Regresi Linear untuk Mengurangi Bias Sistem Penilaian Uraian Singkat

(*Linear Regression for Reducing the Bias of a Short Essay Scoring System*)

Silmi Fauziati[1], Adhistya Erna Permanasari[1], Indriana Hidayah[1], Eko Wahyu Nugroho[1], Bobby Rian Dewangga[1]

***Abstract*—This study is aimed to improve the performance of a short essay scoring system. The improvement is executed by integrating a simple linear regression to the output of a combined cosine similarity method (with weighted term frequency using Term Frequency – Inverse Document Frequency (TF-IDF) method) and term-matching mechanism. The linear regression is conducted by taking the short essay score (resulting from the combined cosine similarity and term matching) as a regressor variable. In order to demonstrate the effectiveness of the proposed scoring system, the performance of the scoring system is measured relative to manual scoring by a lecturer. The results show that prior to linear regression, the scoring system tends to give higher score (biased score) compared to the manual score, which is problematic. The following scoring system with linear regression tackles this problem as the scoring bias is significantly reduced, that is, no tendency to give higher or less score compared to the manual score. That the scoring bias is significantly reduced using a simple approach, linear regression, is expected to contribute in the acceleration of implementing automated essay scoring system on online learning technologies such as e-learning.**

***Intisari*—Makalah ini bertujuan untuk memperbaiki kinerja sistem penilaian tes uraian singkat. Perbaikan kinerja tersebut dilakukan dengan menambahkan regresi linear sederhana pada keluaran gabungan metode *cosine similarity* (dengan pembobotan frekuensi kata berbasis metode *Term Frequency-Inverse Document Frequency* (TF-IDF)) dan mekanisme pencocokan kata. Regresi linear dilakukan dengan menjadikan nilai uraian singkat (hasil *cosine similarity* dan pencocokan kata) sebagai variabel *regressor*. Untuk mengetahui efektivitas sistem penilaian yang diusulkan, diukur kinerja sistem penilaian relatif terhadap nilai manual yang dilakukan oleh dosen. Diperoleh bahwa sebelum dilakukan regresi linear, sistem penilaian cenderung mengeluarkan nilai lebih tinggi (nilai mengalami bias) dibandingkan nilai manual yang dilakukan dosen. Regresi linear memperbaiki kinerja sistem penilaian tersebut dengan mengurangi bias penilaian secara signifikan, yaitu nilai yang diberikan tidak cenderung lebih tinggi maupun lebih rendah daripada nilai manual oleh dosen. Bahwa bias penilaian dapat diturunkan secara signifikan dengan metode yang sederhana, yaitu regresi linear, diharapkan dapat memberikan kontribusi terhadap akselerasi proses penerapan sistem penilaian otomatis untuk tes uraian pada teknologi pembelajaran dalam jaringan seperti *e-learning*.**

***Kata Kunci*—Sistem Penilaian Otomatis, *Cosine Similarity*, TF-IDF, *Regresi Linear*, *E-learning*.**

[1] *Departemen Teknik Elektro dan Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada, Jl. Grafika No. 2 Kampus UGM, Yogyakarta 55281 INDONESIA (telp: +62(274) 552305; e-mail: silmi@ugm.ac.id)*

I. PENDAHULUAN

Di zaman dengan layanan internet yang sudah semakin mudah diperoleh dan dengan perkembangan ilmu dan teknologi yang semakin bertambah pesat seperti sekarang ini, metode dan proses dalam dunia pendidikan telah bergeser. Di lingkungan pendidikan tinggi, misalnya, internet dan teknologi yang ada dapat digunakan untuk mengakses sumber materi belajar oleh mahasiswa maupun bahan materi ajar oleh dosen dengan lebih mudah tanpa diperlukan biaya yang mahal dan usaha yang terlalu besar. Salah satu pendekatan teknologi yang digunakan di dunia pendidikan sebagai sarana untuk membantu proses pembelajaran adalah *e-learning*. Fasilitias *e-learning* ini dapat dimanfaatkan oleh dosen untuk membagikan materi kepada mahasiswa dengan mudah dan cepat dan mengevaluasi penguasaan atau pemahaman materi oleh mahasiswa (contohnya melalui kuis). Selain itu, *e-learning* juga menyediakan tempat untuk mahasiswa mengunggah pekerjaan/ tugas dalam bentuk *soft file* sehingga memudahkan dokumentasi pembelajaran.

Saat ini, *e-learning* yang bersifat adaptif sedang menjadi tren. *E-learning* adaptif merupakan *e-learning* yang menyesuaikan sistem belajar yang dilakukan oleh mahasiswa dengan preferensi mahasiswa yang bersangkutan, meliputi tingkat kemampuan kognitif, strategi dan gaya belajar, materi yang lebih mudah dipahami, perilaku pengguna, dan lain-lain [1]. Pada proses belajar menggunakan *e-learning* tersebut, mahasiswa mendapatkan hasil evaluasi terhadap pemahaman materi dan proses belajar serta umpan balik dari proses belajar mahasiswa. Terdapat dua macam evaluasi yang dilakukan, yaitu *Prior Knowledge Activation* (PKA) dan evaluasi materi. PKA berfungsi untuk menilai tingkat pemahaman awal mahasiswa terkait materi yang akan dipelajari, sedangkan evaluasi materi berfungsi untuk menilai kemampuan pemahaman mahasiswa terhadap materi yang telah dipelajari. Kedua evaluasi tersebut biasanya menggunakan tes berupa pilihan ganda. Namun, tes berupa pilihan ganda kurang mampu menilai tingkat pemahaman materi mahasiswa sebaik tes uraian singkat. Hal ini dikarenakan tes pilihan ganda memungkinkan mahasiswa menerka-nerka jawaban yang benar dari pilihan yang disediakan. Akibat dari pilihan yang telah tersedia tersebut, peluang benar terkaan menjadi lebih tinggi daripada tes uraian singkat. Berbeda dari tes pilihan ganda, tes uraian singkat mampu meminimalkan sifat terkaan tersebut dan mampu mengukur aspek kognitif lebih tinggi sehingga mampu memberikan informasi mengenai materi yang belum dikuasai oleh mahasiswa.

