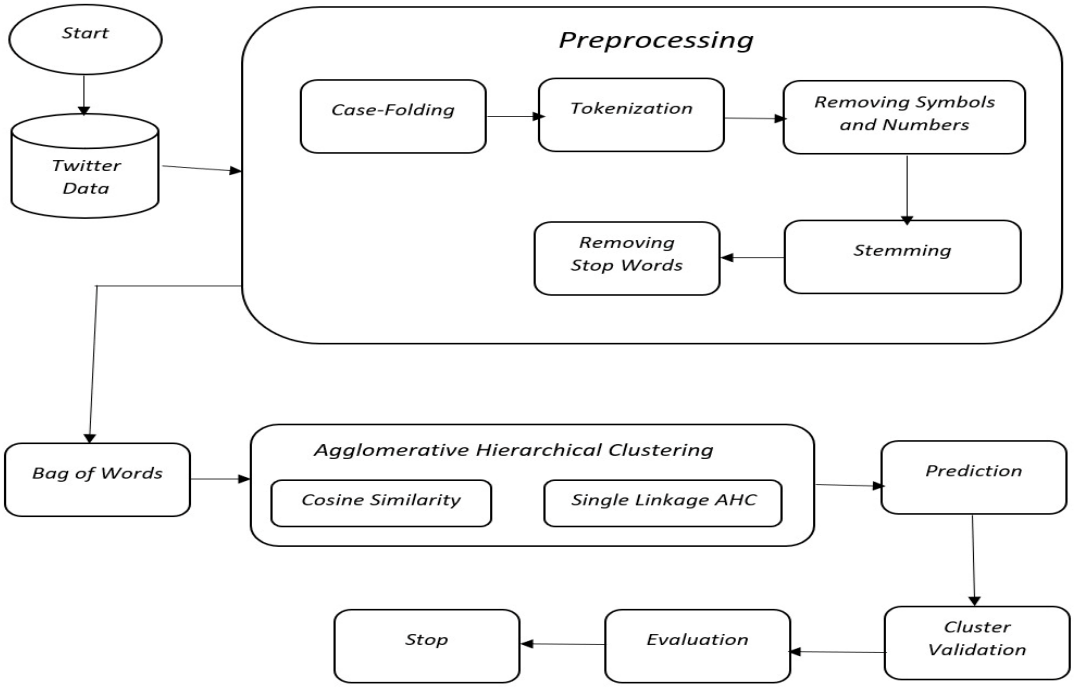
Silhouette Score adalah metode kalkulasi statistik untuk memperkirakan jumlah *cluster* terbaik untuk sebuah *dataset* [17]. *Silhouette Score* menghitung lebar berdasarkan jarak pada setiap titik atau data yang menjadi anggota suatu *cluster*, kemudian lebar diperoleh dengan menghitung rata-rata jarak total.

$$Silhouette\ Score = SI_k = \frac{1}{n} \sum_{i=1}^{n} \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (3)$$

Dimana n adalah jumlah data, a adalah jarak rata-rata antara data i dan data lain i dalam suatu cluster, dan b adalah jarak rata-rata minimum antara data lain dalam cluster yang i berbeda.

III. METODE

Model prediksi kepribadian dibangun dengan menggunakan bahasa pemrograman *Python*. Penggunaan *library scikit-learn* dengan *Python* juga dilakukan dengan mempertimbangkan kelengkapan fitur untuk membangun model pembelajaran mesin dengan metode *Agglomerative Hierarchical Clustering*. Gambaran umum sistem direpresentasikan dalam bentuk diagram blok yang dapat dilihat pada gambar 1



GAMBAR 1
GAMBARAN UMUM SISTEM

Dataset yang digunakan dalam pemodelan ini adalah kuesioner yang didistribusikan secara daring melalui Google *Form*. Kuesioner digunakan untuk memperoleh data pribadi, kategori kepribadian responden untuk dinilai secara manual berdasarkan pertanyaan BFI-10 pada referensi [2] yang diterjemahkan ke dalam bahasa Indonesia, dan persetujuan responden untuk menggunakan *tweet* sebagai data penelitian. Data yang digunakan adalah data dari responden yang mengisi kuisioner sampai dengan tanggal 14 Oktober 2020. Dari 124 akun twitter yang telah diisi oleh responden, hanya 69 *username* yang dapat digunakan karena beberapa responden telah merubah *username* twitter, atau beberapa akun dalam keadaan *private* sehingga akun tersebut tidak dapat di-*crawl*. Selain informasi yang diperoleh dari kuesioner, tahap pemodelan ini juga memanfaatkan korpus yang digunakan pada referensi [9] sebagai *gold standard* untuk penelitian ini

