Volume 5 Nomor 1 Januari 2020

INFORMASI INTERAKTIF

JURNAL INFORMATIKA DAN TEKNOLOGI INFORMASI

PROGRAM STUDI TEKNIK INFORMATIKA – FAKULTAS TEKNIK -UNIVERSITAS JANABADRA

INFORMASI INTERAKTIF	Vol. 5	No. 1	Hal. 1-38	Yogyakarta Januari 2020	ISSN 2527-5240

INFORMASI INTERAKTIF

ISSN 2527-5240

DEWAN EDITORIAL

Alamat Redaksi : Program Studi Teknik Informatika Fakultas Teknik
Universitas Janabadra
Jl. Tentara Rakyat Mataram No. 55-57
Yogyakarta 55231
Telp./Fax : (0274) 543676
E-mail: informasi.interaktif@janabadra.ac.id
Website : http://e-journal.janabadra.ac.id/

Frekuensi Terbit : 3 kali setahun

JURNAL INFORMASI INTERAKTIF merupakan media komunikasi hasil penelitian, studi kasus, dan ulasan ilmiah bagi ilmuwan dan praktisi dibidang Teknik Iinformatika. Diterbitkan oleh Program Studi Teknik Informatika Fakultas Teknik Universitas Janabadra di Yogyakarta, tiga kali setahun pada bulan Januari, Mei dan September.

ISSN 2527-5240

DAFTAR ISI

ISSN 2527-5240

PENGANTAR REDAKSI

Puji syukur kami panjatkan kehadirat Allah Tuhan Yang Maha Kuasa atas terbitnya JURNAL INFORMASI INTERAKTIF Volume 5, Nomor 1, Edisi Januari 2020. Pada edisi kali ini memuat 5 (lima) tulisan hasil penelitian dalam bidang teknik informatika.

Harapan kami semoga naskah yang tersaji dalam JURNAL INFORMASI INTERAKTIF edisi Januari tahun 2020 dapat menambah pengetahuan dan wawasan di bidangnya masing-masing dan bagi penulis, jurnal ini diharapkan menjadi salah satu wadah untuk berbagi hasil-hasil penelitian yang telah dilakukan kepada seluruh akademisi maupun masyarakat pada umumnya.

Redaksi

TEXT MINING DOKUMEN TWEET PADA TWITTER UNTUK KLASIFIKASI KARAKTER CALON KARYAWAN

Saifudin[1)], Kusrini[2)], Hanif Al Fatta[3)]

[1,2,3]Magister Teknik Informatika Universitas AMIKOM Yogyakarta, Depok Sleman
Jl. Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta Indonesia 55283

Email: [1)]saifudin.0941@students.amikom.ac.id, [2)]kusrini@amikom.ac.id, [3)]hanif.a@amikom.ac.id

ABSTRACT

Recruitment is a means to prepare as many workers as possible according to the requirements and qualifications expected by the organization. In recruitment one of the things that is calculated is the character of the prospective employee itself. Companies or organizations usually carry out psychological tests and interviews to get the character of prospective employees. This will make the recruitment process longer and require a lot of money. One way to find out a person's character can be done by looking at the publication of daily activities on various social media. In this study the classification of prospective employees is based on tweets found on twitter. The results of this study are grouping prospective employees based on their characters using the naïve bayes classifier algorithm. From the research that has been done naïve bayes classifier algorithm has an accuracy accuracy of an average of 52% by weighting using the term document frequency.