Kendala dalam tes menggunakan uraian singkat ada dalam proses penilaian jawaban. Dengan jumlah mahasiswa yang banyak, dibutuhkan waktu yang tidak sedikit oleh dosen untuk menilai semua jawaban secara manual. Untuk mengatasi hal ini, dibutuhkan sebuah sistem penilaian secara otomatis untuk menilai uraian singkat pada *e-learning* tersebut.

II. *Automated Essay Scoring*

Penelitian mengenai pembuatan sistem penilaian uraian secara otomatis berbasis teks, atau yang dikenal dengan istilah *Automated Essay Scoring* (AES), sudah mulai dilakukan pada tahun 1968, dengan jumlah lebih dari 170 dan menggunakan metode yang beragam [2]. Salah satu penelitian mengenai AES menerapkan pembobotan *Term Frequency - Inverse Document Frequency* (TF-IDF) dalam proses pemberian nilai otomatis [3]. *Dataset* yang digunakan adalah hasil jawaban evaluasi berbahasa Indonesia. Pada penelitian tersebut, dilakukan pemodelan dengan mengukur kedekatan dokumen jawaban dengan kunci jawaban dalam bentuk nilai kosinus yang diperoleh dari *cosine similarity*.

Penggunaan *cosine similarity* dengan pembobotan TF-IDF sudah pernah diteliti sebelum ini untuk pengelompokan dokumen (*document clustering*) [4]. *Dataset* yang digunakan dalam penelitian tersebut adalah teks jawaban evaluasi mahasiswa dalam bahasa Inggris. Dari penelitian ini diperoleh peningkatan efisiensi pengelompokan dokumen jawaban mahasiswa dan penghematan waktu komputasi.

Penelitian mengenai AES selanjutnya telah dilakukan menggunakan beberapa tes kemiripan teks, antara lain *Longest Common Subsequence* (LCS), *cosine coefficient*, *Jaccard coefficient*, dan *Dice coefficient* yang dipadukan dengan mekanisme pencocokan kata kunci untuk sistem penilaian uraian singkat dalam bahasa Indonesia [5]. Mekanisme pencocokan kata kunci ini bertujuan untuk meminimalkan kelompok kata yang bukan merupakan poin utama penyusun jawaban. Untuk mengukur efek penambahan mekanisme pencocokan kata kunci tersebut, diukur metrik koefisien korelasi dan *Mean of Absolute Error* (MAE). Hasil dari penelitian ini menunjukkan bahwa pencocokan kata kunci mampu meningkatkan kinerja sistem penilaian untuk semua tes kemiripan yang dilakukan, dengan rata-rata peningkatan koefisien korelasi sekitar 8,4% dan penurunan MAE (peningkatan akurasi nilai) sekitar 5,5%.

Penelitian AES lain dilakukan dengan membandingkan efektivitas tes kemiripan *cosine similarity* dan *Jaccard similarity* yang dilakukan dalam dua eksperimen, yaitu tanpa prapengolahan teks dan dengan prapengolahan teks [6]. Hasil pengukuran koefisien korelasi menunjukkan bahwa prapengolahan teks mampu meningkatkan efektivitas kedua tes kemiripan tersebut. Peningkatan koefisien korelasi paling signifikan untuk metode *cosine similarity* yang diperoleh yaitu sekitar 21,6%, dibandingkan dengan metode *Jaccard similarity*, yaitu sekitar 7,5%. Sementara itu, metode *cosine similarity* memiliki koefisien korelasi yang lebih tinggi dibandingkan metode *Jaccard similarity*, baik sebelum (koefisien korelasi 0,51 dibandingkan 0,40) maupun sesudah dilakukan prapengolahan teks (0,62 dibandingkan 0,43). Efek berbagai macam teknik prapengolahan teks terhadap kinerja sistem AES diteliti lebih lanjut melalui eksperimen [7].

Berdasarkan beberapa penelitian di atas, diketahui belum ada penelitian yang mengamati kinerja objektivitas penilaian yang dilakukan oleh sistem AES. Objektivitas penilaian tersebut penting diamati untuk mengetahui penilaian yang diberikan oleh sistem AES mengalami bias relatif terhadap nilai yang diberikan manual oleh dosen atau tidak. Apabila bias bernilai positif, maka sistem penilaian cenderung memberikan nilai lebih tinggi dari nilai manual oleh dosen dan sebaliknya apabila bias bernilai negatif. Sementara itu, diinginkan nilai bias sekecil mungkin mendekati nol untuk keperluan objektivitas penilaian.

Pada makalah ini, sistem penilaian uraian singkat diusulkan dengan menggabungkan metode *cosine similarity* dan mekanisme pencocokan kata yang selanjutnya akan ditingkatkan dengan menambahkan regresi linear. Metode *cosine similarity* dengan pembobotan TF-IDF dipilih karena sudah pernah diteliti sebelum ini untuk sistem AES [3]. Terlebih lagi, *cosine similarity* berhasil memberikan kinerja yang lebih baik dalam mengukur kemiripan teks dibandingkan metode serupa lainnya [6]. Sementara itu, mekanisme pencocokan kata dipilih karena berhasil meningkatkan kinerja penilaian sistem AES berdasarkan penelitian sebelumnya [5]. Gabungan metode *cosine similarity* dan mekanisme pencocokan kata tersebut selanjutnya akan ditambah regresi linear yang motivasinya akan dibahas pada bagian IV ketika membahas hasil. Pada makalah ini juga diukur kinerja objektivitas penilaian oleh sistem, relatif terhadap nilai manual oleh dosen, sebelum dan sesudah ditambahkan regresi linear dengan menggunakan metrik bias.

III. Metodologi

Diagram alir sistem penilaian uraian singkat yang diusulkan pada makalah ini diilustrasikan pada Gbr. 1. Seperti tersaji di Gbr. 1, sistem penilaian uraian singkat yang diusulkan dijalankan melalui beberapa tahapan. Secara garis besar, tahapan-tahapan tersebut adalah pengumpulan data berupa teks jawaban dan kunci jawaban, prapengolahan teks, sistem penilaian berbasis *cosine similarity* dan pencocokan kata, serta peningkatan sistem tersebut berbasis regresi linear. Kemudian, pada tahap terakhir dilakukan perbandingan kinerja antara sistem penilaian sebelum dan sesudah dilakukan regresi linear.

