

Komparasi *Term Weighting* dan *Word Embedding* pada Klasifikasi *Tweet* Pemerintah Daerah

(Comparison of *Term Weighting* and *Word Embedding* on Local Government *Tweet* Classification)

Pande Made Risky Cahya Dinatha¹, Nur Aini Rakhmawati²

Abstract—The emergence of social media encourages the government to use social media to disseminate information to its people. The information must be beneficial for the people to maintain government to citizen relationships. Classification on social media post is possible to categorize the types of posts. The study was conducted on the local government's social media accounts, yet the text processing in the research needs to be explored. Term weighting and word embedding are implemented in this research. The purpose is to compare term weighting term frequency-inverse document frequency, Okapi BM25, and word embedding doc2vec in producing features for the problem of short text classification. This study represents feature selection process, how to assess classification model, and to find the best model to overcome short text classification problem. There are six classes to categorize 1,000 short texts from 91 accounts. The measurements, i.e. precision, recall, f-1, macro-averages, micro-averages, and AUC, were calculated on each model. The result shows that the SVM linear kernel with TF-IDF performs best and slightly better than the logistic regression with 0.572 and 0.766 on macro-average recall and micro-average recall, respectively.

Intisari—Munculnya media sosial mendorong pemerintah untuk memanfaatkan media sosial sebagai sarana penyebaran informasi. Informasi yang diberikan haruslah bermanfaat bagi masyarakat dalam rangka meningkatkan hubungan *government to citizen*. Klasifikasi terhadap unggahan media sosial pemerintah daerah dapat dilakukan untuk mengetahui jenis informasi yang diunggah. Penelitian klasifikasi unggahan media sosial pada studi kasus pemerintah daerah di Indonesia telah berhasil dilakukan, tetapi pengolahan teks untuk membangun model klasifikasinya masih dapat dieksplorasi. Metode pengolahan teks yang dibahas di dalam makalah ini adalah *term weighting* dan *word embedding*. Tujuan makalah ini adalah membandingkan *term weighting term frequency-inverse document frequency*, Okapi BM25, dan *word embedding doc2vec* dalam menghasilkan fitur untuk mengatasi masalah klasifikasi teks pendek. Makalah ini merepresentasikan teks sebagai fitur untuk melakukan klasifikasi, mengetahui kinerja model klasifikasi yang telah menerapkan teknik tersebut, dan membandingkan kinerja setiap model klasifikasi untuk mengetahui metode terbaik di dalam studi kasus klasifikasi unggahan media sosial pemerintah daerah di Indonesia. Terdapat enam kelas untuk mengklasifikasi 1.000 teks pendek dari 91 akun pmda. Pengukuran *precision*, *recall*, *f-1*, *macro-average*, *micro-average*, dan AUC dilakukan pada masing-masing model. Hasil menunjukkan bahwa model TF-IDF bersama SVM linear

memberikan hasil yang lebih baik dibandingkan *logistic regression* dengan skor 0,572 dan 0,766 pada pengukuran *macro-average recall* dan *micro-average recall*.

Kata Kunci— *Classification*, *Term Weighting*, *Word Embedding*, *Social Media*, *Short Text*, *Twitter*.

I. PENDAHULUAN

Internet menjadi hal yang penting dalam mengimplementasikan pemerintahan elektronik atau yang biasa disebut *e-government*. Media sosial memiliki peran yang penting apabila dimanfaatkan oleh pemerintah untuk mewujudkan komunikasi *Government to Citizen* (G2C) dan membangun pemerintah yang lebih terbuka, partisipatif, dan inovatif melalui *open government*. Media sosial menyediakan platform untuk berbicara atau berbagi informasi dengan mudah dan murah. Pemerintah memanfaatkan media sosial Twitter untuk menyebarkan informasi ke publik. Akun yang diteliti pada makalah ini adalah akun pemerintah daerah setingkat provinsi dan kabupaten. Pemerintah setingkat ini sangat dekat dengan masyarakat dan penyebaran informasinya harus dilakukan dengan sangat baik. Makalah ini bermaksud melihat penggunaan media sosial dari sisi penyebaran informasi dengan metode pembelajaran mesin. Terdapat berbagai macam metode pembelajaran mesin dan dalam makalah ini dicari metode terbaik sesuai dengan studi kasus yang dihadapi.

Perkembangan *e-government* adalah penggunaan layanan jejaring sosial di dalam pemerintahan. *E-government* didefinisikan sebagai penggunaan atau pemanfaatan teknologi informasi oleh lembaga pemerintahan untuk meningkatkan pelayanan pada masyarakat, pelaku bisnis, dan sekaligus memfasilitasi kerja sama dengan lembaga pemerintah lainnya. Ada tiga strategi untuk membangun hubungan organisasi-publik melalui jejaring sosial atau media sosial [1]. Tiga strategi tersebut adalah *disclosure* atau keterbukaan, diseminasi informasi, dan keterlibatan. Keterbukaan berfokus pada penciptaan transparansi dan kepercayaan, diseminasi informasi memberi nilai kepada pengguna atau masyarakat seputar pemerintahan, sedangkan keterlibatan merupakan upaya untuk melibatkan pengguna di halaman organisasi. Di sisi lain, pengguna dapat memberikan informasi yang bermanfaat. Dalam konteks hubungan pemerintah-publik, pengguna dapat memberikan laporan langsung, keluhan layanan publik, pendapat kepada pemerintah, dan lain-lain.

