Indonesian Journal of Applied Statistics ISSN 2621–086X
Volume 3 No. 2 November 2020

Implementasi *Text Mining* Pada Analisis Sentimen Pengguna Twitter Terhadap *Marketplace* di Indonesia Menggunakan Algoritma *Support Vector Machine*

Dyah Auliya Agustina[1], Sri Subanti[2], Etik Zukhronah[3]
[1,2,3]Program Studi Statistika, Universitas Sebelas Maret

[1]dyahauliya31@student.uns.ac.id, [2]srisubanti@staff.uns.ac.id, [3]etikzukhronah@staff.uns.ac.id

Abstract. In this digital era, technology development has changed the behavior of society from buy offline to online. One of this behavioral changes is marked by the growth of global marketplace including in Indonesia. The big marketplaces in Indonesia that have received a lot of public response on social media are Tokopedia, Shopee, and Bukalapak. This research determines the public sentiment toward both the service and issues surrounding these three marketplaces on media social especially Twitter. Public opinion is classified into a positive or negative sentiment. The data used in this study is obtained from Twitter API (Application Programming Interface) using keyword Shopee, Tokopedia, and Bukalapak. Preprocessing texts are divided into five steps: cleansing, case folding, stemming, stopwords, and tokenizing. Training and testing data are divided using *k*-fold cross validation method, while visualization the characteristic of text is using word cloud. Research shows that public are posting tweet more positive sentiment than negative one. The perfomance of classification shows that the best *G*-mean and AUC value for Bukalapak testing data are 0.85 and 0.86 in the first fold. While the best *G*-mean and AUC value for Shopee testing data are 0.76 and 0.77 in the seventh fold and the best *G*-mean and AUC value for Tokopedia testing data are 0.82 and 0.83 in the sixth fold.

Keywords : sentiment analysis, marketplace, support vector machine, twitter

1. Pendahuluan

Pada era digital saat ini, masyarakat dapat memanfaatkan kemudahan dan keefektifan dalam berinteraksi antara satu sama lain. Pengguna internet di dunia pada tahun 2019 sebesar 4,4 milyar pengguna sedangkan di Indonesia sebesar 107,2 juta pengguna [1]. Kegiatan *online* yang populer di Indonesia adalah penggunaan media sosial dan perpesanan seluler. Salah satu media sosial yang banyak diminati oleh masyarakat Indonesia adalah Twitter, dimana memungkinkan penggunanya untuk mengirim dan membaca pesan singkat yang disebut *tweet*.

Perkembangan teknologi juga telah mengubah perilaku masyarakat untuk berbelanja melalui *offline* ke *online*. Salah satu perubahan perilaku ini ditandai dengan pertumbuhan *marketplace* yang mengalami peningkatan di seluruh dunia termasuk Indonesia. Di Indonesia banyak terdapat *marketplace* dari skala kecil hingga skala besar. Tiga *marketplace* besar di Indonesia berdasarkan *ranking* rata-rata pengunjung setiap

kuartal, *ranking* aplikasi, pengikut media sosial, dan jumlah karyawan pada tahun 2020 diantaranya adalah Tokopedia, Shopee, dan Bukalapak.

Kajian yang dilakukan oleh Imelda dan Affandes [2] menggunakan metode *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF) untuk klasifikasi *tweet* berdasarkan kategori iklan dan kategori bukan iklan. Maulana dan Pratiwi [3] melakukan penelitian tentang pembangunan infrastruktur di Indonesia menggunakan metode *Boosting Support Vector Machine* dimana hasil klasifikasi dibuat dengan visualisasi *wordcloud* secara interaktif menggunakan aplikasi berbasis web R *Shiny*. Peneliti melakukan pengklasifikasian mengenai analisis sentimen pengguna Twitter terhadap *marketplace* di Indonesia berdasarkan dua kelas yaitu sentimen positif dan sentimen negatif. Metode yang digunakan adalah SVM karena SVM dapat diterapkan pada *tweet entity* dengan tingkat akurasi relatif lebih baik dibandingkan metode klasifikasi lainnya [4]. Penelitian dilakukan untuk mengetahui performa klasifikasi yang diberikan oleh algoritma SVM menggunakan *kernel linier* dan *kernel radial basis function*.

2. Landasan Teori

2.1. *Text Mining*. *Text mining* merupakan teknik yang digunakan untuk menangani klasifikasi, *clustering*, *information extraction*, dan *information retrival*. Perbedaan *text mining* dan *data mining* adalah pola yang digunakan *text mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur, sedangkan data mining pola yang diambil dari *database* yang terstruktur [5].