A. Pengumpulan Teks Jawaban Mahasiswa dan Kunci Jawaban dari Dosen

Dataset yang digunakan pada sistem penilaian uraian singkat berupa teks kunci jawaban yang dibuat oleh dosen dan beberapa teks jawaban dari mahasiswa. Teks kunci jawaban dan teks jawaban tersebut berkaitan dengan pertanyaan tentang salah satu materi pada mata kuliah Algoritme dan Struktur Data.

B. Prapengolahan Teks

Prapengolahan teks bertujuan untuk mengubah data teks dengan struktur tidak teratur menjadi teratur [8]. *Dataset* atau teks yang tersusun dari kata dan karakter yang sangat beragam dirapikan untuk meningkatkan kualitas sistem penilaian. Untuk

Gbr. 1 Diagram alir sistem penilaian uraian singkat yang diusulkan.

keperluan ini, prapengolahan teks terdiri atas tahapan normalisasi teks, *stopword removal*, dan *stemming*.

1) Normalisasi Teks: Pada tahapan ini, semua huruf kapital pada kalimat asli diubah menjadi huruf kecil serta tanda baca dan simbol dihapus. Hasil dari normalisasi teks adalah teks yang lebih seragam, yaitu tersusun dari huruf kecil semua serta tanpa tanda baca dan simbol.

2) Stopword Removal: Pada tahapan ini, kata-kata yang termasuk dalam *stopword list* dihapus. *Stopword list* merupakan kata-kata umum yang tidak mempunyai konteks semantik yang signifikan pada suatu kalimat [9]. Beberapa contoh kata yang termasuk dalam *stopword list* adalah kata sambung dan keterangan, seperti ‘yang’, ‘namun’, ‘para’, ‘jika’, ‘ketika’, dan ‘yaitu’.

3) Stemming: Pada tahapan ini, setiap kata berimbuhan di dalam teks diubah menjadi bentuk kata dasar, sesuai definisi *stemming* itu sendiri, yaitu proses pemisahan kata dasar dari imbuhannya [10], [11]. Contohnya, kata berimbuhan ‘membantu’ diubah menjadi bentuk dasar ‘bantu’ dan ‘dirancang’ diubah menjadi bentuk dasar ‘rancang’.

C. Pembobotan Kata Berbasis TF-IDF

Pembobotan kata dilakukan dengan metode TF-IDF yang berbasis pada informasi jumlah kemunculan kata (*term frequency*) dan jumlah dokumen yang mengandung suatu kata (*document frequency*) [12]–[14]. Untuk melakukan pembobotan berbasis TF-IDF, diperlukan beberapa formula berikut

$$\begin{aligned} W_{t,d} &= TF_{t,d} \cdot IDF_{t,d}, \\ TF_{t,d} &= f(t,d), \\ IDF_{t,d} &= \log \frac{D}{DF_t} + 1 \end{aligned} \tag{1}$$

dengan $W_{t,d}$ adalah jumlah kemunculan (frekuensi) suatu kata yang telah dibobot, $TF_{t,d}$ adalah jumlah kemunculan suatu kata t di dalam suatu teks d, DF_t adalah jumlah teks yang mengandung suatu kata t, $IDF_{t,d}$ adalah *inverse document frequency*, dan D adalah jumlah teks yang digunakan.

D. Cosine Similarity dan Pencocokan Kata

1) Cosine Similarity: Pada tahapan ini, diukur tingkat kemiripan teks kunci jawaban dan teks jawaban menggunakan *cosine similarity*. Formula untuk mencari nilai *cosine similarity* γ ditulis sebagai berikut

$$\gamma = \frac{\boldsymbol{x_1^T x_2}}{\|\boldsymbol{x_1}\|\|\boldsymbol{x_2}\|} \tag{2}$$

dengan $\boldsymbol{x_1}$ dan $\boldsymbol{x_2}$ secara berturut-turut adalah vektor yang berisikan komponen-komponen bernilai frekuensi kata yang terbobot TF-IDF baik pada teks kunci jawaban dan teks jawaban. Nilai *cosine similarity* γ ini berkisar dari 0 sampai 1. Semakin mirip kedua teks, nilai *cosine similarity* semakin menuju nilai 1 dan sebaliknya.

2) Pencocokan Kata: Pada tahapan ini, terlebih dahulu dilakukan ekstraksi terhadap setiap kata yang telah melalui prapengolahan teks, yaitu apabila terdapat kata yang muncul lebih dari sekali, hanya diambil satu kata. Nilai pencocokan kata dihitung dengan menghitung rasio jumlah kata yang sama antara dua buah teks (jawaban dan kunci jawaban) dengan jumlah kata pada teks kunci jawaban.

E. Regresi Linear

Pada tahapan ini, dilakukan regresi linear dengan variabel *regressor* adalah nilai sistem hasil penghitungan dari *cosine similarity* dan pencocokan kata. Regresi linear yang diterapkan memiliki bentuk model sebagai berikut

$$\boldsymbol{y} = w_0 + w_1 \boldsymbol{u} + \boldsymbol{e} \tag{3}$$

dengan $\boldsymbol{u}$ adalah vektor *regressor* (mewakili sekumpulan nilai sistem), $\boldsymbol{y}$ adalah vektor observasi (nilai manual dari dosen), w_0 dan w_1 adalah parameter model, dan $\boldsymbol{e}$ adalah selisih aditif. Proses identifikasi parameter w_0 dan w_1 menghasilkan estimasi parameter $\hat{w}_0$ dan $\hat{w}_1$ optimal yang dihitung berbasis kriteria *least square* sebagai berikut [15]

$$\begin{aligned} \hat{\boldsymbol{w}} &= (\boldsymbol{\Phi^T \Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}, \\ \hat{\boldsymbol{w}} &= [\hat{w}_0 \quad \hat{w}_1]^T, \boldsymbol{\Phi} = [\boldsymbol{1} \quad \boldsymbol{u}] \end{aligned} \tag{4}$$

dengan $\hat{\boldsymbol{w}}$ adalah vektor estimasi parameter dan $\boldsymbol{\Phi}$ adalah matriks *regressor*.