Penelitian pada penggunaan *social network service* dalam konteks pemerintah-publik sudah dilakukan [2]. Penelitian ini dilakukan pada konten Facebook Kotamadya Karlstad. Studi

^{1,2} Departemen Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Jl. Raya ITS, Sukolilo, Surabaya, 60111, INDONESIA (tlp/fax.: 031-5999944; email: ¹risky.17052@mhs.its.ac.id, ²nur.aini@is.its.ac.id)

TABEL I
KATEGORI *POST*

Kategori	Deskripsi	Contoh Dokumen
Edukasi warga	Peningkatan pengetahuan kepada warga, membantu warga untuk meningkatkan wawasan.	jadi orang bukan kerja mudah tugas berat mulia pundak tugas bentuk generasi muda masa depan bangsa
Informasi peristiwa	Pemberitahuan atau pengumuman mengenai suatu peristiwa pada daerah tersebut.	kunjung umum kota surabaya gedung balai pemuda jalan gubernur suryo no surabaya senin-minggu pukul wib hari libur nasional tutup literasi
Promosi daerah	Mempromosikan bahwa daerah tersebut menjadi tempat yang nyaman untuk ditinggali atau bekerja, kelebihan daerah tersebut.	halo halo halo selamat pagi lur yuk dukung raih guangzhou award ken dong buat saat lur
Informasi layanan	Pemberitahuan atau informasi mengenai layanan yang diberikan atau yang dimiliki oleh daerah tersebut.	dinurahmani periksa dpt laku puskesmas cara gratis jenis periksa liput periksa fisik periksa laboratorium gol darah kadar darah merah haemoglobin sifilis hepatitis
Permintaan informasi/opini kepada warga	Permintaan pendapat kepada warga, permintaan opini kepada warga terhadap suatu hal yang bersangkutan terhadap daerah tersebut atau mengenai hal lainnya.	terima kasih saran masuk jadi bahan evaluasi
Pemberitahuan pemeliharaan	Pemberitahuan pemeliharaan kerusakan, terjadinya kerusakan atau malfungsi suatu layanan yang ada pada daerah tersebut.	pinginjadibaik pemkotsurabaya mohon maaf hasil koordinasi dinas bersih ruang buka hijau kota surabaya saat profil tank bocor sedang baik usaha hari tanggal november sdh normal sedia air terima

ini berfokus pada unggahan pemerintah dan unggahan pengguna di dinding akun pemerintah. Konten lalu dievaluasi dan dikategorikan. Ada enam kategori yang didapat, yaitu mendidik warga, pemasaran acara, pemasaran kota, pemasaran layanan, meminta informasi dari warga, dan informasi pemeliharaan layanan, seperti seperti ditunjukkan pada Tabel I. Kategori ini dijadikan sebagai kelas untuk mengklasifikasikan penggunaan media sosial pemerintah Indonesia di Twitter. Tujuan makalah ini adalah menemukan *classifier* yang paling cocok dengan studi kasus yang dihadapi. Kategori yang telah dijelaskan sebelumnya sesuai untuk diterapkan karena memiliki strategi yang sama untuk menyebarkan informasi seperti pada Instruksi Presiden Republik Indonesia nomor 3 tahun 2003 [2].

Makalah ini menerapkan metode ekstraksi fitur dan klasifikasi pembelajaran mesin. Ekstraksi fitur yang diterapkan adalah *term weighting* dan *word embedding*. Metode tersebut dapat menghasilkan fitur yang dapat digunakan sebagai data masukan untuk proses klasifikasi. Makalah ini membandingkan *term weight Term Frequency-Inverse Document Frequency (TF-IDF)*, Okapi BM25, dan *word embedding doc2vec* dalam memproduksi fitur untuk klasifikasi teks pendek menggunakan *Support Vector Machine (SVM)* dan *Logistic Regression (LR) classifier*. *Tweet* yang sudah dikumpulkan perlu diolah terlebih dahulu agar lebih mudah diproses pada tahap ekstraksi fitur dan klasifikasi. Teknik yang digunakan adalah *lowercase*, *tokenizing*, *stopword*, dan *stemming*.

Penelitian tentang komparasi seleksi fitur untuk klasifikasi teks pendek menunjukkan *term weighting* dan *word embedding* dapat memengaruhi kinerja klasifikasi [3]. TF-IDF, *doc2vec*, dan *word2vec* menunjukkan hasil yang baik pada kinerja klasifikasi. TF-IDF dapat digunakan untuk pembobotan kata dan mereduksi ukuran dokumen [4]. TF-IDF menunjukkan hasil yang baik dibandingkan dengan *bag-of-words* dan *n-grams* pada teks pendek menggunakan klasifikasi SVM dan LR

[5]. TF-IDF mendapatkan hasil yang lebih baik dibandingkan dengan *document frequency difference* dan *chi-square* [6]. Okapi BM25 memiliki nilai IDF sehingga dapat digunakan untuk membobotkan kata [7], [8]. *Doc2vec* menghasilkan *vector space model*, tetapi tidak dapat digunakan untuk mereduksi dokumen seperti pada TF-IDF dan *doc2vec*. Setiap dokumen memiliki nilai *similarity* antar dokumen, sehingga dapat digunakan sebagai masukan untuk proses klasifikasi.