2.2. Analisis Sentimen. Analisis sentimen merupakan studi komputasi dari pengolahan bahasa alami dan komputasi linguistik dimana menganalisis pendapat, penilaian, evaluasi, sikap, emosi, dan sentimen terhadap suatu entitas seperti produk, jasa, individu, organisasi, peristiwa, topik, dan atribut lainnya. Data pengujian biasanya berupa ulasan produk secara *online* dengan menyatakan perasaan emosional sedih, senang, atau marah sehingga dapat dilihat apakah mendapatkan penilaian positif atau negatif [6].

2.3. Twitter. Twitter adalah layanan komunikasi bagi teman dan keluarga untuk tetap terhubung melalui pertukaran pesan yang cepat dan sering dengan menggunggah *tweet* yang dapat berisi foto, video, tautan, dan teks. Sedangkan Twitter *API* (*Application Programming Interface*) adalah akses programatik kepada perusahaan, pengembang, dan pengguna ke data Twitter [7].

2.4. *Marketplace*. *Marketplace* merupakan tempat melakukan kegiatan transaksi dan bisnis antara penjual dan pembeli melalui media *online* berbasis internet. *Marketplace* memiliki konsep seperti pasar tradisional dimana pemilik *marketplace* menyediakan tempat bagi para penjual dan membantu melakukan transaksi dengan lebih mudah sehingga tidak bertanggung jawab atas barang-barang yang dijual [8].

2.5. Praproses Teks. Tahap ini adalah tahapan dimana data disiapkan untuk siap dianalisis lebih lanjut dalam proses *text mining* [5]. Ada beberapa tahap dalam *preprocessing* ini, antara lain *case folding*, *cleansing*, *stemming*, *stopwords*, dan *tokenizing*.

2.6. Pembobotan Kata. *Term weighting* adalah metode pembobotan yang dilakukan untuk ekstraksi data teks dan diperlukan karena masing-masing kata yang berbeda dalam suatu dokumen memiliki tingkat kepentingan yang berbeda [9]. Selanjutnya rumus untuk menentukan bobot *Term Frequency-Inverse Document Frequency* (*TF-IDF*) dinyatakan dengan persamaan:

$$idf_t = \log \frac{D}{dft} \quad (1)$$

$$W_{t,d} = tf_{t,d} \text{ x } idf_t \quad (2)$$

dengan $W_{t,\mathrm{d}}$ adalah pembobotan *TF-IDF*, $tf_{t,d}$ adalah bobot kata *t* dalam setiap dokumen *d*, *D* adalah banyaknya dokumen, df_t adalah jumlah dokumen yang mengandung *term*, dan idf_t adalah bobot *inverse* dari nilai *df*.

2.7. *Support Vector Machine* (SVM). SVM adalah sistem pembelajaran yang menggunakan fungsi-fungsi linier pada ruang hipotesis dalam sebuah *feature space* berdimensi tinggi dan dilatih dengan algoritma pembelajaran pada teori optimisasi dengan mengimplementasikan *learning bias* yang berasal dari teori pembelajaran statistik [10]. SVM menemukan *hyperplane* dengan memaksimalkan jarak antar kelas (margin) agar memiliki kemampuan generalisasi yang tinggi terhadap data-data yang akan datang [11]. *Hyperplane* membagi himpunan data menjadi dua kelas secara sama dimana jarak antara *hyperplane* dengan objek-objek data berbeda kelas terluar (berdekatan) yang diberi warna hitam atau putih adalah sama persis, seperti diilustrasikan pada Gambar 1. Objek-objek data terluar yang paling dekat dengan *hyperplane* disebut *support vector*.

Misalkan data pada himpunan data latih dinotasikan sebagai $x_i \in R^d$ sedangkan label kelas dinyatakan sebagai $y_i \in \{-1,+1\}$ untuk i = 1,2, ..., n, dimana n adalah jumlah data.