TABEL I
PRAPENGOLAHAN TEKS UNTUK TEKS KUNCI JAWABAN (T1) DAN SALAH SATU CONTOH TEKS JAWABAN (T2)

Tahapan Prapengolahan Teks	Hasil
Kalimat asli	**T1**: Vertex yaitu node atau titik yang nantinya akan dihubungkan dengan edge. Edge, yaitu garis penghubung. loop yaitu edge yang menghubungkan vertex dengan dirinya sendiri. **T2**: Vertex adalah data atau node dari graph, sedangkan edge adalah garis yang menghubungkan antara vertex
Normalisasi teks	**T1**: vertex yaitu node atau titik yang nantinya akan dihubungkan dengan edge edge yaitu garis penghubung loop yaitu edge yang menghubungkan vertex dengan dirinya sendiri **T2**: vertex adalah data atau node dari graph sedangkan edge adalah garis yang menghubungkan antara vertex
Stopword removal	**T1**: vertex node titik dihubungkan edge edge garis penghubung loop edge menghubungkan vertex **T2**: vertex data node graph edge garis menghubungkan vertex
Stemming	**T1**: vertex node titik hubung edge edge garis hubung loop edge hubung vertex **T2**: vertex data node graph edge garis hubung vertex

F. Perbandingan Kinerja Sistem Penilaian Sebelum dan Sesudah Dilakukan Regresi Linear

Pada tahapan ini, kinerja sistem penilaian dalam menilai uraian singkat dibandingkan antara sebelum dan sesudah dilakukan regresi linear. Kinerja yang dibandingkan di sini relatif terhadap nilai manual yang diberikan oleh dosen, dengan asumsi dosen memberikan nilai secara objektif. Penilaian yang objektif tersebut dimungkinkan dengan melakukan penilaian uraian singkat berdasarkan kunci jawaban yang telah dibuat sebelumnya.

Pengukuran kinerja dilakukan secara objektif dengan menghitung dua metrik, yaitu MAE, yang memiliki informasi akurasi penilaian; dan bias, yang memiliki informasi nilai sistem cenderung lebih tinggi (*overestimation*) atau lebih rendah (*underestimation*) relatif terhadap nilai manual yang diberikan oleh dosen.

IV. HASIL DAN PEMBAHASAN

A. Prapengolahan Teks

Pada bagian ini, diambil teks kunci jawaban dan salah satu teks jawaban untuk ditunjukkan proses prapengolahan teks: normalisasi teks, *stopword removal*, dan *stemming.* Contoh prapengolahan teks tersebut disajikan pada Tabel I, dengan **T1** adalah teks kunci jawaban dan **T2** adalah teks jawaban. Pada Tabel I, dapat dicermati bahwa prapengolahan teks berhasil dilakukan. Pada tahapan normalisasi teks, huruf besar telah diganti menjadi huruf kecil dan tanda baca titik dihilangkan. Pada tahapan *stopword removal*, kata sambung dan keterangan yang kurang relevan, seperti 'yaitu', 'atau', 'yang', 'nantinya', 'akan', dan 'dengan' dihilangkan. Pada tahapan *stemming*, kata-kata yang memiliki imbuhan diubah menjadi bentuk dasar, seperti kata berimbuhan 'dihubungkan' yang diubah menjadi bentuk dasar 'hubung'.

TABEL II
HASIL PEMBOBOTAN FREKUENSI KATA BERBASIS METODE TF-IDF UNTUK TEKS KUNCI JAWABAN (T1) DAN SALAH SATU CONTOH TEKS JAWABAN (T2) YANG TELAH DILAKUKAN PRAPENGOLAHAN TEKS

Kata	TF		DF	IDF	TF – IDF	
	T1	T2			T1	T2
vertex	2	2	2	1	2	2
data	0	1	1	1,301	0	1,301
node	1	1	2	1	1	1
graph	0	1	1	1,301	0	1,301
edge	3	1	2	1	3	1
garis	1	1	2	1	1	1
hubung	3	1	2	1	3	1
titik	1	0	1	1,301	1,301	0
loop	1	0	1	1,301	1,301	0

B. Pembobotan Kata Berbasis Metode TF-IDF

Pada bagian ini ditunjukkan pembobotan kata berbasis metode TF-IDF untuk teks kunci jawaban **T1** dan teks jawaban **T2** yang telah dikenai prapengolahan teks (lihat teks **T1** dan teks **T2** di Tabel I pada baris *Stemming*). Hasil pembobotan tersaji di dalam Tabel II.

Pada Tabel II, kolom TF-IDF menunjukkan jumlah kemunculan kata (TF) yang dibobot dengan nilai IDF untuk masing-masing teks kunci jawaban **T1** dan kunci jawaban **T2**. Penghitungan TF-IDF ini dilakukan dengan menghitung $W_{t,d}$ pada (1) dengan $t \in \{\text{'vertex', 'data', 'node', ..., 'loop'}\}, d = \{\mathbf{T1}, \mathbf{T2}\}$, dan $D = 2$.

C. Penerapan Sistem dengan Cosine Similarity Pencocokan Kata untuk Penghitungan Nilai Uraian Singkat

Data pada kolom TF-IDF pada Tabel II kemudian digunakan untuk menghitung nilai *cosine similarity* dan nilai pencocokan kata.