Penelitian tentang klasifikasi teks pendek menggunakan SVM pada unggahan akun pemerintah daerah di Indonesia telah dilakukan [9]. Penelitian ini menerapkan kelas atau kategori yang ditunjukkan pada penelitian sebelumnya [2]. Penelitian tentang pengolahan teks masih dilakukan dan memiliki potensi untuk meningkatkan kinerja klasifikasi. Penerapan metode ekstraksi fitur dapat meningkatkan kinerja klasifikasi dan penelitian ini membandingkan kinerja klasifikasi untuk melihat metode terbaik pada studi kasus yang diangkat. Evaluasi model diukur menggunakan *precision*, *recall*, *f-1*, dan *Receiver Operating Characteristic (ROC)*. Pengukuran kinerja klasifikasi mengutamakan pengukuran *recall* karena kesalahan prediksi *false negative* lebih dapat diterima pada studi kasus yang dihadapi. Alasannya adalah terdapat *tweet* yang memiliki dua label atau lebih.

II. METODOLOGI

Tahapan penelitian dalam makalah ini dibagi menjadi beberapa bagian, yaitu *tweet scraping*, *data cleaning*, *text processing*, dan klasifikasi.

A. *Tweet Scraping*

Pengumpulan data dilakukan menggunakan *TweetScraper*, yaitu *web crawler* tanpa API Twitter. *Crawler* mendapat 1.000 *tweet* dari 91 akun media sosial pemerintah daerah sepanjang Maret 2019. *Dataset* diberi label secara manual. Terdapat enam kelas untuk membedakan setiap kategori *tweet* seperti pada

Tabel I. Kelas ini ditentukan dari penelitian tentang penggunaan Facebook di pemerintah, yaitu sebuah studi kasus informasi konten penggunaan Facebook di pemerintah. Parameter yang digunakan untuk melakukan *scraping tweet* adalah “*from*”, “*since*”, dan “*until*”. *From* mendeskripsikan akun yang dituju, sedangkan *since* dan *until* adalah kurun waktu *tweet* yang akan diambil. Contohnya adalah “*from: @SapawargaSby since:2019-04-28 until:2019-04-30*”.

B. Data Cleaning

Tahap pengolahan data dimulai dari deteksi bahasa pada *dataset*. Hanya *tweet* berbahasa Indonesia yang digunakan dalam makalah ini. Deteksi bahasa dilakukan menggunakan *library Polyglot* yang terdapat di dalam modul bahasa pemrograman Python. Proses pembersihan *dataset* dari tautan dan simbol menggunakan modul *RegEx*. Salah satu contoh *tweet* yang dibersihkan adalah “Personel Gabungan Bersihkan Lokasi Kebakaran RW 07 Krukut\n <http://barat.jakarta.go.id/v12/?p=berita&id=2748> \u00a0 \u2026 pic.twitter.com/Nay3CBfTU3” menjadi “personel gabung sih lokasi bakar rw 07 krukut”.

Tokenizing atau *tokenization* adalah proses membagi dokumen dan memasukkannya kembali berdasarkan kata yang menyusunnya. Tujuannya untuk mendapatkan karakteristik sebuah dokumen sehingga mempermudah pengolahan kata. *Casefolding* mengubah penulisan teks ke dalam bentuk *lowercase* atau penulisan huruf kecil. *Casefolding* dilakukan untuk menyetarakan setiap kata sehingga kata-kata yang sama tetapi memiliki kapitalisasi berbeda dianggap satu kata yang sama. *Stemming* menguraikan kata ke bentuk kata dasarnya. Setiap imbuhan yang melekat pada kata dihilangkan dengan tujuan untuk menyetarakan kata. Proses *tokenizing*, *casefolding*, dan *stemming* dilakukan menggunakan modul Sastrawi. Salah satu contoh proses ini adalah “Sahabat Sukabumi, kemarin, Wali Kota Sukabumi, H. Achmad Fahmi, memberikan materi kepemimpinan dan motivasi kepada 100 pemuda dari berbagai daerah di Indonesia pada acara Pemimpin Muda Nusantara. *Strategic Forum 2019* yg dilaksanakan di Jakarta. pic.twitter.com/BjQWlJaVP” menjadi “sahabat sukabumi kemarin wali kota sukabumi h achmad fahmi beri materi pimpin dan motivasi 100 pemuda dari bagai daerah di indonesia pada acara pimpin muda nusantara *strategic forum 2019* yg laksana jakarta”. Dokumen disaring menggunakan *stopword* bahasa Indonesia. Terdapat 133 kata di dalam daftar *stopword*. Kata yang dianggap tidak memiliki makna dihilangkan, seperti kata “dan”, “atau”, dan “yang”. Proses ini menggunakan modul Sastrawi. Salah satu contoh penggunaannya adalah “sahabat sukabumi kemarin wali kota sukabumi h achmad fahmi beri materi pimpin dan motivasi 100 pemuda dari bagai daerah di indonesia pada acara pimpin muda nusantara *strategic forum 2019* yg laksana jakarta” menjadi “sahabat sukabumi kemarin wali kota sukabumi h achmad fahmi beri materi pimpin motivasi 100 pemuda bagai daerah indonesia acara pimpin muda nusantara *strategic forum 2019* yg laksana jakarta”.