Kedua kelas -1 dan +1 diasumsikan dapat dipisahkan secara sempurna oleh *hyperplane* berdimensi *d*, yang definisikan sebagai berikut

$$w \cdot x + b = 0 \tag{3}$$

nilai *w* dapat dihitung dengan

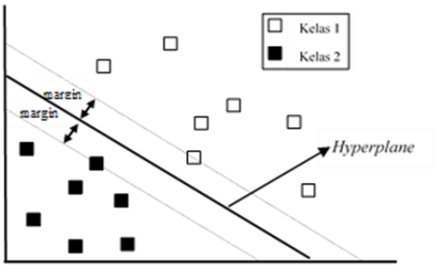
$$w = \sum_{i=1}^{n} a_i y_i x_i \tag{4}$$

dan nilai *b* dapat dihitung dengan

$$b = y_i - w^T x \tag{5}$$

Setelah menemukan vektor *w* dan skalar *b*, maka diperoleh persamaan *hyperplane* yang paling optimum. Fungsi keputusan SVM dapat dirumuskan sebagai berikut

$$f(x) = \begin{cases} -1, & \text{jika} \quad w.x + b \le -1 \\ +1, & \text{jika} \quad w.x + b > 1 \end{cases} \tag{6}$$



Gambar 1. *Hyperplane* Memisahkan Dua Kelas [11]

Formulasi tersebut menggunakan asumsi bahwa kedua kelas terpisah secara linier, namun himpunan data pada umumnya terpisah secara nonlinier. Masalah ini dapat diselesaikan dengan menggunakan konsep *kernel trick* pada ruang berdimensi lebih tinggi. Proses pembelajaran pada SVM untuk menemukan *support vector* hanya bergantung pada *dot product* dari data pada *feature space*, yaitu Φ_i, Φ_j. Umumnya, transformasi Φ tidak diketahui dan sulit dipahami sehingga perhitungan *dot product* dapat digantikan dengan fungsi kernel $K(x_i, x_j)$ yang mendefinisikan secara implisit fungsi transformasi Φ yang disebut *kernel trick*, dirumuskan dengan

$$K(x_i, x_j) = \Phi_i(x_i) \cdot \Phi_j(x_j) \tag{7}$$

Kernel Gaussian (*Radial Basis Function/RBF*) disarankan diuji pertama kali karena memiliki performansi yang sama dengan kernel linier pada parameter tertentu seperti parameter *C* [12]. Peran dari parameter *C* adalah meminimalkan kesalahan pelatihan dan mengurangi kompleksitas model. Fungsi kernel yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Fungsi Kernel Pada *SVM*

Fungsi Kernel	Rumus $K(x, x_k)$	Parameter
Linier	$K(x, x_k) = x_k^T x$	C
RBF	$K(x, x_k) = \exp\{-\gamma \mid x - x_k \mid^2 x_k^T$	γ dan C

2.8. Pembagian Data Latih dan Data Uji. *K-Fold Cross Validation* adalah metode mempartisi himpunan data *D* secara acak menjadi *k-fold* yang saling bebas f_1, f_2, ..., f_k, sehingga *fold* berisi 1/*k* pada masing-masing bagian data. Selanjutnya membangun *k* himpunan data D_1, D_2, ..., D_k yang masing-masing berisi (*k*-1) *fold* untuk data latih, 1 *fold* untuk data uji [11].

2.9. Pengukuran Kualitas Uji. Tahap ini dilakukan untuk mengetahui performa klasifikasi yang dilakukan dan didapatkan dalam sebuah *set confusion matrix* [11] yang disajikan pada Tabel 2.

Tabel 2. *Confusion Matrix*

Kelas Aktual	Kelas Hasil Prediksi		Jumlah
	Ya	Tidak	
Ya	*TP*	*FN*	*p*
Tidak	*FP*	*TN*	*N*

dengan *true positive* (*TP*) adalah tuple positif yang dilabeli dengan benar oleh *classifier*, *false negative* (*FN*) adalah tuple positif yang salah dilabeli oleh *classifier*, *true negative* (*TN*) adalah tuple negatif yang dilabeli dengan benar oleh *classifier*, dan *false positive* (*FP*) adalah tuple negatif yang dilabeli dengan benar oleh *classifier*. Dokumen yang memiliki data *balance* menggunakan akurasi untuk menghitung ketepatan klasifikasi. Selanjutnya rumus akurasi, sensitivitas, dan spesifitas disajikan sebagai berikut

$$\text{Akurasi} = \frac{TP + TN}{p + N} \tag{8}$$

$$\text{Sensitivitas} = \frac{TP}{p} \tag{9}$$

$$\text{Spesifisitas} = \frac{TN}{N} \tag{10}$$

Untuk data *imbalance*, pengukuran ketepatan klasifikasi yang digunakan adalah *G-mean* dan nilai *Area Under Curve* (*AUC*) yang merupakan indikator performansi kurva *Receiver Operating Charateristic* (*ROC*) yang dapat meringkas kinerja *classifier* menjadi

