ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 12, Nomor 4, Tahun 2023, Halaman 560 - 569
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS SENTIMEN KEBIJAKAN PENYELENGGARA SISTEM ELEKTRONIK LINGKUP PRIVAT MENGGUNAKAN *PENALIZED LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE*

Nur Afnita Amalia[1*], Iut Tri Utami[2], Yuciana Wilandari[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** : afnitaamalia12@students.undip.ac.id

DOI: 10.14710/J.GAUSS.12.4.560-569

Article Info:
Received: 2023-03-25
Accepted: 2024-07-01
Available Online: 2024-07-10

Keywords:
Electronic System Operator; Sentiment Analysis; Penalized Logistic Regression; Support Vector Machine.

Abstract: The implementation of the Electronic System Operator (ESO) regulation, which imposes blocking sanctions on several ESOs that do not register, has caused a variety of opinions from the public, especially on social media Twitter to raise the hashtag #BlokirKominfo. In this research, sentiment analysis was carry outed to determine the response of Twitter users to the implementation of ESO regulations by MoCI. Sentiment analysis is a textual information extraction process that classifies sentiment into positive and negative categories. The steps that are used including crawling data, text preprocessing, labeling, feature selection, term weighting with TF-IDF and classification using the Penalized Logistic Regression (PLR) with the L1 regularization and Support Vector Machine (SVM) with the RBF kernel. Sentiment classification in PLR is basically finding the optimal weight parameter. The idea of SVM sentiment classification is to find the best hyperplane to separate the data points. Evaluation of classification performance uses the accuracy value calculated through the confusion matrix. The highest percentage of accuracy in sentiment classification results using the PLR is 84,12% and SVM is 83,53%. It means that the PLR algorithm works better than the SVM algorithm in classifying public sentiment towards the implementation of ESO regulations on Twitter.

1. PENDAHULUAN

Regulasi Peraturan Menteri Komunikasi dan Informatika Republik Indonesia Nomor 5 Tahun 2020 tentang Penyelenggara Sistem Elektronik Lingkup Privat mewajibkan pendaftaran Sistem Elektronik oleh setiap PSE sebelum melakukan penawaran digital atau melakukan bisnis di Indonesia. Pendaftaran Sistem Elektronik merupakan wujud komitmen dari Penyelenggara Sistem Elektronik (PSE) untuk bekerja sama dengan Pemerintah memberikan perlindungan yang lebih efektif bagi pengguna internet, termasuk perlindungan konsumen, perlindungan informasi pribadi pengguna dan perlindungan ruang digital yang lebih aman dan produktif (KemKominfo, 2022). Sementara bagi para PSE Lingkup Privat yang tidak terdaftar setelah mendapat peringatan, selanjutnya diterapkan sanksi berupa pemutusan akses atau pemblokiran (Rahmawati, 2022). Pemblokiran yang dilakukan oleh Kemkominfo kepada sejumlah perusahaan PSE menarik perhatian publik dan memicu beragam opini publik terkait pemblokiran tersebut, khususnya pada media sosial *Twitter* sehingga menaikkan tagar #BlokirKominfo.

Twitter merupakan sebuah layanan *microblogging* dan jejaring sosial yang digunakan untuk mengirim pesan (*tweet*) serta menjadi ruang untuk interaksi sesama pengguna (Ihsan *et al.*, 2021). *Twitter* menjadi media sosial dengan peringkat lima teratas yang banyak digunakan oleh masyarakat Indonesia (Rezeki *et al.*, 2020). Pengguna media sosial *Twitter* di Indonesia sebanyak 19,5 juta pengguna dari total 500 juta pengguna global (Ihsan *et al.*, 2021). Indonesia sebagai negara demokrasi menempatkan rakyat untuk berperan dalam proses pengambilan keputusan oleh pemerintah, sehingga tanggapan

masyarakat juga perlu diperhatikan sebagai bentuk evaluasi kinerja pemerintah dan jajarannya. Identifikasi pro dan kontra melalui analisis sentimen terhadap *tweet* mengenai peraturan PSE Lingkup Privat perlu dilakukan untuk mencerminkan tanggapan publik tentang implementasi peraturan tersebut. Analisis Sentimen merupakan salah satu bidang penelitian pada *text mining* dengan tujuan untuk mengekstrak atribut dari sebuah komentar (opini, sentimen, dan emosi) yang disampaikan secara tekstual untuk mendapatkan informasi (Wardhani *et al*., 2020).