Dokumen yang sudah diolah sesuai dengan tahap yang dijelaskan sebelumnya dapat diproses dengan metode *term weighting* TF-IDF, Okapi BM25, atau *word embedding*

doc2vec. Setiap metode ini dapat menghasilkan *vector space model*. *Vector space model* yang dihasilkan dapat digunakan sebagai fitur untuk pelatihan dan pengujian klasifikasi.

C. Text Processing

Term weighting adalah sebuah prosedur dalam proses pengindeksan teks untuk memberikan bobot pada setiap kata yang ada di sebuah dokumen [10]. *Term weighting* memberi bobot pada kata berupa angka berdasarkan kepentingan kata tersebut di dalam dokumen untuk meningkatkan keefektifan pengambilan informasi pada sebuah sistem. Hal ini dilakukan karena tidak semua kata yang ada pada sebuah dokumen memiliki tingkat kepentingan yang sama. Proses pemberian bobot kata dipengaruhi oleh tiga faktor utama, yaitu *term frequency*, *inverse document frequency* (IDF), dan *document length*. *Term frequency* adalah jumlah kemunculan kata pada sebuah dokumen, IDF merupakan pembobotan jumlah dokumen pada korpus dan jumlah dokumen dengan kemunculan suatu kata, sedangkan *document length* adalah panjang dokumen atau jumlah kata yang ada di dokumen tersebut. TF-IDF sangat sering diimplementasikan pada penelitian terdahulu. Teknik ini menghitung kemunculan kata dan membobotkannya untuk mengurutkan kata berdasarkan tingkat kepentingannya. TF-IDF memiliki perhitungan seperti berikut [7], [8].

$$IDF = \ln \frac{1+N}{1+df(t)} + 1 \quad (1)$$

$$TFIDF = TF * IDF \quad (2)$$

$$div = \sqrt{\sum TFIDF^2} \quad (3)$$

$$TFIDF I2 = TFIDF_t / div \quad (4)$$

dengan N adalah jumlah dokumen pada korpus, $df(t)$ adalah jumlah dokumen dengan kemunculan suatu kata t , dan TF adalah jumlah kemunculan suatu kata. *TFIDF* tersebut menggunakan normalisasi $I2$ yang tersedia pada modul *scikit-learn*. Tiap kata pada dokumen menghasilkan nilai dari (4). Nilai ini ditotal untuk mendapat nilai kata dari sebuah korpus. Semakin tinggi nilai yang muncul, semakin penting kata tersebut dianggap di dalam korpus. Kata dengan nilai *threshold* di bawah 0,32 dihilangkan dari korpus. Nilai ini diambil dari pengamatan hasil *TFIDF* di *dataset*. Dokumen diubah bentuknya menjadi vektor menggunakan modul *sklearn CountVectorizer*. *Dataset* yang diolah dari proses ini sudah digunakan bersamaan dengan label untuk proses klasifikasi.

Term weighting Okapi BM25 dapat memberikan peringkat terhadap dokumen dalam urutan probabilitas relevansi terhadap *query* yang diajukan. Metode ini hampir sama dengan TF-IDF, tetapi memiliki parameter empiris yang telah ditetapkan [7], [8]. Implementasi Okapi BM25 di dalam makalah ini diubah dengan hanya menggunakan nilai *term frequency* dan perhitungan IDF yang dimiliki oleh Okapi BM25. Perhitungan IDF Okapi BM25 ditunjukkan dalam (5).

$$IDF = \ln \frac{N-df(t)+0.5}{df(t)+0.5} \quad (5)$$

Perhitungan *term frequency* pada metode ini sama dengan *TFIDF*, yaitu hanya menghitung kemunculan kata pada

dokumen. IDF pada Okapi BM25 berbeda tanpa menerapkan normalisasi (5). Kemudian, kata dengan *threshold* di bawah 0,32 dihilangkan dari korpus. Dokumen diubah menjadi vektor menggunakan modul *sklearn CountVectorizer*.

Metode *doc2vec* diperkenalkan untuk merepresentasikan dokumen teks ke bentuk numerik [11]. Metode ini dapat digunakan untuk mencari kemiripan antar dokumen. *CountVectorizer* tidak digunakan pada proses ini. Model yang dibangun menggunakan metode ini menghasilkan vektor yang dapat digunakan sebagai fitur untuk proses klasifikasi.

Proses *data cleaning* dan pengolahan teks dijelaskan secara singkat sebagai berikut. Yang pertama adalah mendeteksi bahasa Indonesia. Kemudian, tautan dan simbol yang melekat pada teks dihilangkan, lalu dilakukan *tokenizing* dan *casefolding*. Langkah selanjutnya adalah *stemming* bahasa Indonesia, lalu *stopword* diterapkan untuk menghilangkan kata yang tidak perlu. Terakhir, dilakukan proses ekstraksi fitur dengan metode *term weighting* dan *word embedding*.

D. Klasifikasi

SVM merupakan klasifikasi *supervised learning*. Konsep klasifikasi adalah melakukan pembelajaran terhadap algoritme klasifikasi dengan memasukkan data latih sebagai dasar untuk melakukan proses klasifikasi terhadap data uji, sehingga dapat melakukan pengelompokan berdasarkan ciri tertentu. Klasifikasi SVM sering digunakan untuk melakukan klasifikasi teks [3], [4], [11], [12]. Penelitian terdahulu menunjukkan *classifier* ini memiliki kinerja yang baik dan bersaing dengan *classifier* lainnya dalam mengklasifikasi teks. *Kernel* yang digunakan adalah *linear*, *polynomial*, RBF, dan *sigmoid*. *Kernel* ini dapat digunakan untuk melakukan klasifikasi *One Versus Rest* (OVR). OVR digunakan karena kelas bersifat *multiclass*.