TABEL 3
CONTOH DATA TWITTER SEBELUM PRE-PROCESSING

Twitter *Username*	*Timestamp*	*Tweet*	Bahasa
@kingkon666	2020-10-14 09:06:15	“b’udah bulan oktober 2020, sepatu compass masih susah cari harga retail?”	in
@eksantiass	2020-10-08 05:37:34	“b’setelah lima semester, *at least* km tau ada nama eksantias di dunia ini ahaha”’	in
@jipawww	2020-10-06 08:22:28	b’@grumpydumbty Demo masak sih aku mau’	in

TABEL 4
CONTOH DATA TWITTER SEBELUM *PRE-PROCESSING*

Twitter *Username*	Token
@kingkon666	[‘sudah’, ‘bulan’, ‘oktober’, ‘2020’, ‘,’, ‘sepatu’, ‘compass’, ‘masih’, ‘susah’, ‘cari’, ‘harga’, ‘retail’, ‘?’]
@eksantiass	['setelah', 'lima', 'semester', ',', *'at', 'least',* 'km', 'tau', 'ada', 'nama', 'eksantias', 'd', 'dunia', 'ini', 'ahaha']
@jipawww	['demo', 'masak', 'sih', 'aku', 'mau']

Tabel 3 dan Tabel 4 menunjukan perubahan bentuk data dari sebelum *preprocessing* dan setelah *preprocessing*. Satu- satunya fitur data twitter yang digunakan sebagai masukan untuk model hanya *username* dan *tweet* dalam bentuk *token*.

TABEL 5
CONTOH KORPUS KATA YANG
BERLABEL *BIG FIVE PERSONALITY*

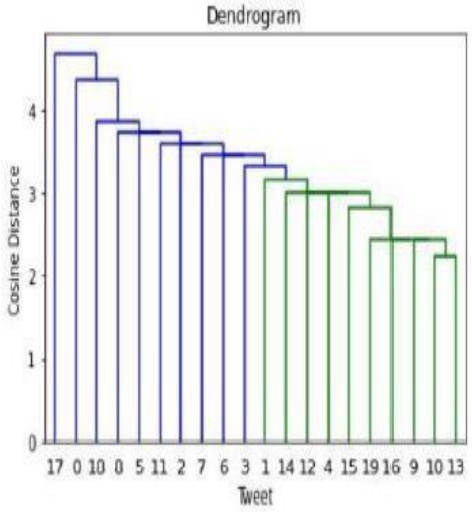
Text	***Meaning***	***Traits***
belajar	mau belajar	*Conscientiousness*
kasih	memberi	*Agreeableness*
luv	sayang	*Extraversion*
terlambat	lambat	*Neuroticism*
hmmm	berpikir	*Openness*

Tabel 5 menunjukan karakteristik korpus kata berlabel *Big Five Personality Traits* yang digunakan sebagai *gold standard* dalam penelitian ini, di mana korpus tersebut berisi 6306 kata yang telah dikelompokkan oleh psikolog berdasarkan *Big Five Personality* [9].

IV. HASIL DAN PEMBAHASAN

Kuesioner yang dibagikan berhasil mengumpulkan 124 responden yang bersedia dilibatkan dalam penelitian ini. Namun, hanya 69 data pengguna yang benar-benar sesuai dan dapat digunakan. Hal ini dikarenakan beberapa *username* milik responden telah diubah, beberapa akun belum mengunggah *tweet*, dan masih terdapat akun yang bersifat *private*.

Hasil jawaban kuesioner *BFI-10* untuk masing-masing responden kemudian dihitung dengan menggunakan perhitungan skala *likert* [13]. Perhitungan tersebut menghasilkan data kategori kepribadian masing-masing responden berdasarkan lima besar ciri kepribadian (*O, C, E, A, N*), data tersebut nantinya akan digunakan untuk menghitung keakuratan model yang dibangun.