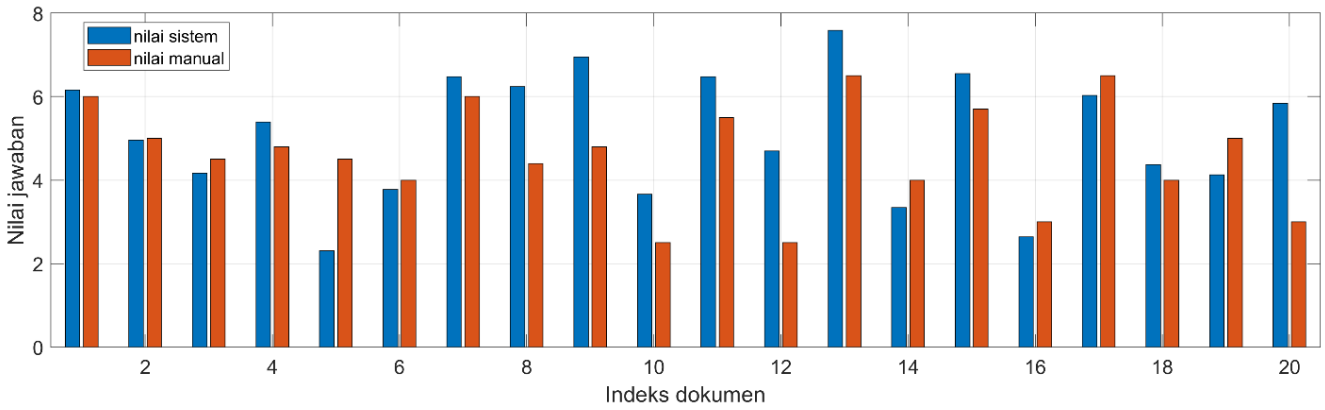
1) Cosine Similarity: Untuk mempermudah penghitungan nilai *cosine similarity*, nilai TF-IDF untuk masing-masing teks kunci jawaban **T1** dan teks jawaban **T2** pada Tabel I disajikan dalam bentuk vektor sebagai berikut

$$\boldsymbol{x_1} = [2 \quad 0 \quad 1 \quad 0 \quad 3 \quad 1 \quad 3 \quad 1{,}301 \quad 1{,}301]^T$$

$$\boldsymbol{x_2} = [2 \quad 1{,}301 \quad 1 \quad 1{,}301 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0]^T$$

dengan $\boldsymbol{x_1}$ adalah vektor TF-IDF untuk teks kunci jawaban **T1** dan $\boldsymbol{x_2}$ adalah vektor TF-IDF untuk teks jawaban **T2**. Dengan menggunakan formula *cosine similarity* pada (2), dihasilkan nilai *cosine similarity* γ sebesar

$$\gamma = \frac{\boldsymbol{x_1^T x_2}}{\|\boldsymbol{x_1}\|\|\boldsymbol{x_2}\|} = 0{,}6796.$$



Gbr. 2 Keseluruhan nilai jawaban: nilai sistem hasil *cosine similarity* dan pencocokan kata (biru) dan nilai manual dari dosen (oranye).

2) Pencocokan Kata: Nilai pencocokan kata dihitung dengan membagi jumlah kata yang sama antara teks kunci jawaban **T1** dan teks jawaban **T2** dengan jumlah kata pada teks kunci jawaban **T1** pada Tabel II.

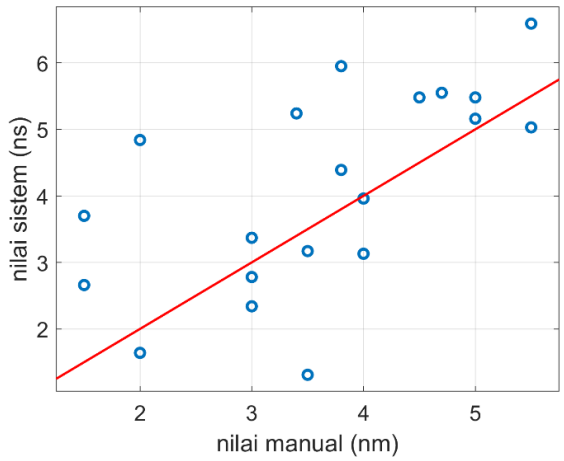
Perlu dicatat bahwa kata yang dihitung tidak memiliki perulangan. Contohnya, jumlah kata 'vertex' akan dihitung satu meskipun pada teks **T1**, kata tersebut muncul sebanyak dua kali. Untuk kasus ini, jumlah kata yang sama antara teks **T1** dan **T2** tersebut adalah lima, sedangkan jumlah kata **T1** tersebut adalah tujuh, sehingga diperoleh nilai pencocokan kata sebesar $\beta = 5/7 = 0{,}7143$.

Setelah menghitung nilai *cosine similarity* γ dan nilai pencocokan kata β, nilai akhir dihitung dengan mencari rata-rata kedua nilai γ dan β yang kemudian dikali dengan faktor 10. Penyekalaan dengan faktor 10 ini berfungsi untuk membuat nilai akhir tersebut berada di rentang nilai 0 sampai 10 (rentang nilai *cosine similarity* dan pencocokan kata adalah 0 sampai 1). Penyekalaan dengan faktor 10 ini diperlukan agar nilai akhir yang diperoleh nanti dapat dengan adil dibandingkan dengan nilai manual yang diberikan oleh dosen yang juga memiliki rentang nilai 0 sampai 10 untuk keperluan pengukuran kinerja sistem penilaian. Penghitungan tersebut menghasilkan nilai akhir 6,969. Nilai akhir yang diperoleh dari *cosine similarity* dan pencocokan kata tersebut selanjutnya disebut dengan istilah nilai sistem (nilai uraian singkat dari sistem).

D. Kinerja Sistem Penilaian dengan Cosine Similarity dan Pencocokan Kata

Setelah dilakukan prapengolahan teks, pembobotan kata berbasis metode TF-IDF, dan penghitungan nilai sistem n_s berbasis *cosine similarity* serta pencocokan kata untuk semua teks jawaban, pada bagian ini diukur kinerja nilai sistem n_s tersebut relatif terhadap nilai manual n_m yang diberikan oleh dosen. Keseluruhan nilai sistem n_s dan nilai manual n_m untuk teks ke-1 hingga teks ke-20 dibandingkan dengan diagram batang pada Gbr. 2 dan dengan *bivariate plot* pada Gbr. 3.

Pada Gbr. 2, terlihat bahwa hampir 40% total teks memiliki nilai sistem n_s yang mendekati nilai manual n_m. Sementara itu, *bivariate plot* pada Gbr. 3 menunjukkan bahwa sekitar 60% titik berada di atas garis merah, yang mengindikasikan bahwa nilai sistem n_s cenderung menilai jawaban lebih tinggi daripada nilai manual n_m.



Gbr. 3 *Bivariate plot* untuk keseluruhan nilai jawaban: nilai sistem (hasil *cosine similarity* dan pencocokan kata) vs nilai manual dari dosen.