LR dijelaskan oleh J.S. Kramer pada artikelnya yang berjudul *The Origin of Logistic Regression* [13]. LR mewakili nilai variabel dengan angka biner satu atau nol, benar atau salah. LIBLINEAR LR dengan *multiclass* (OVR) cocok digunakan pada *dataset* linear. LR mengelompokkan data dengan membaginya dalam garis lurus. Penentuan nilai parameter terbaik untuk SVM dan LR (*C*, *gamma*, dan iterasi) menggunakan *GridSearchCV* dengan *cross-validation*.

Evaluasi kinerja *classifier* dilihat dari *micro-average precision*, *macro-average recall*, *micro-average precision*, *macro-average recall*, *f-measure*, ROC, dan nilai *Area Under Curve* (AUC) [14]. Nilai AUC didapat dari area di bawah ROC, sedangkan ROC didapatkan dari penggambaran grafik *True Positive Rate* (TPR) dan *False Positive Rate* (FPR).

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \frac{Pre \times Rec}{Pre + Rec} \quad (8)$$

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

$$Macro - Precision = \frac{\sum Precision_i}{N} \quad (11)$$

$$Micro - Precision = \frac{\sum TP_i}{\sum TP_i + FP_i} \quad (12)$$

$$Macro - Recall = \frac{\sum Recall_i}{N} \quad (13)$$

$$Micro - Recall = \frac{\sum TP_i}{\sum TP_i + FN_i} \quad (14)$$

$$Macro - F1 = \frac{(Ma - Pre \times Ma - Rec)}{(Ma - Pre + Ma - Rec)} \quad (15)$$

$$Micro - F1 = \frac{(Mi - Pre \times Mi - Rec)}{(Mi - Pre + Mi - Rec)} \quad (16)$$

$$AUC = \frac{\sum_i (FPR_i - FPR_{i-1}) * TPR_{i-1} + TPR_i}{2} \quad (17)$$

Nilai *precision*, *recall*, *f-1*, TPR, dan FPR adalah hasil dari perhitungan *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) dari setiap kelas. *Precision* atau disebut juga *confidence* menunjukkan bagian dari prediksi positif yang *real positive* atau untuk mengukur ketepatan prediksi klasifikasi. *Recall* atau disebut *sensitivity* menunjukkan bagian dari *real positive* yang diprediksi positif atau menghitung banyaknya dokumen yang diprediksi secara benar pada suatu kelas. Perhitungan TPR and FPR ditunjukkan pada (9) dan (10). Nilai *N* pada (11) and (13) adalah jumlah dari setiap kelas. Penjumlahan dari *precision* dibagi jumlah kelas untuk menghitung *macro-average precision* pada (11) and penjumlahan *recall* dibagi jumlah kelas untuk menghitung *macro-average recall* pada (13). *Micro-average precision* dan *micro-average recall* dihitung menggunakan (12) dan (14). Persamaan (15) hanya menggunakan nilai *macro* untuk menghitung *macro f-1* dan (16) hanya menggunakan nilai *micro* untuk menghitung *micro f-1*. AUC dihitung dengan mengukur luas area trapesium yang ada di bawah kurva ROC, seperti pada (17).