GAMBAR 2
DENDOGRAM SEBAGAI REPRESENTASI *CLUSTER*
UNTUK *TWEET* DARI SETIAP AKUN

Model yang telah dibangun berhasil mengklasifikasikan 20,1% dari total 69 pengguna yang dianalisis dengan *silhouette score* rata-rata -0,23 untuk 5 parameter *cluster* dengan menghitung jarak *cosinus* dan *single-linkage* seperti yang ditampilkan pada gambar 2. *Silhouette score* yang kurang dari 0 berarti ada data yang salah *cluster*. Hal ini disebabkan banyaknya variasi bentuk kata unik yang digunakan dalam setiap *tweet* dan terbatasnya jumlah dan variasi kata yang terdaftar dalam kata korpus dengan label *Big Five Personality Traits*, yang dibuat secara manual hanya untuk tujuan penelitian yang dilakukan sebagai referensi [9].

Penelitian ini menunjukkan bahwa selain pilihan metode *machine learning* yang digunakan saat membangun model, bahasa fitur kebahasaan yang diolah juga mempengaruhi kinerja model. Performa model yang mengolah ciri kebahasaan bahasa Indonesia tidak memberikan hasil yang lebih baik jika dibandingkan dengan model yang mengolah ciri kebahasaan menggunakan bahasa yang terdaftar di *Linguistic Enquiry Word Count* (LIWC)

Sebagai pendekatan lain untuk mengoptimalkan kinerja, pembuatan korpus manual oleh tenaga ahli, perlu dilakukan karena *tweet* yang ditangani adalah bahasa Indonesia. Hal ini menjadi tantangan karena saat ini belum ada kosakata tertutup seperti LIWC yang secara umum dapat digunakan untuk mengolah fitur linguistik dalam bahasa Indonesia.

Penggunaan LIWC untuk mengolah fitur linguistik dalam bahasa tertentu sangat umum, dimana *LIWC* adalah alat untuk memproses fitur linguistik dengan pendekatan *closed-vocabulary* tersedia untuk bahasa inggris, mandarin, belanda, arab, jerman, italia, korea, norwegia, portugis, spanyol. Selain *LIWC*, alat yang mirip dengan pendekatan *closed-vocabulary* lainnya yang umum digunakan adalah *MRC, NRC, SentiStrength*, dan *SPLICE* [18].

V. KESIMPULAN

Model yang dibangun dengan menggunakan metode *Agglomerative Hierarchical Clustering* tidak memberikan hasil kinerja yang unggul dibandingkan dengan metode *supervised learning* lainnya yang juga menyelesaikan masalah serupa dengan *dataset tweet* Indonesia, dimana model *Agglomerative Hierarchical Cluster* yang telah dibangun memiliki akurasi sebesar 20,1% dengan rata- rata *silhouette score* -0.23.

Tantangan lain dalam mengimplementasikan pengenalan tipe kepribadian otomatis dengan *machine learning* adalah ketersediaan alat untuk memproses fitur linguistik seperti *LIWC (Linguistic Enquiry Word Count)*, yang saat ini hanya tersedia untuk bahasa inggris, mandarin, belanda, arab, jerman, italia, korea, norwegia, portugis, spanyol. Hal ini menyebabkan fitur linguistik dengan bahasa yang belum terdaftar pada LIWC membutuhkan biaya lebih untuk mengoptimalkan hasil model yang dibuat.

Optimalisasi pengolahan fitur linguistik dengan bahasa-bahasa yang tidak dapat ditangani oleh LIWC dapat dilakukan dengan pendekatan lain seperti penerjemahan, atau melibatkan pakar untuk membangun alat *closed-vocabulary* yang mirip dengan LIWC guna memudahkan pemrosesan fitur linguistik secara umum untuk sebuah bahasa tertentu. Hal ini juga dapat mengurangi biaya dan waktu yang dibutuhkan jika setiap masalah serupa masih membutuhkan keterlibatan pakar psikologi untuk menilai kategori kepribadian mulai dari pembentukan korpus kata, hingga penilaian yang komprehensif untuk setiap individu yang diteliti.