satu nilai [13]. Skala untuk interpretasi nilai *AUC* disajikan dalam Tabel 3. Berikut merupakan rumus *G-mean* dan *AUC*:

$$G\text{-}mean = \sqrt{Sensitivitas \times Spesifitas} \quad (11)$$

$$AUC = \frac{1}{2}(Sensitivitas + Spesifitas) \quad (12)$$

Tabel 3. Interpretasi Nilai *AUC*

Nilai AUC	Kinerja Model
0,5-0,6	Buruk
0,6-0,7	Cukup
0,7-0,8	Baik
0,8-0,9	Sangat Baik
0,9-1,0	Sempurna

2.10. *Word Cloud*. *Word Cloud* merupakan karakteristik suatu teks yang terdiri dari kata-kata yang banyak muncul dalam analisis. Selain itu, *word cloud* merupakan representasi grafis dari sebuah dokumen dengan *plotting* kata-kata yang sering muncul pada dokumen di ruang dua dimensi dan terkumpul seperti gumpalan awan [14].

3. Metode Penelitian

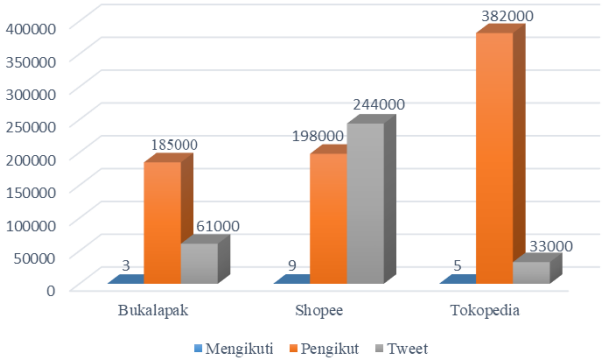
3.1 Metode Pengumpulan Data. Data yang digunakan dalam penelitian ini adalah data primer yang berupa data teks hasil *crawling* (ekstraksi pola informasi dan pengetahuan) sejumlah besar sumber data tak terstruktur melalui Twitter *API* menggunakan kata kunci Bukalapak, Shopee, dan Tokopedia pada bulan Januari-Maret 2020 sebanyak 15.000 data pada setiap kata kunci.

3.2 Langkah Penelitian. Tahapan atau langkah-langkah dalam penelitian ini adalah

1) Pengumpulan dan pelabelan data.
2) Praproses teks.
3) Pembobotan data menggunakan metode *TF-IDF*.
4) Pembagian data latih dan data uji menggunakan metode *k-fold cross validation*.
5) Klasifikasi data dengan menggunakan fungsi kernel *trick*.
6) Evaluasi hasil klasifikasi.
7) Visualisasi data menggunakan *word cloud*.
8) Kesimpulan.

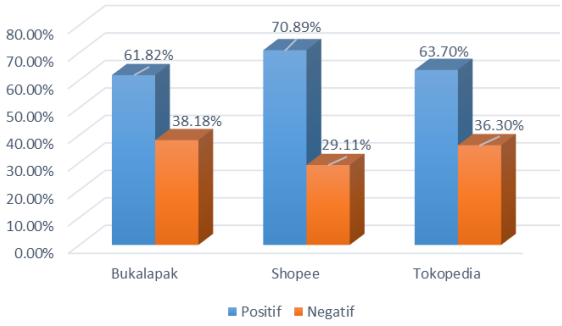
4. Hasil dan Pembahasan

4.1 Karakteristik Data. Tahap pertama sebelum dilakukan analisis yang lebih lanjut adalah analisis untuk karakteristik data. Gambar 2 merupakan perbandingan sosial media Twitter dari masing-masing *marketplace*.



Gambar 2. Grafik Perbandingan antar *Official Account* Twitter Setiap *Marketplace*

Gambar 2 menunjukkan bahwa akun Tokopedia paling banyak memiliki pengikut yaitu sebesar 382.000 pengikut karena Tokopedia memiliki *brand ambassador* yang banyak digemari anak muda. Sedangkan akun Shopee memiliki jumlah *tweet* paling banyak sebesar 244.000 *tweet* karena interaksi antara *admin* dan pengikut sangat tinggi. Setelah melakukan *crawling*, tidak semua *tweet* mengandung sentimen sehingga peneliti melakukan klasifikasi sentimen secara manual dan berdasarkan persepsi peneliti. Dari 15.000 *tweet* yang diperoleh dari setiap kata kunci didapat 3.185 *tweet* yang mengandung sentimen pada kata kunci Bukalapak, 2.360 pada kata kunci Shopee dan 4.338 *tweet* pada kata kunci Tokopedia. Selanjutnya dapat diketahui persentase *tweet* yang mengandung sentimen postif dan sentimen negatif pada Gambar 3.