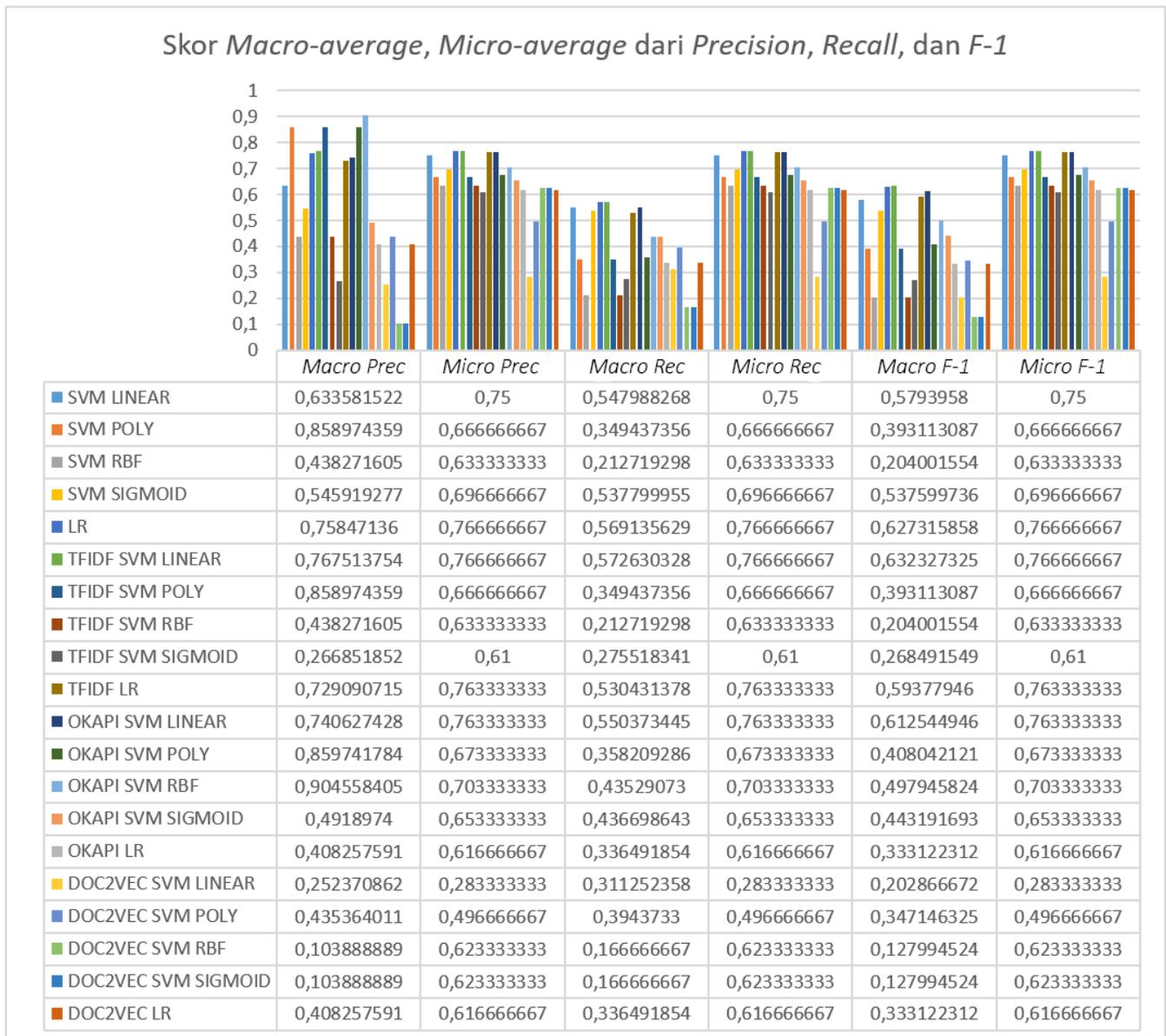
III. HASIL DAN PEMBAHASAN

A. Dataset

Dataset dibagi menjadi set pelatihan dan set pengujian. Sebanyak 70% dari *dataset* digunakan sebagai data latih dan 30% digunakan sebagai data uji. Terdapat 1.000 *tweet* pada *dataset* sehingga jumlah set pelatihan adalah 700 *tweet* dan set pengujian sebanyak 300 *tweet*. Persebaran data pada tiap kategori tidak seimbang. Persebaran *tweet* edukasi warga sebesar 10,5% atau 105 *tweet*, *tweet* informasi peristiwa sebesar 61,9% atau 619 *tweet*, *tweet* promosi daerah sebesar 6,7% atau 67 *tweet*, *tweet* informasi layanan sebesar 14,2% atau 142 *tweet*, *tweet* permintaan informasi atau opini kepada warga sebesar 4,1% atau 41 *tweet*, dan *tweet* pemberitahuan pemeliharaan sebesar 2,6% atau 26 *tweet*.

B. Pengujian Model

Terdapat dua puluh model yang diuji pada makalah ini. Lima model menggunakan metode klasifikasi seperti model SVM linear, SVM *poly*, SVM RBF, SVM *sigmoid*, dan LR. Lima belas model lainnya adalah model yang menggunakan metode pengolahan teks dan klasifikasi seperti TFIDF SVM linear dan



Gbr. 1 Skor pengukuran *macro-* dan *micro-average precision, recall, dan f-1*.

Okapi LR. Parameter pada saat dilakukan klasifikasi, seperti iterasi dan nilai C, ditentukan dari nilai tertinggi pada pengukuran *recall*. Setiap model yang diuji ditunjukkan pada Gbr. 1 dan Gbr. 2.