Keywords: *Naïve Bayes Classifier, TFIDF, Character*

1. PENDAHULUAN

Salah satu hal yang penting dalam sebuah perusahaan atau organisasi adalah menejemen sumber daya manusia. Unsur penting dalam menejemen sumber daya manusia adalah rekrutmen. Rekrutmen merupakan sarana untuk menyiapkan sebanyak-banyaknya tenaga pekerja yang sesuai dengan syarat dan kualifikasi yang diharapkan oleh organisasi untuk menyelesaikan pekerjaan yang sudah disiapkan (*job discription*). Keberhasilan rekrutmen pekerja, dalam organisasi sangat menentukan keberhasilan terwujudnya tujuan organisasi itu sendiri. Rekrutmen yang baik adalah rekrutmen yang bisa menjawab kebutuhan pekerjaan yang disiapkan oleh organisasi. Karena begitu pentingnya proses rektumen ini, tidak mustahil bahwa tugas manajemen sumber daya berikutnya menjadi lebih mudah dan berjalan dengan baik [1]. Dalam rekrutmen hal yang sangat penting adalah karater dari calon pegawai itu sendiri. Perusahaan atau organisasi biasanya melakukan test psikologi dan wawancara untuk mendapatkan karakter dari calon pegawainya.

Untuk mengetahui karater seseorang dapat dilakukukan dengan melihat publikasi kegiatan sehari-hari di dalam *social media*. Twitter dalah salah satu media social terbesar dengan penggunan lebih dari 300 juta pengguna. Setiap hari lebih dari 100 juta tweet dibagikan oleh penggunanya [2]. Dengan pertumbuhan pengguna social media yang semakin pesat hal tersebut dapat dimanfaatkan diberbagai bidang. Salah satu pemanfaatannya adalah untuk *Automatic Personality Recognition* (APR). Tujuan dari APR adalah untuk secara otomatis mengklasifikasikan karakter kepribadian seseorang menggunakan model kepribadian seperti *Big Five*, dan *Myers-Brigg Type Indicator* [3].

Dari beberapa rivew yang dilakukan terhadap beberapa paper, banyak paper yang melakukan penelitian dengan menggunakan berbagai macam algoritma diantaranya K-NN, Naïve Bayes, *Stochastic Gradient Descent* (SGD) dan lain-lain, terdapat perbedaan akurasi yang didapat dari penelitian tersebut.rata-rata akurasi yang didapat dari penelitian tersebut kurang dari 80% dengan berbagai cara dan kondisi yang diterapkan pada algorima. Jumlah dataset yang dipakai pada proses training sangat berpengaruh terhadap hasil akurasi yang didapat saat proses penelian. Selain jumlah dataset banyak faktor yang dapat mempengaruhi hasil akurasi dari algoritma yang digunakan, diantaranya adalah metode-metode yang dipakai seperti text preprocessing. *Text preprocessing* yang dilakukan peneliti sebelumnya belum sepenuhnya sempurna karena masih belum bisa menangani masalah seperti penghapusan awalan dan akhiran secara efektif pada tahap stemming. Selain itu penggunaan query expansion sangat sabgat berpebgaruh terhadap keberhasilan algoritma yang dipakai, dengan

pengguaan query expansion dapat meningkatkan akurasi yang diperoleh.

Berdasarkan latar belakang tersebut, penggabungan dua metode yang diusulkan akan digunakan untuk klasifikasi karekter calon karyawar berdasarkan tweet, penulis merasa perlu melakukan penelitian untuk mengimplementasikan dan menguji akurasi dari naïve bayes classifier klasifikasi karekter calon karyawan berdasarkan tweet.

1.1 Klasifikasi

Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas atau konsep data. Model diturunkan berdasarkan analisis satu set data. Misalnya dalam sebuah penjualan elektronik anda ingin mengklasifikasi satu set besar item dalam toko berdasarkan tiga jenis yaitu respon pembeli terhadap penjualan yakni baik, biasa dan tidak ada respon. Anda ingin mendapatkan model untuk masing-masing dari tiga kelas ini berdasarkan fitur deskripsi dari item seperti harga, jenis dan kategori.

Klasifikasi yang dihasilkan harus secara maksimal membedakan masing-masing kelas dari yang lain dan mengdeskripsikan keputusan. Keputusannya misalnya dapat mengidentifikasi harga sebagai faktor tunggal yang paling membedakan ketiga kelas tersebut. Artinya disamping harga, fitur lain dapat membedakan objek dari masing-masing kelas dari satu sama lain termasuk merek dan jenis [4].