Agar pengukuran kinerja sistem penilaian yang telah dibuat lebih objektif, dilakukan penghitungan metrik MAE dan bias. Metrik MAE dihitung untuk mengukur akurasi nilai sistem n_s, sedangkan metrik bias dihitung untuk mengukur kecenderungan penilaian sistem termasuk *overestimation* atau *underestimation*. Hasil penghitungan kedua metrik tersebut menghasilkan nilai sebagai berikut

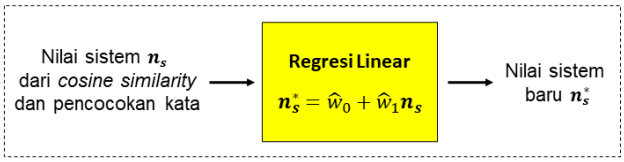
$$MAE = \frac{1}{20}\Sigma_{i=1}^{20}\left|n_{s_i} - n_{m_i}\right| = 0{,}993$$

$$Bias = \frac{1}{20}\Sigma_{i=20}^{20}\left(n_{s_i} - n_{m_i}\right) = 0{,}479.$$

Nilai MAE yang diperoleh memiliki arti bahwa secara rata-rata, selisih absolut antara nilai sistem n_s dengan nilai manual n_m adalah sebesar 0,993 poin. Sementara itu, nilai bias yang positif mengindikasikan bahwa sistem penilaian cenderung menilai jawaban lebih tinggi daripada nilai manual.

E. Peningkatan Sistem Berbasis Regresi Linear

Pada bagian ini, sistem penilaian yang telah dibuat dengan memanfaatkan *cosine similarity* dan pencocokan kata akan ditingkatkan berbasis regresi linear yang diilustrasikan melalui Gbr. 4. Motivasi dilakukannya regresi linear adalah penghitungan koefisien korelasi Pearson ρ antara nilai akhir n_s



Gbr. 4 Ilustrasi implementasi regresi linear untuk sistem penilaian dengan *cosine similarity* dan pencocokan kata.

dari sistem dengan nilai manual n_m dari dosen yang tergolong cukup baik, yaitu

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = 0{,}6228 \quad (5)$$

dengan X dan Y secara berturut-turut adalah variabel acak yang mewakili nilai sistem n_s dan nilai manual n_m, sedangkan σ_X dan σ_Y secara berturut-turut adalah deviasi standar variabel acak X dan Y. Nilai koefisien korelasi $\rho = 0{,}6228$, yang dalam makalah ini menunjukkan tingkat hubungan linear antara nilai sistem n_s dan nilai manual n_m [16], memang bukan termasuk kategori sangat baik ($\rho \geq 0{,}75$) [6], [17]. Namun, nilai tersebut dapat dikatakan cukup merepresentasikan hubungan linear karena koefisien $\rho \approx 0{,}63$ berada di atas 0,5 dan lebih dekat dengan 0,7 [18].

Pada Gbr. 4, regresi linear dapat dipandang sebagai metode untuk mengubah vektor nilai sistem $\boldsymbol{n_s}$ (yang diperoleh dengan memanfaatkan *cosine similarity* dan pencocokan kata) menjadi suatu vektor nilai sistem baru $\boldsymbol{n_s^*}$**.** Pengubahan tersebut dilakukan melalui estimasi parameter $\widehat{\boldsymbol{w}} = [\hat{w}_0 \quad \hat{w}_1]^T$ yang dihitung melalui (4) dengan vektor *regressor* $\boldsymbol{u} = \boldsymbol{n_s}$ dan vektor observasi $\boldsymbol{y} = \boldsymbol{n_m}$.

Untuk mengukur kinerja hasil regresi linear terhadap nilai sistem $\boldsymbol{n_s}$ ini, dilakukan *Leave-One-Out Cross Validation* (LOOCV) yang dipilih karena data yang berhasil diperoleh untuk penelitian ini relatif sedikit, yaitu sejumlah dua puluh teks. Sesuai namanya, sementara data/teks (berjumlah sembilan belas) yang lain digunakan untuk keperluan identifikasi parameter $\boldsymbol{w}$ melalui (4)**,** LOOCV menyisakan satu data/teks (*leave one out*) untuk keperluan validasi atau pengukuran kinerja (*cross validation*). Mekanisme tersebut akan terus diulang sampai semua teks digunakan untuk validasi sehingga akan diperoleh total dua puluh iterasi dari LOOCV tersebut. Hasil LOOCV yang diperoleh ditunjukkan pada Gbr. 5 untuk pengukuran metrik MAE dan Gbr. 6 untuk pengukuran metrik bias.

Dapat diamati pada Gbr. 5 bahwa hampir 40% total iterasi memiliki selisih absolut nilai sistem baru n_s^* dengan nilai manual n_m di bawah $MAE = 1$, meskipun selisih absolut tertinggi mencapai $MAE = 2$. Pada Gbr. 6 terlihat bahwa terdapat 45% nilai bias positif dan 55% nilai bias negatif. Untuk penilaian kinerja yang lebih objektif, dilakukan penghitungan rata-rata nilai MAE dan bias untuk semua iterasi yang selanjutnya disebut, berturut-turut, dengan MAE dan bias dari nilai sistem baru n_s^*.

Hasil penghitungan MAE dan bias nilai sistem baru n_s^* (setelah regresi linear) kemudian dibandingkan dengan MAE dan bias nilai sistem n_s (sebelum dilakukan regresi linear) sebagaimana terangkum dalam Tabel III. Terlihat bahwa