Logistic Regression merupakan salah satu algoritma yang dapat digunakan untuk melakukan klasifikasi sentimen. Menurut Molnar (2019) implementasi algoritma *Logistic Regression* dalam mengklasifikasikan sebuah dataset dapat dikatakan cukup signifikan karena sekaligus memiliki kemampuan untuk memberikan probabilitas. Pada beberapa kasus klasifikasi menggunakan *Logistic Regression* yang mengalami *overfitting* akan memberikan performa klasifikasi yang buruk, sehingga dapat digunakan algoritma *Penalized Logistic Regression* sebagai alternatif. *Penalized Logistic Regression* merupakan metode *supervised learning* yang sesuai untuk data berdimensi tinggi karena secara implisit melakukan pemilihan fitur dengan menambahkan kendala pada persamaannya (Kassambara, 2019).

Pada teknik *supervised learning*, algoritma *Support Vector Machine* merupakan algoritma yang banyak dipakai karena memiliki akurasi klasifikasi yang baik (Alsaeedi dan Khan, 2019). Algoritma *Support Vector Machine* bekerja dengan mencari *hyperplane* terbaik sebagai pembatas antara kelas data di ruang input (Nugroho, 2007 dalam Prasetyo, 2014). Menurut Alsaeedi dan Khan (2019) algoritma SVM terbukti bekerja lebih efektif dalam berbagai permasalahan klasifikasi teks seperti pada analisis sentimen. Penelitian ini membahas mengenai analisis sentimen kebijakan PSE Lingkup Privat pada *Twitter* dengan algoritma *Penalized Logistic Regression* dan *Support Vector Machine* yang bertujuan untuk membandingkan hasil kinerja klasifikasi sentimen dengan kedua algoritma tersebut.

2. TINJAUAN PUSTAKA

Twitter adalah sebuah aplikasi *online* yang memungkinkan antar pengguna untuk berkomunikasi dan berinteraksi dalam jaringan sosial. *Twitter* menjadi salah satu aplikasi yang banyak diminati karena penggunaan *Twitter* cukup mudah dan sederhana. *Tweet* merupakan istilah pada *Twitter* yang berarti teks yang ditulis oleh pengguna baik berupa informasi berita, ungkapan ekspresi, ungkapan aspirasi atau opini mengenai topik tertentu atau hal yang menjadi perbincangan utama (Darwis *et al*., 2020).

Sentiment Analysis, juga dikenal sebagai *Opinion Mining* merupakan bidang ilmu yang mempelajari opini, sentimen, evaluasi, penilaian, sikap dan perasaan tentang sesuatu hal seperti barang, pelayanan, seseorang, organisasi, peristiwa dan topik tertentu (Liu, 2012). Sentimen dapat didefinisikan sebagai perasaan, pandangan atau pendapat yang diungkapkan seseorang baik secara tertulis maupun lisan berupa pendapat positif atau negatif (Tyagi dan Sharma, 2018). *Text Mining* merupakan sebuah teknik untuk memecahkan masalah dalam klasifikasi, pengelompokan, penyajian dan ekstraksi informasi (Rahutomo *et al*., 2018). *Text Preprocessing* merupakan tahap pertama dari pengolahan teks yang mengubah dokumen menjadi data terstruktur sesuai dengan kebutuhan analisis untuk diproses lebih lanjut dalam proses *Text Mining.* Tahapan *text preprocessing* diantaranya adalah *case folding, cleansing, normalize* dan *remove duplicate* (Testiana dan Erlina, 2022).