Gambar 3. Perbandingan *Tweet* Sentimen Antar *Marketplace*

Gambar 3 menunjukkan *tweet* yang diperoleh pada setiap kata kunci lebih banyak mengarah ke sentimen positif daripada sentimen negatif. Selain itu, dapat diketahui bahwa perbandingan antara *tweet* yang mengandung sentimen positif dan sentimen negatif pada setiap kata kunci cenderung tidak seimbang, sehingga pengukuran ketepatan klasifikasi data pada ketiga kata kunci tersebut menggunakan *G-mean* dan *AUC*.

4.2 Praproses Teks. Sebelum dianalisis lebih lanjut, data *tweet* mengenai *marketplace* dan dijabarkan melalui simulasi pada Tabel 4 menggunakan data *tweet* Tokopedia.

Tabel 4. Simulasi Praproses Teks Data

Tahapan	Hasil Praproses
Tweet	@tokopedia @TokopediaCare ini kenapa dari pagi tadi sampe skr ga bisa2 beli tiket? Hadeehhh https://t.co/jwIImJu4Xs
Case folding	@tokopedia @tokopediacare ini kenapa dari pagi tadi sampe skr ga bisa beli tiket hadeehhh httpstcojwiimjuxs
Cleansing	ini kenapa dari pagi tadi sampe skr ga bisa beli tiket hadeehhh
Stemming	ini kenapa dari pagi tadi sampe skr ga bisa beli tiket hadeh
Stopwords & tokenizing	[‘pagi’ ‘sampe’ ‘skr’ ‘ga’ ‘bisa’ ‘beli’ ‘tiket’]

4.3 Klasifikasi Menggunakan *Support Vector Machine*

4.3.1 Menggunakan Kernel Linier. Ketepatan klasifikasi yang diberikan kernel linier untuk menentukan parameter *C* pada masing-masing *marketplace* disajikan pada Tabel 5. Berdasarkan Tabel 5 didapatkan hasil ketepatan klasifikasi untuk menentukan parameter. Pada data Tokopedia, ketepatan klasifikasi optimum menggunakan parameter *C* sebesar 100 sedangkan pada data Shopee dan Bukalapak menggunakan parameter sebesar 1. Selanjutnya dilakukan pembagian data latih dan data uji menggunakan metode *k-fold cross validation*. Pada analisis ini, rata-rata nilai akurasi data Bukalapak sebesar 83,7% dengan nilai akurasi tertinggi pada *fold* ke-1 yaitu sebesar 88%. Sedangkan rata-rata nilai akurasi data Shopee sebesar 80% dengan nilai akurasi tertinggi pada *fold* ke-7 sebesar 83% dan rata-rata nilai akurasi data Tokopedia sebesar 81% dengan nilai akurasi tertinggi pada *fold* ke-9 sebesar 85%. Perhitungan ketepatan klasifikasi berdasarkan *fold* dengan nilai akurasi tertinggi disajikan pada Tabel 6.

Tabel 6 menunjukkan akurasi pada data uji ketiga *marketplace* di atas 80%, Karena pada data ketiga *marketplace* cenderung *imbalance* maka digunakan nilai *G-mean*

dan *AUC* dimana masing-masing memiliki nilai sebesar 0,84 dan 0,85 untuk data uji Bukalapak, 0,71 dan 0,74 untuk data uji Shopee, dan 0,80 dan 0,81 untuk data uji Tokopedia, Mengacu pada Tabel 3 dapat disimpulkan bahwa kinerja klasifikasi data uji *tweet* Bukalapak dan Tokopedia dikatakan sangat baik, sedangkan untuk data uji *tweet* Shopee dikatakan baik jika ditinjau dari nilai *AUC*.