TABEL III
PERBANDINGAN KINERJA SISTEM PENILAIAN SEBELUM DAN SESUDAH DILAKUKAN REGRESI LINEAR

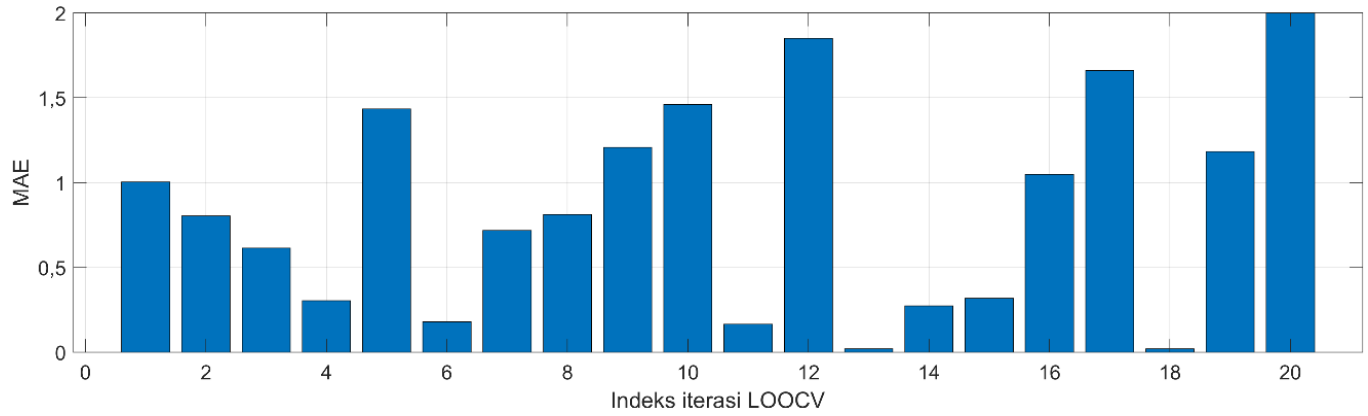
Metrik Kinerja	Sebelum Regresi Linear	Sesudah Regresi Linear
MAE	0,993	0,854
Bias	0,479	-0,014

kinerja sistem penilaian mengalami peningkatan setelah dilakukan regresi linear. Nilai MAE yang mengalami penurunan setelah dilakukan regresi linear menunjukkan bahwa akurasi sistem penilaian mengalami peningkatan, meskipun tidak terlalu signifikan. Sementara itu, nilai bias mengalami penurunan signifikan, yaitu mencapai nilai yang dekat dengan nol, yang mengindikasikan bahwa sistem cenderung menilai jawaban dengan lebih objektif (tidak cenderung lebih tinggi maupun lebih rendah) setelah dilakukan regresi linear. Adapun bias sebenarnya bernilai negatif pada Tabel III tersebut dapat ditoleransi mengingat nilai -0,014 dekat dengan nilai nol.

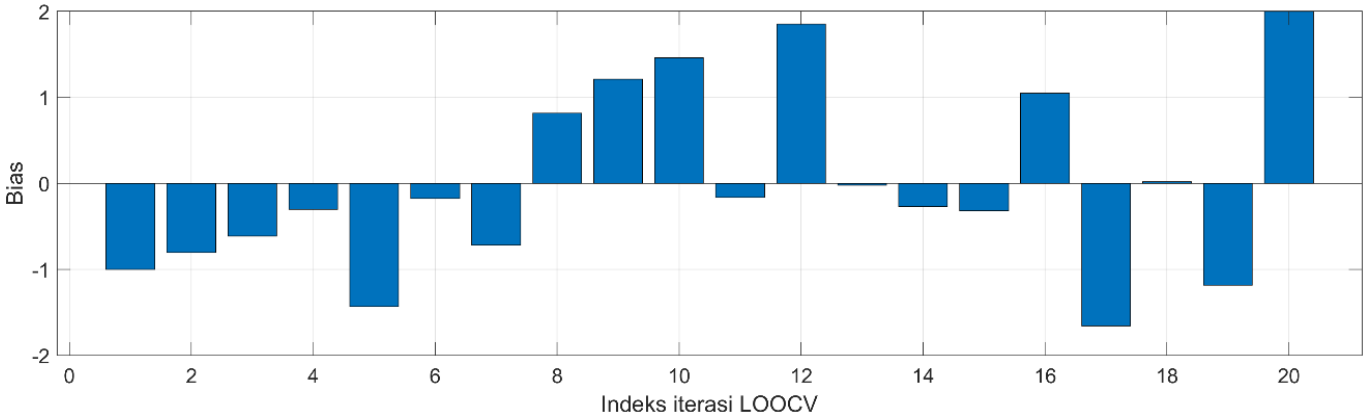
Perlu diketahui bahwa dalam makalah ini, nilai manual n_m hanya diberikan oleh satu dosen dengan asumsi bahwa penilaian tersebut dilakukan secara objektif. Penilaian secara objektif ini penting mengingat hal tersebut berpengaruh langsung terhadap nilai koefisien korelasi pada (5) yang menentukan tingkat linearitas hubungan nilai sistem n_s relatif terhadap nilai manual n_m. Penilaian manual yang tidak objektif oleh dosen dapat menurunkan konsistensi nilai manual n_m. Penurunan konsistensi nilai manual n_m ini dapat menurunkan nilai koefisien korelasi sehingga turut menurunkan tingkat linearitas hubungan kedua nilai n_s dan n_m. Hal ini akan berakibat pada penurunan relevansi penggunaan regresi linear. Selain itu, hasil pengukuran kinerja nilai sistem n_s relatif terhadap nilai manual n_m menjadi kurang valid.

Untuk mengatasi permasalahan objektivitas penilaian, penilaian manual oleh dosen sebisa mungkin mengacu pada kunci jawaban yang telah dibuat sebelumnya. Apabila uraian singkat dinilai oleh lebih dari satu dosen, penilaian oleh dosen-dosen tersebut tetap perlu mengacu pada kunci jawaban yang telah dibuat untuk meminimalkan keberagaman penilaian manual. Minimalisasi keberagaman nilai manual n_m ini dimaksudkan untuk tidak terlalu memengaruhi koefisien korelasi ρ pada (5) (tingkat linearitas hubungan kedua nilai n_s dan n_m) sehingga regresi linear masih relevan dilakukan. Selain itu, apabila jawaban uraian singkat sangat dimungkinkan beragam, dapat dibuat beberapa alternatif kunci jawaban sehingga bisa memberikan nilai sistem n_s dan nilai manual n_m yang lebih sesuai seperti yang telah dilakukan sebelumnya [5].

Sebagai informasi tambahan, koefisien korelasi untuk sistem AES dengan metode *cosine similarity* atau *cosine coefficient* cukup tinggi, yaitu antara 0,6 dan 0,7 pada penelitian lain [5], [6]. Hal tersebut mengindikasikan bahwa regresi linear dapat diterapkan untuk sistem AES pada penelitian tersebut. Selanjutnya, *framework* penambahan regresi linear yang diusulkan pada makalah ini diharapkan dapat ditambahkan pada sistem AES lain untuk menguji pengurangan bias dapat diperoleh secara signifikan seperti yang diperoleh pada penelitian ini atau tidak. Pengurangan bias sistem penilaian



Gbr. 5 Metrik MAE nilai sistem baru (setelah dilakukan regresi linear) relatif terhadap nilai manual dari dosen untuk setiap iterasi *Leave-One-Out Cross Validation* (LOOCV).