1.2 Naive Bayes

Naive Bayes Bayesian Classification (NBC) adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class [5]. NBC merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi [6]. Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap pengujian/ klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel, sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap pengujian/ klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi. Berikut ini pada gambar 4 tahapan dalam proses algoritma naïve bayes.

1.3 Naive Bayes

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi/evaluasi. *K-Fold Cross Validation* membagi himpunan menjadi k bagian secara acak dan saling bebas. Data sebanyak (k-1) fold digunakan untuk mentraining model sedangkan 1 fold digunakan untuk melakukan testing. Validasi dilakukan sebanyak k kali hingga semua data dalam dataset diujikan pada model [7].

1.4 Big Five

The Big Five digunakan untuk pencarian fitur pada beberapa media sosial seperti FriendFeed, Facebook dan Twitter. Feature yang digunakan biasanya diadopsi dari *Linguistic Inquiry and Word Count* (LIWC) tool dan MRC Psycholinguistic Database. Feature yang digunakan dalam penelitian ini juga diadopsi dari gabungan LIWC dan MRC sebanyak 22 feature [8].

2. TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Pada penelitian yang dilakukan Joda Pahlawan Romadhona Tanjung dkk (2017) melakukan klasifikasi tweet pada twitter dengan menggunkan metode *Fuzzy KNearest Neighbour* (Fuzzy K-NN) dan Query Expansion Berbasis Apriori. Proses pengklasifikasian sebuah tweets sukar dilakukan karena tweets berupa *short-text*. Oleh karena itu, sebelum dilakukan proses klasifikasi sebuah tweets dilakukan preprocessing dan proses ekspansi kata terlebih dahulu dengan algoritme Query Expansion agar memberikan hasil maksimal pada proses klasifikasi. Penggunaan query expansion sangat berpengaruh dalam pemodelan, karena dengan menggunkan query expansion dapat meningkatkan akurasi algoritma yang sangat signifikan. Akurasi tertinggi yang didapatkan jika menggunakan query expansion adalah sebesar 76%. Kemudian akurasi tertinggi

yang diperoleh jika tanpa menggunakan query expansion adalah sebesar 58% [9].

Penelitian terdahulu yang dilakukan oleh Gabriel Yakub dkk (2018) tentang teknik optimasi untuk pengenalan kepribadian otomatis (APR) berdasarkan Twitter dalam Bahasa Indonesia, bahasa ibu orang Indonesia. Ada tiga algoritma pembelajaran mesin yang digunakan dalam penelitian ini, yaitu Stochastic Gradient Descent (SGD), dan dua algoritma pembelajaran ensemble, Gradient Boosting (XGBoost) dan susun (super learner). Dalam penelitian tersebut berhasil meningkatkan algoritma pembelajaran mesin untuk memprediksi kepribadian pengguna secara otomatis. Dalam penelitian juga pendekatan dapat diadaptasi oleh penelitian lain untuk mengatasi masalah yang sama. Perbaikan lebih lanjut dapat dilakukan dengan memperluas kumpulan data dan menggunakan algoritma pembelajaran mendalam untuk memprediksi kepribadian pengguna secara otomatis (Adi, Tandio, Ong, & Suhartono, 2018).

Dalam beberapa tahun terakir banyak penelitian tentang klasifikasi karakter seseorang berdasarkan tweet pada twitter salah satunya yang dilakukan Anisha Yata dkk (2018). Penelitian tersebut bertujuan mengotomatiskan prediksi kepribadian pengguna oleh mengimplementasikan multi-label classifier pada data tekstual. Penelitian dilakukan dengan membandingkan beberapa algoritma untuk dicari tingkat akurasi tertinggi. Beberapa metode klasifikasi yang dipakai seperti Naive-Bayes, K- Nearest Neighbors dan Support Vector Machine bersama dengan multi-label classifier seperti Binary Relevance, Chains Classifier dan Random k-Label [10].