Tabel 5. Penentuan Parameter SVM

C	Ketepatan Klasifikasi		
	Bukalapak	Shopee	Tokopedia
0,01	0,621	0,712	0,,641
0,1	0,763	0,715	0,739
1	**0,851**	**0,803**	**0,868**
10	0,826	0,785	0,865
100	0,808	0,779	0,867
1000	0,808	0,779	0,867
10000	0,808	0,779	0,867

Tabel 6. Ketepatan Klasifikasi *SVM* Kernel Linier

Data		Akurasi	Sensitivitas	Spesifisitas	*G-mean*	*AUC*
Latih	Bukalapak	0,9494	0,9701	0,9159	0,9426	0,9430
	Shopee	0,9453	0,9847	0,8495	0,9146	0,9171
	Tokopedia	0,9480	0,9731	0,9000	0,9358	0,9365
Uji	Bukalapak	0,8809	0,9746	0,7295	0,8432	0,8521
	Shopee	0,8347	0,9641	0,5217	0,7092	0,7429
	Tokopedia	0,8545	0,9439	0,6824	0,8026	0,8131

4.3.2 Menggunakan Kernel *Radial Basis Function* (RBF). Parameter yang digunakan pada klasifikasi SVM kernel RBF adalah *C* dan gamma. Hasil ketepatan klasifikasi yang diberikan kernel RBF untuk menentukan parameter pada masing-masing *marketplace* disajikan pada Tabel 7.

Pada data Bukalapak dan Tokopedia, ketepatan klasifikasi optimum yang didapat menggunakan parameter *C* sebesar 10 dan parameter gamma sebesar 1. Sedangkan untuk data Shopee menggunakan parameter *C* sebesar 10 dan parameter gamma sebesar 0,1. Selanjutnya dilakukan pembagian data latih dan data uji menggunakan metode *k-fold*

cross validation. Pada analisis ini, rata-rata nilai akurasi data Bukalapak sebesar 84% dengan nilai akurasi tertinggi pada *fold* ke-1 yaitu sebesar 89%. Sedangkan rata-rata nilai akurasi data Shopee sebesar 81% dengan nilai akurasi tertinggi pada *fold* ke-7 sebesar 83% dan data Tokopedia sebesar 82% dengan nilai akurasi tertinggi pada *fold* ke-6 sebesar 86%. Selanjutnya adalah perhitungan ketepatan klasifikasi berdasarkan *fold* dengan nilai akurasi tertinggi yang disajikan pada Tabel 8.

Tabel 7. Penentuan Parameter SVM untuk Semua *Marketplace*

Marketplace	*C*	Ketepatan Klasifikasi						
		$\gamma = 1000$	$\gamma = 100$	$\gamma = 10$	$\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$
Bukalapak	0,01	0,621	0,621	0,621	0,621	0,621	0,621	0,621
	0,1	0,621	0,621	0,621	0,659	0,621	0,621	0,621
	1	0,632	0,632	0,634	0,843	0,802	0,621	0,621
	10	0,632	0,632	0,637	**0,848**	0,848	0,81	0,621
	100	0,632	0,632	0,637	0,848	0,827	0,847	0,81
	1000	0,632	0,632	0,637	0,848	0,825	0,814	0,848
	10000	0,632	0,632	0,637	0,848	0,825	0,808	0,813
Shopee	0,01	0,712	0,712	0,712	0,712	0,712	0,712	0,712
	0,1	0,712	0,712	0,712	0,712	0,712	0,712	0,712
	1	0,717	0,717	0,718	0,771	0,723	0,712	0,712
	10	0,717	0,717	0,718	0,801	**0,803**	0,729	0,712
	100	0,717	0,717	0,718	0,801	0,79	0,798	0,73
	1000	0,717	0,717	0,718	0,801	0,791	0,783	0,799
	10000	0,717	0,717	0,718	0,801	0,791	0,782	0,782
Tokopedia	0,01	0,651	0,651	0,651	0,651	0,651	0,651	0,651
	0,1	0,662	0,662	0,669	0,651	0,651	0,651	0,651
	1	0,671	0,671	0,673	0,828	0,767	0,651	0,651
	10	0,671	0,671	0,673	**0,842**	0,84	0,775	0,651
	100	0,671	0,671	0,673	0,842	0,827	0,839	0,777
	1000	0,671	0,671	0,673	0,842	0,827	0,817	0,839
	10000	0,671	0,671	0,673	0,842	0,827	0,811	0,817

Tabel 8 menunjukkan bahwa akurasi pada data uji ketiga *marketplace* di atas 80%, Data uji *tweet* Bukalapak, Shopee, dan Tokopedia cenderung *imbalance* maka digunakan nilai *G-mean* dan *AUC* sebesar 0,85 dan 0,86 untuk data uji Bukalapak, 0,76 dan 0,77 untuk data uji Shopee, dan 0,82 dan 0,83 untuk data uji Tokopedia, Mengacu

pada Tabel 3 dapat disimpulkan bahwa kinerja klasifikasi data uji *tweet* Bukalapak dan Tokopedia dikatakan sangat baik, sedangkan untuk data uji *tweet* Shopee dikatakan baik jika ditinjau dari nilai *AUC*.