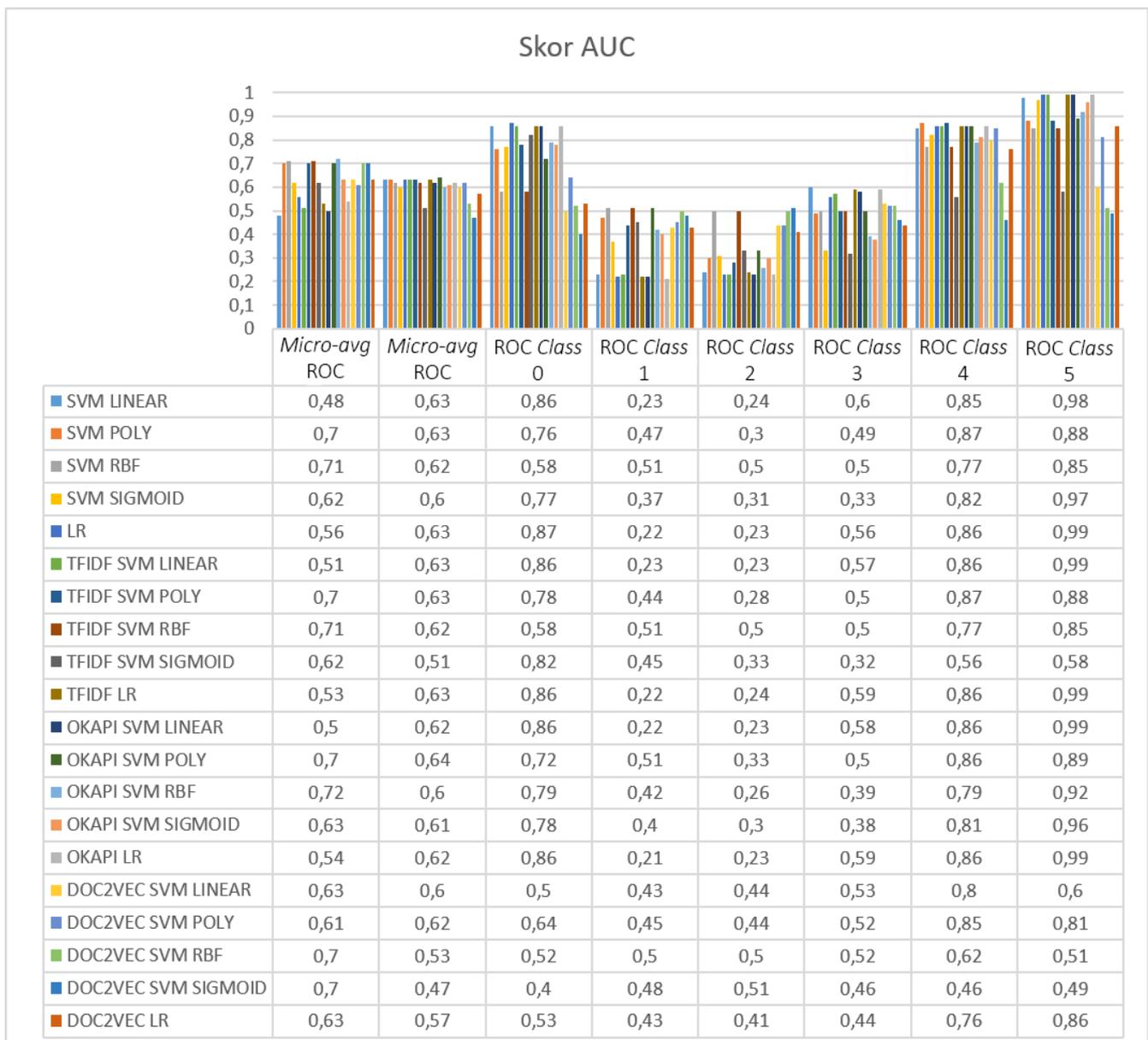
C. *Precision, Recall, dan F-1*

Pengukuran *macro* dan *micro* pada *precision, recall, dan f-1* digunakan untuk mengukur klasifikasi *multiclass*. Pengukuran ini tidak dipengaruhi oleh distribusi data yang tidak seimbang. Seperti yang terlihat pada Gbr. 1, model linear TF-IDF SVM menunjukkan hasil terbaik pada pengukuran *recall*. LR bekerja dengan baik tanpa *term weighting* dan fitur *word embedding*. *Kernel SVM* linear lebih unggul dibandingkan dengan model lainnya dengan skor 0,572 pada *macro-average recall* dan 0,766 pada *micro-average recall*. Penelitian sebelumnya juga menunjukkan SVM dengan TF-IDF dan LR mendapatkan hasil

yang lebih baik jika dibandingkan dengan *doc2vec* [3]. Okapi BM25 dalam makalah ini mendapatkan nilai tertinggi pada *macro-average recall* sebesar 0,55 dan *micro-average recall* sebesar 0,763, tetapi masih belum dapat memperoleh skor yang lebih baik dari TFIDF. *Doc2vec* mendapatkan skor yang paling buruk jika dibandingkan dengan TFIDF dan Okapi BM25, yaitu 0,394 pada *macro-average recall* dan 0,623 pada *micro-average recall*.

D. *ROC dan AUC*

Kurva ROC dapat merepresentasikan kinerja klasifikasi *multiclass* [15]. Setiap kelas memiliki nilai TPR dan FPR, sehingga perhitungan AUC untuk setiap kelas dapat dilakukan. Nilai AUC untuk setiap *classifier* ditampilkan pada Gbr. 2, dengan kelas 0 untuk edukasi warga, kelas 1 untuk kategori informasi peristiwa, kelas 2 untuk kategori promosi daerah,



Gbr. 2 Skor AUC.

kelas 3 untuk kategori informasi layanan, kelas 4 untuk kategori permintaan informasi, dan kelas 5 untuk kategori pemberitahuan pemeliharaan.

Gbr. 2 menunjukkan bahwa *kernel* SVM RBF mendapatkan skor tertinggi pada pengukuran *micro-average* ROC sebesar 0,72 dan model polinomial Okapi SVM mendapatkan skor tertinggi pada pengukuran *macro-average* ROC sebesar 0,64. TFIDF SVM linear mendapatkan skor *micro-average* ROC dan *macro-average* ROC sebesar 0,51 dan 0,63. Pengukuran ROC AUC mewakili kinerja model dari berbagai *threshold*, sehingga tidak merefleksikan nilai dari suatu *threshold* yang spesifik. ROC AUC juga bergantung pada TPR dan FPR. Beberapa model mendapatkan FPR rendah (nol *false positive*) tetapi juga memiliki kinerja lebih buruk karena model tidak dapat memprediksi sampel positif yang sebenarnya.