Penelitian dari Veronica Ong dkk (2017) melakukan klasifikasi personality berdasarkan twitter dengan Bahasa Indonesia. Dalam penelitian tersebut berupaya membangun sistem prediksi kepribadian berdasarkan pada pengguna twitter informasi untuk Bahasa Indonesia, bahasa asli Indonesia. Sistem prediksi kepribadian dibangun Dukungan Mesin Vektor dan XGBoost dilatih dengan 329 contoh (pengguna). Hasil evaluasi menggunakan salib 10 kali lipat validasi menunjukkan bahwa sistem berhasil mencapai tertinggi akurasi rata-rata 76,2310% dengan Support Vector Machine dan 97,9962% dengan XGBoost [11].

Penelitian yang bertujuan untuk untuk menunjukkan efek dari algoritma yang diusulkan dalam mengklasifikasikan dokumen yang ditulis dalam bahasa Arab (2018). Dalam penelitian tersebut algoritma AC hybrid baru (HAC) diusulkan. HAC menerapkan kekuatan algoritma Naïve Bayes (NB) untuk mengurangi jumlah aturan klasifikasi dan menghasilkan beberapa aturan yang mewakili setiap nilai atribut. Dua percobaan dilakukan pada dataset teks Arab menggunakan enam algoritma yang berbeda, yaitu J48, NB, berbasis klasifikasi pada asosiasi (CBA), klasifikasi multi-kelas berdasarkan aturan asosiasi (MCAR), pakar multi-kelas klasifikasi berdasarkan aturan asosiasi (EMCAR), dan algoritma klasifikasi asosiatif cepat (FACA). Hasil percobaan menunjukkan bahwa pendekatan HAC menghasilkan akurasi klasifikasi yang lebih tinggi dibandingkan MCAR, CBA, EMCAR, FACA, J48 dan NB dengan keuntungan 3,95%, 6,58%, 3,48%, 1,18%, 5,37% dan 8,05% masing-masing. Selanjutnya, pada dataset Reuters-21578, hasilnya menunjukkan bahwa Algoritma HAC telah kinerja yang sangat baik dan stabil dalam hal akurasi klasifikasi [12].

3. METODE PENELITIAN

Dalam penelitian ini menggunakan metode pengumpulan data dan perancangan sistem. Dalam pengumpulan data terdapat dua cara yaitu dengan pengumpulan data primer dan data sekunder. Dalam pengumpulan data primer, penulis memperoleh data dari twitter yaitu berupa data tweet dari user yang telah diklasifikasi karakternya terlebih dahulu oleh psikolog. Dalam data sekunder penulis memperoleh data dari jurnal atau literature terkait.

Gambaran umum terkait penelitian yang akan dilakukan adalah pada gambar 1 sebagai berikut:

a. Identifikasi Maslah

Identifikasi masalah, dalam proses ini terdapat pemilihan dan perumusan masalah serta identifikasi dan pemberian definisi operasional dari masalah yang dihadapi.

b. Tahapan Review

Berisi bagaimana mencari sumber data, contoh data, penentuan object penelitian dan jenis dataset yang digunakan

c. Pengumpulan Data

Langkah selanjutnya bagaimana mengumpulkan data baik dari jenis dataset, jumlah dataset dan variasi dataset yang diperlukan.

d. Pengolahan Data

Data yang digunakan sebagai dataset adalah tweet yang telah dikumpulkan dari beberapa orang dan telah di klasifikasi berdasarkan karakter orang tersebut.

Preprocessing, dalam pengolahan data langkah pertama yang dilakukan adalah melakukan preprocessing. Langkah ini dilakukan melalui beberapa tahapan yaitu [4]:

1. Tokenizing
   Dalam tahapan ini dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya
2. Filtering
   Dalam tahap ini dilaukakan tahap mengambil kata-kata penting dari hasil tokenizing.
3. Stemming
   Merupakan tahapan mencari kata dasar dari tiap kata hasil filtering, dalam tahap ini dilakukan proses pengembalian berbagai bentuk kata ke dalam satu representasi yang sama
4. Weighting
   Tahapan ini dilakukan dengan tujuan untuk memberikan suatu nilai/bobot pada term pada hasil stemming
5. Ektraksi Fitur
6. Multinominal Naïve Bayes
   Multinomial Naive Bayes merupakan sebuah metode yang bekerja dengan cara menghitung frekuensi setiap term pada dokumen.

e. Query Expansion

Query Expansion yang berguna untuk mengingkatkan kualitas query yang dimasukkan.

f. Naïve Bayes Classifier

Naive Bayes Bayesian Classification (NBC) adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class [5]. NBC merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi [6].