Tabel 8. Ketepatan Klasifikasi SVM Kernel RBF

Data		Akurasi	Sensitivitas	Spesifisitas	*G-mean*	*AUC*
Latih	Bukalapak	0,9990	1,0000	0,9973	0,9986	0,9986
	Shopee	0,9940	0,9934	0,9612	0,9771	0,9773
	Tokopedia	0,9840	0,9996	0,9993	0,9994	0,9994
Uji	Bukalapak	0,8871	0,9746	0,7459	0,8526	0,8603
	Shopee	0,8347	0,9222	0,6232	0,7581	0,7727
	Tokopedia	0,8618	0,9404	0,7114	0,8179	0,8259

4.4 Perbandingan Klasifikasi SVM Menggunakan Kernel Linier dan Kernel *Radial Basis Function* (RBF). Berdasarkan analisis klasifikasi SVM menggunakan kernel linier dan kernel RBF diperoleh perbedaan parameter dan hasil ketepatan klasifikasi masing-masing kernel. Perbandingan rata-rata nilai akurasi dan *AUC* pada data uji pada kernel linier dan kernel RBF disajikan pada Tabel 9.

Tabel 9. Perbandingan SVM Kernel Linier dan Kernel RBF Pada Data Uji

Marketplace	Parameter	Kernel	Akurasi	*G-mean*	*AUC*
Bukalapak	$C = 1$	Linier	0,880	0,843	0,852
	$C = 10, \gamma = 1$	*RBF*	**0,887**	**0,852**	**0,860**
Shopee	$C = 1$	Linier	0,834	0,709	0,743
	$C = 10, \gamma = 0{,}1$	*RBF*	**0,835**	**0,758**	**0,773**
Tokopedia	$C = 1$	Linier	0,854	0,803	0,813
	$C = 10, \gamma = 1$	*RBF*	**0,861**	**0,818**	**0,826**

Tabel 9 menunjukkan bahwa perbandingan ketepatan klasifikasi data uji terbaik pada masing-masing *marketplace* menggunakan SVM kernel RBF memiliki nilai yang lebih tinggi daripada menggunakan SVM kernel linier sehingga dapat disimpulkan bahwa klasifikasi menggunakan SVM kernel RBF lebih baik daripada menggunakan SVM kernel linier.

4.5 Visualisasi *Word Cloud.* Visualisasi dengan *word cloud* untuk masing-masing sentimen pada data *tweet* Bukalapak yang disajikan pada Gambar 4.

(a) (b)

Gambar 4. *Word Cloud* Bukalapak Sentimen Positif (a) dan Sentimen Negatif (b)

Pada sentimen positif, kata kunci terbanyak yang digunakan adalah kata ‘bantu’. Hal ini dikarenakan para pelanggan Bukalapak di media sosial Twitter saling membantu dalam permainan potong harga yang diadakan oleh Bukalapak. Sedangkan pada sentimen negatif, kata kunci terbanyak adalah kata ‘mohon’. Hal tersebut dikarenakan pihak Bukalapak meminta maaf atas kendala yang dilalui para pelanggan Bukalapak di media sosial Twitter.

Visualisasi dengan *word cloud* untuk masing-masing sentimen pada data *tweet* Shopee disajikan pada Gambar 5.

(a) (b)

Gambar 5. *Word Cloud* Shopee Sentimen Positif (a) dan Sentimen Negatif (b)

Pada sentimen positif, kata kunci terbanyak adalah kata ‘jual’. Hal ini dikarenakan jika para pedagang di media sosial Twitter menawarkan produk yang dijual, maka respon yang paling banyak dari pengguna media sosial Twitter lain adalah menanyakan apakah produk tersebut juga tersedia di Shopee. Sedangkan pada sentimen negatif, kata kunci terbanyak adalah kata ‘beli’. Hal tersebut dikarenakan barang yang ditawarkan di Shopee banyak yang berasal dari luar negeri antara lain Cina dan Korea. Data *tweet* Shopee dikumpulkan dari bulan Januari 2020 hingga Maret 2020 dimana pada rentang waktu tersebut hampir seluruh dunia mengalami pandemi Covid-19 sehingga para pelanggan Shopee khawatir atas barang yang datang dari Cina.