IV. KESIMPULAN

Di dalam makalah ini didemonstrasikan klasifikasi teks pendek pada *tweet* media sosial pemerintah daerah. Terdapat 1.000 *tweet* yang diambil dari 91 akun Twitter pemerintah daerah. *Tweet* diolah sebelum dilakukan klasifikasi dengan menghilangkan *noise*, menghilangkan tautan atau *link* dan simbol, *tokenizing*, *casefolding*, *stemming*, dan akhirnya *stopwording*. TF-IDF dan Okapi BM25 membutuhkan *CountVectorizer transformer* untuk mengonversi data tekstual menjadi data numerik. *Doc2vec* dapat menghasilkan vektornya sendiri dan memiliki fitur yang lebih kompleks dengan menghadirkan data numerik dan nilai kesamaan semantik antar dokumen. Kinerja *classifier* diukur menggunakan *precision*, *recall*, *f-1*, TPR, FPR, ROC, dan AUC. Berdasarkan hasil

percobaan dan analisis dapat diambil beberapa kesimpulan. *Threshold* terbaik pada penerapan TF-IDF adalah 0,32 dan *threshold* terbaik pada Okapi BM25 adalah 0,2. Sementara itu, *doc2vec* mendapatkan hasil yang buruk di seluruh pengukuran. Penelitian lebih lanjut diperlukan untuk menemukan pengaturan parameter terbaik untuk membangun model *doc2vec* yang lebih baik. Model TFIDF SVM linear bekerja paling baik dengan skor skor 0,572 pada *macro-average recall* dan 0,766 pada *micro-average recall*, sedangkan model LR adalah terbaik kedua dengan skor 0,569 pada *macro-average recall* dan 0,766 pada *micro-average recall*. SVM linear pada makalah ini adalah yang paling dapat diandalkan untuk melakukan klasifikasi teks pendek dan efektif terhadap model yang bersifat *linearly separable high dimensional space*. Diketahui juga bahwa *tweet* yang diunggah sangat memengaruhi klasifikasi dan kejelasan informasi yang diberikan. Dari hasil analisis, dapat dikatakan bahwa pemerintah daerah yang memanfaatkan Twitter harus memperhatikan penulisan kata pada *tweet* agar tidak disingkat dan dapat menggunakan *thread* apabila *tweet* terlalu panjang, karena penulisan yang singkat dapat memengaruhi pembobotan kata dan ciri kelas pada korpus, sehingga proses klasifikasi dan penyebaran informasi menjadi lebih jelas. Selain itu, perlu disertakan *headline* tautan atau *caption* atau judul ketika membagikan tautan, foto, atau video, agar *tweet* lebih jelas.

REFERENSI

- [1] R.D. Waters, E. Burnett, A. Lamm, dan J. Lucas, "Engaging Stakeholders Through Social Networking: How Nonprofit Organizations are Using Facebook," *Public Relations Review*, Vol. 35, No. 2, hal. 102-106, 2009.
- [2] M. Magnusson, P. Bellström, dan C. Thoren, "Facebook Usage in Government - A Case Study of Information Content," *Proceeding of the Eighteenth Americas Conference on Information Systems*, 2012, hal. 1-9.
- [3] Y. Wang, Z. Zhou, S. Jin, D. Liu, dan M. Lu, "Comparisons and Selections of Features and Classifiers for Short Text Classification," *International Conference on Artificial Intelligence Applications and Technologies 2017*, 2018, hal. 1-7.
- [4] N. Indriani, E. Rainarli, dan K.E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *Jurnal INFOTEL*, Vol. 9, No. 4, hal. 416-421, 2017.
- [5] V. Mickevicius, T. Krilavicius, dan V. Morkevicius, "Classification of Short Legal Lithuanian Texts," *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing 2015*, 2015, hal. 106-111.
- [6] T. Parlar, S.A. Özel, dan F. Song, "Interactions Between Term Weighting and Feature Selection Methods on the Sentiment Analysis of Turkish Reviews," dalam *Computational Linguistics and Intelligent Text Processing. CICLing 2016, Lecture Notes in Computer Science*, Vol. 9624, A. Gelbukh, Eds., Cham, Switzerland: Springer, 2018, hal. 335-346.
- [7] K.S. Jones, S. Walker, dan S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 1," *Information Processing and Management*, Vol. 36, No. 6, hal. 779-808, 2000.
- [8] K.S. Jones, S. Walker, dan S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2," *Information Processing and Management*, Vol. 36, No. 6, hal. 809-840, 2000.
- [9] G.K. Prakoso, "Rancang Bangun Aplikasi untuk Klasifikasi Post pada Media sosial Pemerintah Daerah di Indonesia Menggunakan Support Vector Machine (SVM)," B.IS. tesis, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, 2018.
- [10] H. Hinterberger, *Encyclopedia of Database Systems*, 1st ed., New York, USA: Springer US, 2009.
- [11] Q.V. Le dan T. Mikolov, "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014, Vol. 32, hal. II-1188-II-1196.
- [12] H. Shin dan J. Paek, "Automatic Task Classification via Support Vector Machine and Crowdsourcing," *Mobile Information Systems*, Vol. 2018 hal. 1-9, 2018.
- [13] J.S. Cramer, "The Origin of Logistic Regression," Tinbergen Institute, Amsterdam, The Netherlands, Working Paper, No. 2002-119/4, Nov. 2002.
- [14] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Flinders University of South Australia, Adelaide, Australia, Technical Report SIE-07-001, hal. 37-63, Des. 2007.
- [15] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, Vol. 27, No. 8, hal. 861-874, 2006.