Persamaan Metode Naive Bayes seperti pada persamaan 1 dibawah ini :

$$P(H|X)(P(H|X).P(H))/(P(X)) \quad (1)$$

Keterangan :
X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

g. Testing

pengujian dilakukan dengan dataset yang belum pernah digunakan untuk training sehingga data yang teruji merupakan databaru dan belum dipelajari sistem sebelumnya.

h. Validasi

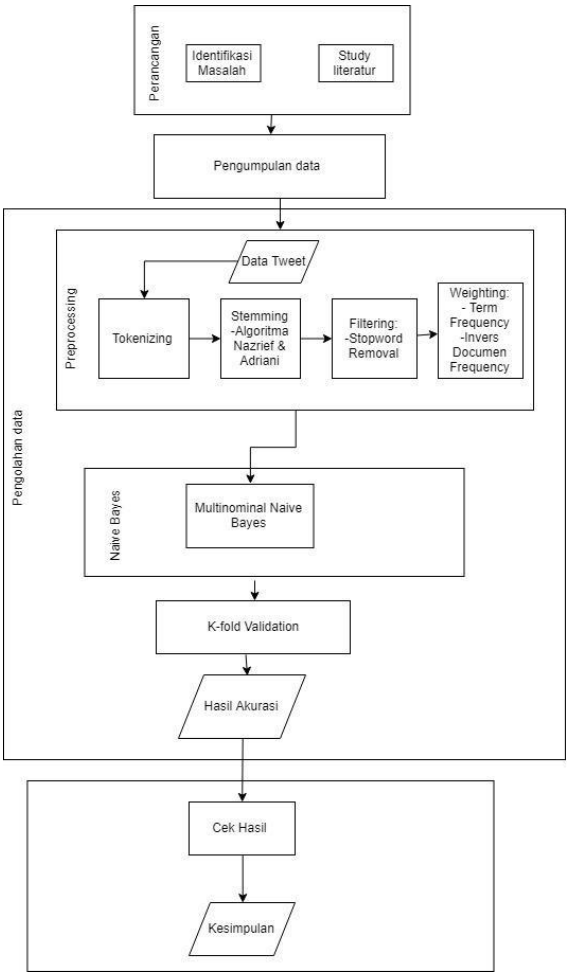
validasi menggunakan k-fold cross validation dimana dataset tertentu dipilih untuk digunakan sebagai skema untuk pengujian dengan data sebanyak (k-1) fold.

i. Kesimpulan dan evaluasi,

dalam kesimpulan menyajikan hasil dari penelitian yang telah dilakukan dan bagaimana evaluasi yang sebaiknya dilakukan.

4. HASIL DAN PEMBAHASAN

Hasil dalam penelitian ini terangkum dalam Gambar 1 dibawah ini.



Gambar 1. Alur penelitian

4.1 Input dan Preprocessing Data

Data yang digunakan berupa data *tweet* dari user yang telah diklasifikasi sebelumnya oleh psikiater. Klasifikasi karater yang digunkan adalah klasifikasi *big five.* Class yang dipakai adalah *constentiousness, openess to experiences, emotional stability, agreeableness* dan *extraversion.* Berikut hasil pengumpulan data seperti pada tabel 1.