Gbr. 6 Metrik bias nilai sistem baru (setelah dilakukan regresi linear) relatif terhadap nilai manual dari dosen untuk setiap iterasi *Leave-One-Out Cross Validation* (LOOCV).

secara signifikan tersebut diharapkan mampu mengakselerasi proses diberlakukannya sistem penilaian otomatis pada sistem belajar *e-learning* untuk tes uraian singkat sehingga mempermudah tugas penilai atau dosen.

V. KESIMPULAN

Regresi linear mampu mengurangi bias sistem penilaian uraian singkat yang diusulkan pada makalah ini secara signifikan. Sistem penilaian uraian singkat tersebut menggabungkan metode *cosine similarity* dengan pembobotan TF-IDF dan mekanisme pencocokan kata. Pengurangan bias yang signifikan ini menunjukkan bahwa nilai uraian singkat yang diberikan sistem tidak memiliki kecenderungan lebih tinggi *(overestimation*) maupun lebih rendah (*underestimation*) relatif terhadap nilai manual oleh dosen sehingga meningkatkan kinerja sistem penilaian dari sisi objektivitas penilaian. Dalam melakukan regresi linear, penilaian manual oleh dosen sebisa mungkin dilakukan secara objektif, dengan cara penilaian tersebut mengacu pada kunci jawaban yang telah dibuat. Hal ini terutama berlaku apabila uraian singkat dinilai oleh lebih dari satu dosen untuk meminimalkan keberagaman nilai antar dosen. Apabila jawaban uraian singkat sangat dimungkinkan beragam, dapat ditambahkan beberapa alternatif kunci jawaban untuk memberikan nilai sistem dan nilai manual yang sesuai. Diharapkan *framework* penambahan regresi linear ini dapat diterapkan pada sistem penilaian uraian singkat lain untuk mengetahui tingkat pengurangan bias yang dihasilkan. Pengurangan bias sistem penilaian secara signifikan diharapkan mampu mengakselerasi proses diberlakukannya sistem penilaian otomatis pada sistem belajar *e-learning* untuk tes uraian singkat sehingga mempermudah tugas penilai/dosen.

REFERENSI

[1] Z. Melicheríková dan A. Busikova, "Adaptive E-learning - A tool to Overcome Disadvantages of E-learning," *Proc. Int. Conf. Emerg. eLearning Technol. Appl.*, 2012, hal. 263–266.

[2] Y. Li dan Y. Yan, "An Effective Automated Essay Scoring System Using Support Vector Regression," *Proc. - 2012 5th Int. Conf. Intell. Comput. Technol. Autom. ICICTA*, 2012, hal. 65–68.

[3] A.R. Lahitani, A.E. Permanasari, dan N.A. Setiawan, "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment," *Proc. 2016 4th Int. Conf. Cyber IT Serv. Manag. CITSM 2016*, 2016, hal. 1-6.

[4] K.P.N.V. Satya dan J.V.R. Murthy, "Clustering Based on Cosine Similarity Measure," *Int. J. Eng. Sci. Adv. Technol.*, Vol. 2, No. 3, hal. 508–512, 2012.

[5] U. Hasanah, A.E. Permanasari, S.S. Kusumawardani, dan F.S. Pribadi, "A Scoring Rubric for Automatic Short Answer Grading System," *Telkomnika (Telecommunication Comput. Electron. Control.)*, Vol. 17, No. 2, hal. 763–770, 2019.

[6] U. Hasanah dan D.A. Mutiara, "Perbandingan Metode Cosine Similarity dan Jaccard Similarity untuk Penilaian Otomatis Jawaban Pendek," *Semin. Nas. Sist. Inf. dan Tek. Inform.*, 2019, hal. 1255–1263.

[7] U. Hasanah, T. Astuti, R. Wahyudi, Z. Rifai, dan R.A. Pambudi, "An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian," *Proc. - 2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. (ICITISEE)*, 2018, hal. 230–234.

[8] K. Subbu dan V. Gurusamy, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, Vol. 5, No. 1, hal. 7–16, 2014.

[9] J.K. Raulji dan J.R. Saini, "Stop-word Removal Algorithm and Its Implementation for Sanskrit Language," *Int. J. Comput. Appl.*, Vol. 150, No. 2, hal. 15–17, 2016.

[10] V. Gupta, N. Joshi, dan I. Mathur, "Design and Development of a Rule-based Urdu Lemmatizer," *Proc. Int. Conf. on ICT for Sustain. Dev.*, 2016, hal. 161–169.

[11] A.P. Wibawa, F.A. Dwiyanto, I.A.E. Zaeni, R.K. Nurrohman, dan A. Afandi, "Stemming Javanese Affix Words Using Nazief and Adriani Modifications," *J. Inform.*, Vol. 14, No. 1, hal. 36–42, 2020.

[12] S. Qaiser dan R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, Vol. 181, No. 1, hal. 25–29, 2018.

[13] C.P. Medina dan M.R.R. Ramon, "Using TF-IDF to Determine Word Relevance in Document Queries," *New Educ. Rev.*, Vol. 42, No. 4, hal. 40–51, 2015.

[14] B. Laurensz dan E. Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 10, No. 2, hal. 118–123, 2021.

[15] R. Johansson, *System Modeling and Identification*. New Jersey, USA: Prentice Hall, 1993.

[16] K.H. Zou, K. Tuncali, dan S.G. Silverman, "Correlation and Simple Linear Regression," *J. Vet. Clin.*, Vol. 27, No. 4, hal. 427–434, 2010.

[17] J.L. Fleiss, B. Levin, dan M.C. Paik, *Statistical Methods for Rates and Proportions*, Hoboken, USA: John Wiley and Sons, 2013.

[18] D. Rumsey, *Statistics II for Dummies*. Hoboken, USA: Wiley Publishing, Inc., 2009.