REFERENSI

[1] Ahmad, N. dan Siddique, J., "*Personality Assessment using Twitter Tweets*," Procedia *Computer Science*, pp.1964-1973,2017. doi.org/10.1016/j.procs.2017.08.067

[2] Goldberg, L.R., "*An alternative "Description of personality": The Big-Five Factor structure," Personality and Personality Disorders: The Science of Mental Health*, pp. 34, 2013. doi.org/10.4324/9780203822845

[3] Quercia, D., Michal, K., David, S. dan Jon, C., "*Our twitter profiles, our selves: Predicting personality with twitter*," Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE *International Conference on Social Computing*, PASSAT/SocialCom 2011, pp. 180-185, 2011. doi**:** 10.1109/PASSAT/SocialCom.2011.26

[4] Jaimes Moreno, D. R., Carlos Gomez, J., Almanza-Ojeda, D. L. dan Ibarra-Manzano, M. A., "*Prediction of personality traits in twitter users with latent features*," CONIELECOMP 2019 - 2019 *International Conference on Electronics, Communications and Computers*, Issue 3, pp. 176-181, 2019. doi:10.1109/CONIELECOMP.2019.8673242

[5] Ihsan, Z. dan Furnham, A., "*The new technologies in personality assessment: A review," Consulting Psychology Journa*l, 70(2), pp. 147–166, 2018. doi:10.1037/cpb0000106

[6] Mairesse, F., Walker, M.A., Mehl, M.R. dan Moore, R.K., "*Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*," *Journal of Artificial Intelligence Research*, 30, pp. 457-500, 2007.

[7] Yusra, Fikry, M., Syarfianto, R., Mai Candra,R., Budianta, E., "Klasifikasi Kepribadian *Big Five* Pengguna Twitter dengan Metode Naïve Bayes," *In* Seminar Nasional Teknologi Informasi Komunikasi dan Industri, pp. 317-321, 2018.

[8] Mann, A.K, dan Kaur, N., "*Review paper on clustering techniques*," *Global Journal of Computer Science and Technology*, 2013.

[9] Alamsyah, A., Bastikarana, R.S., Ramadhanti, A.R. dan Widiyanesti, S., "*Recognizing Personality from Social Media Linguistic Cues: A Case Study of Brand Ambassador Personality," 2020 8th International Conference on Information and Communication Technology* (ICoICT), pp. 1-5, Juni 2020. doi**:** 10.1109/ICoICT49345.2020.9166221

[10] Jeremy, N. H., Prasetyo, C. dan Suhartono, D., "Identifying *personality traits for Indonesian user from twitter dataset*," *International Journal of Fuzzy Logic and Intelligent Systems*, 19(4): pp.283-289, 2019.

[11] Adi, G.Y.N., Tandio, M.H., Ong, V. dan Suhartono, D., "*Optimization for automatic personality recognition on Twitter in Bahasa Indonesia," Procedia Computer Science*, 135: pp.473-480,2018. doi.org/10.1016/j.procs.2018.08.199

[12] Rammstedt, B. dan John, O.P., "*Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," Journal of research in Personality*, 41(1), pp.203-212, 2007.

[13] Wu, C.H., "*An Empirical Study on the Transformation of Likert-scale data to numerical scores," Applied Mathematical Sciences*, 1(58), pp.2851-2862, 2007.

[14] Celli, F. and Rossi, L., "*The Role of Emotional Stability in Twitter Conversations," In Proceedings of the Workshop on Semantic Analysis in Social Media*, pp. 10-17, April 2012.

[15] Sahu, L. dan Mohan, B.R., "*An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop," In 2014 9th International Conference on Industrial and Information Systems* (ICIIS), pp. 1-5, 2014.

[16] Sasirekha, K. dan Baby, P., "*Agglomerative Hierarchical Clustering Algorithm-A Review," International Journal of Scientific and Research Publications*, pp.83, 2013

[17] Saitta, S., Raphael, B., dan Smith, I.F., "*A bounded index for cluster validity," In International workshop on machine learning and data mining in pattern recognition*, pp. 174-187, July 2007. doi.org/10.1007/978-3-540-73499-4_14

[18] Ong, V., Rahmanto, A.D., Williem dan Suhartono, D., "*Exploring personality prediction from text on social media: a literature review," Internetworking* Indonesia, 9(1): pp. 65-70, 2017.