Tabel 1**.** Data Uji

NO	User	Karater
1	User 1	*constentiousness*
2	User 2	*openess to experiences*
3	User 3	*emotional stability*
4	User 4	*openess to experiences*
5	User 5	*emotional stability*
6	User 6	*agreeableness*
7	User 7	*extraversion*
..	…..	…..
..	…..	…….
38	User 10	*extraversion*

a. *Tokenizing*

Tahapan pemotongan string input berdasarkan tiap kata yang menyusunnya. Dalam tokenizing juga dapat membuang beberapa karakter yang dianggap sebagai tanda baca.

Tabel 2. Hasil *Tokenizing*

Word	Atribute	Total Occurences
Abadi	Abadi	2.0
Ada	Ada	4.0
Aduuhh	Aduuhh	1.0
Agama	Agama	1.0
Ahmad	Ahmad	2.0
Air	Air	2.0
Akhir	Akhir	2.0
Akhirnya	Akhirnya	1.0
……...	……..	……..
yapp	yapp	1.0
yang	yang	78.0

b. *Filtering*

Tahap filtering adalah tahap mengambil kata-kata penting dari hasil tokenizing banyak algoritma yang bisa digunakan dalam tahap ini.

Tabel 3. Hasil *Filtering*

Word	Atribute	Total Occurences
Abadi	Abadi	2.0
Ada	Ada	4.0
Aduuhh	Aduuhh	1.0
Agama	Agama	1.0
Belajar	Belajar	4.0
Belum	Belum	1.0
Bencana	Bencana	1.0
Beraktifitas	Beraktifitas	2.0
……..	……	……..
yang	yang	78.0

c. *Stemming*

Menurut Peter Willet (1997) stemming adalah proses untuk menggabungkan atau memecahkan setiap varian-varian suatu kata menjadi kata dasar. Tahapan stemming adalah tahapan mencari kata dasar dari tiap kata hasil filtering, dalam tahap ini dilakukan proses pengembalian berbagai bentuk kata ke dalam satu representasi yang sama.

Tabel 4. Hasil *Steaming*

Word	Hasil
Abadi	Abadi
Ada	Ada
Aduuhh	Aduuhh
Agama	Agama
Belajar	Belajar
Belum	Belum
Bencana	Bencana
Beraktifitas	Aktivittas
……..	……
yang	yang

4.2 Pembobotan Kata (*term weighting*)

Term frequency (TF) meruakan jumlah atau frekuensi kemunculan sebuat term di dalam sebuah dokumen. Semakin sering sebuah term muncul dalam sebuah dokumen maka sekain tinggi pula bobotnya.

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata. Metode ini merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi

dokumen yang bersangkutan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat [13]. Unutk menghitung TFIDF dalam penelitaian ini menggunkan persamaan 2 dibawah

$$IDFj = \log(D/dfj) \quad (2)$$

Dimana D adalah jumlah semua dokumen dalam koleksi sedangkan dfj adalah jumlah dokumen yang mengandung term (tj)

4.3 Kesimpulan (*conclusion*)

Kesimpulan yang diperoleh dari hasil *training*, *testing* dan *validation* yang dilakukan adalah memiliki hasil akurasi. Akurasi yang diperoleh di penelitian ini depangaruhi oleh pre-processing dan pembobotan. Dengan teknik pembobotan yang berbeda akan menghasilkan hasil akurasi algoritma yang berbeda juga.

4.4 Analisis Hasil Penelitian

Dari beberapa pengujian yang dilakukan dengan dua model pembobotan yang berbeda yaitu dengan *Term Frequency* dan menggunakan *Term Frequency Inverse Document Frequency* (IDF). Data yang digunakan dalam pengujian sebanyak 38 user, dengan setiap user diambil *tweet*-nya. Untuk penggujian yang dialakukan dengan menggunkan *k-fold validation* dengan jumlah fold sebanyak 10. Dari pengujian deperoleh hasil akurasi seperti pada tabel di bawah. Tabel 5 menjukan hasil dari pengujian dengan pembobotan *Term Frequency*

Tabel 5**.** Hasil Pengujian Pertama

	true con	**true otx**	**true es**	**true agr**	**true ext**	**class precision**
pred. con	5	1	1	2	0	55.56%
pred. otx	1	2	1	0	2	33.33%
pred. es	1	2	2	0	0	40.00%
pred. agr	3	1	2	9	1	56.25%
pred. ext	1	0	2	2	7	58.33%
classr ecall	45.45%	33.33%	25.00%	69.23%	70.00%	

Dari tabel 5 diatas hasil rata-rata akurasi yang didapatkan yaitu sebesar 52 % dengan jumlah fold sebanyak 10.