Visualisasi dengan *word cloud* untuk masing-masing sentimen pada data *tweet* Tokopedia disajikan pada Gambar 6.

(a) (b)

Gambar 6 *Word Cloud* Tokopedia Sentimen Positif (a) dan Sentimen Negatif (b)

Pada sentimen positif, kata kunci terbanyak yang digunakan adalah kata ‘beli’. Berbeda dengan Shopee, kata beli dalam Tokopedia mendapat sentimen positif dikarenakan pada masa pandemi Covid-19 para pengguna media sosial Twitter dapat dengan mudah membeli barang yang dibutuhkan setiap orang pada masa pandemi yaitu *hand sanitizer*, masker, dan vitamin. Sedangkan pada sentimen negatif, kata kunci terbanyak yang digunakan adalah kata ‘kendala’. Hal tersebut dikarenakan pihak Tokopedia cukup sering mengalami kendala seputar transaksi dan pengiriman barang yang tertunda karena adanya Pembatasan Sosial Berskala Besar (PSBB).

5. Kesimpulan

Berdasarkan analisis yang telah dilakukan disimpulkan bahwa *tweet* yang diperoleh dengan kata kunci Bukalapak, Shopee, dan Tokopedia lebih banyak mengarah ke sentimen positif daripada sentimen negatif, Selain itu, performa klasifikasi menunjukkan nilai *G-mean* dan *AUC* terbaik untuk data uji Bukalapak sebesar 0,85 dan 0,86 pada *fold* pertama, untuk data uji Shopee sebesar 0,76 dan 0,77 pada *fold* ke tujuh dan untuk data uji Tokopedia sebesar 0,82 dan 0,83 pada *fold* ke enam. Secara keseluruhan, klasifikasi menggunakan kernel RBF lebih baik dibandingan menggunakan kernel linier.

Daftar Pustaka

[1] Statista, *databooks.katadata.co.id*, diakses pada 20 Januari 2020.
[2] Imelda dan Affandes, M. Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) pada Klasifikasi Tweet. *Jurnal Sains, Teknologi, dan Industri*. Vol. 12, No. 2. 2015.

[3] Maulana, A. and Pratiwi, H. Sentiment Analysis of Public Towards Infrastructure Development in Indonesia on Twitter Media Using Boosting Support Vector Machine Method. International Conference on Science (ICSAS). *AIP Conf. Proc.* 2202, 020082-1-020082-13. 2019.

[4] Joachims, T. Text Categorization with Support Vector Machine: Learning with Many Relevant Features. *Proceedings of The 10th European Conference on Machine Learning*. Pages 137-142. 1998.

[5] Feldman, R. and Sanger, J. *The Text Mining Handbook: Advance Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York. 2007.

[6] Liu, B. *Handbook of Natural Language Processing 2nd Edition*. CRC Press, Boca Raton. 2010.

[7] Twitter Help, http://help.twitter.com/, diakses pada 21 Januari 2020.

[8] Sofiani, I. dan Nurhidayat, A. I. Rancang Bangun Aplikasi E-Marketplace Hasil Pertanian Berbasis Website dengan Menggunakan Framework Codeigniter. *Jurnal Managemen Informatika*, Vol. 10, No. 01. 2019.

[9] Robertson, S. Understanding Inverse Document Frequency: On Theoritical Arguments for IDF. *Journal of Documentation*, Vol. 60, No. 5, 510. 2004.

[10] Christianini, N. and Taylor, J. *An Introduction to Support Vector Machine and Other Kernel-based Learning Methods.* Cambridge University Press, Cambridge.

[11] Suyanto. 2018. *Data Mining untuk Klasifikasi dan Klasterisasi Data Edisi Revisi.* Informatika, Jakarta. 2000.

[12] Gunn, S. R. *Support Vector Machine for Classification and Regression*. University of Southampton, Southampton. 1998.

[13] Bekkar, M., Djemaa, H. K., and Alitouch, T. A. Evaluation Measure for Models Assesment over Imbalanced Data Sets. *Journal of Information Engineering and Aplications*, 3, 2738. 2013.

[14] Castella, Q. and Sutton, C. Word Storm: Multiples of Word Clouds for Visual Comparison of Documents. *International World Wide Web Conference Committee* (*IW3C2*). ACM 978-1-4503-2744-2/14/04. 2014.