Tabel 6**.** Hasil Pengujian Kedua

	true con	**true otx**	**true es**	**true agr**	**true ext**	**class precision**
pred. con	5	2	1	2	0	50.00 %
pred. otx	1	1	2	0	2	16.67 %
pred. es	1	2	2	0	0	40.00 %
pred. agr	3	1	1	9	1	60.00 %
pred. ext	1	0	2	2	7	58.33 %
classre call	45.45%	16.67%	25.00%	69.23%	70.00%	

Dari tabel 5 diatas hasil rata-rata akurasi yang didapatkan yaitu sebesar 50 % dengan jumlah fold sebanyak 10.

5. KESIMPULAN

Kesimpulan yang diperoleh dari hasil penelitian tersebut adalah sebagai berikut:

1. *Pre-processing* sangat mempengaruhi hasil dari akurasi algoritma yang dipakai, terutama pada proses pembobotan.
2. Hasil pembobotan dengan menggunakan TFID memiliki hasil akurasi yang lebih baik dibandingan dengan menggunkan *term-frekuensi.*
3. Nilai akurasi tertinggi yaitu 52 % dengan menggunkan pengujian k-flod validation.

Nilai akurasi terendah adalah sebesar 50% dengan jumlah data yang dipakai sebanyak 38 user dengan *class precision* rata- rata 48,7 % dan *recall* sebesar 48,6 %.

DAFTAR PUSTAKA

[1] M. Iskarim, "Rekrutmen Pegawai Menuju Kinerja Organisasi," in *Jurnal Manajemen Pendidikan Islam*, 2017.

[2] Y. . I. Claudy, R. S. Perdana and M. A. Fauzi, "Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2018.

[3] G. . Y. N. Adi, M. H. Tandio, V. Ong and D. Suhartono, "Optimization for Automatic Personality Recognition," *3rd International Conference on Computer Science and Computational Intelligence*, 2018.

[4] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morghan Kaufamnn Publisher , 2011.

[5] Kusrini and E. T. Luthfi, Algoritma Data Mining. Edisi 1, Yogyakarta: Andi Offset, 2009.

[6] B. Santosa, Data Mining: Teknik Pemanfaatan Data untuk keperluan bisnis, Yogyakarta: Graha Ilmu., 2007.

[7] Suyanto, Data Mining untuk Klasifikasi dan Klasterisasi Data, Bandung: Penerbit Informatika, 2017.

[8] N. Febrianto, I. Prasetia and A. Wijaya, "Pembuatan Sistem Prediksi Kepribadian “The Big Five Traits” dari," 2016.

[9] J. . P. R. Tanjung, M. . A. Fauzi and I. , "Klasifikasi Tweets Pada Twitter Dengan Menggunakan Metode Fuzzy KNearest Neighbour (Fuzzy K-NN) dan Query Expansion Berbasis Apriori," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2017.

[10] A. Yata, P. Kante, T. Sravani and B. Malathi, "Personality Recognition using Multi-Label Classification," *International Research Journal of Engineering and Technology (IRJET)*, 2018.

[11] V. Ong, A. D. S. Rahmanto, W. . D. Suhartono, A. E. Nugroho, E. W. Andangsari and M. N. Suprayogi, Personality Prediction Based on Twitter Information, Proceedings of the Federated Conference on Computer Science and Information Systems, 2017.

[12] W. Hadi, Q. A. Al-Radaideh and . S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification," *Applied Soft Computing*, 2018.

[13] Robertson and Stephen, "Understanding Inverse Document Frequency: On theoretical arguments for IDF," *Journal of Documentation*, vol. Vol. 60.