Pelabelan data teks menjadi sentimen positif ataupun negatif dilakukan dengan menggunakan kamus *lexicon*, yaitu kamus *Indonesia Sentiment* (InSet), *Sentistrength, boosterwords* dan negasi. Kamus InSet yang dikembangkan oleh Koto dan Rahmaningtyas (2018) memberikan nilai polaritas sentimen secara manual yang diperkuat dengan membandingkan kata sinonim pada suatu kata tersebut. Kamus *Sentistrength* yang

dikembangkan oleh Wahid dan Azhari (2016) memberi nilai polaritas sentimen secara manual yang dikerjakan oleh kelompok ahli linguistik Universitas Gadjah Mada. Kedua kamus tersebut memiliki dua kelas sentimen positif dan negatif yang mana memiliki nilai polaritas diantara -5 (sangat negatif) dan +5 (sangat positif). Proses pengurangan dimensi dalam sebuah dokumen teks dengan cara menghilangkan kata yang tidak sesuai agar meningkatkan efektifitas dan akurasi dalam proses klasifikasi disebut *Feature Selection*. Tahapan *Feature Selection* terdiri dari proses *Stopword Removal, Stemming* dan *Tokenizing* (Testiana and Erlina, 2022).

Term weight atau dapat disebut dengan pembobotan kata ditujukan untuk memberikan bobot setiap kata (*term)* pada dokumen teks. *Term Frequency-Inverse Document Frequency* (TF-IDF) menjadi metode yang paling umum digunakan untuk melakukan pembobotan kata (Sabrila *et al.*, 2022).

$$W_{j,i} = \frac{n_{j,i}}{\sum_p n_{p,i}} . log_2 \left(\frac{D}{d_j}\right) \quad (1)$$

dengan $W_{j,i}$ merupakan bobot TF-IDF *term* ke-j dalam dokumen ke-i. $n_{j,i}$ merupakan jumlah *term* ke-j yang diberikan dalam dokumen ke-i. p merupakan banyaknya *term* yang terbentuk, $\sum_p n_{p,i}$ adalah jumlah seluruh *term* yang ada pada dokumen ke-i. D merupakan jumlah keseluruhan dokumen teks, sedangkan jumlah dokumen dengan *term* ke-j disimbolkan sebagai d_j.

Klasifikasi dengan model *Logistic Regression* bukan menggunakan garis lurus atau *hyperplane*, melainkan menggunakan fungsi logistik yang memprediksi dua nilai maksimum antara 0 dan 1. Fungsi logistik untuk proses klasifikasi adalah sebagai berikut (Andrew, 2004):

$$p(y = 1|\boldsymbol{x}_i; \boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{w}^T x_i}} \quad (2)$$

dengan $\boldsymbol{w} \in \mathbb{R}^p$ adalah parameter model berupa vektor bobot. Pada dasarnya klasifikasi menggunakan regresi logistik adalah optimalisasi parameter $\boldsymbol{w}$ yang menemukan solusi dari persamaan optimalisasi berikut (Andrew, 2004):

$$arg \max_{\boldsymbol{w}} \sum_{i=1}^{m} \log p(y_i|\boldsymbol{x}_i; \boldsymbol{w}) \quad (3)$$

Nilai $\boldsymbol{w}$ sebagai parameter dalam model dapat dioptimalisasikan dengan meminimumkan *logloss error function*. Fungsi tersebut pada dasarnya memberi tahu seberapa jauh hasil estimasi dari nilai sebenarnya.

$$J(\boldsymbol{w}) = -\frac{1}{m}\sum_{i=1}^{m}\{y_i \log \sigma(\boldsymbol{w}^T \boldsymbol{x}_i) + (1 - y_i) \log[1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)]\} \quad (4)$$

Model regresi logistik klasik memiliki performa yang buruk ketika berada dalam situasi dimana terdapat kumpulan variabel dalam jumlah besar yang melebihi jumlah sampelnya. Alternatif yang dapat dilakukan adalah dengan menggunakan *Penalized Logistic Regression* yang memungkinkan untuk membuat model yang terpenalti dengan menambahkan kendala pada persamaannya (Kassambara, 2019). *Penalized Logistic Regression* merupakan metode *supervised learning* yang sesuai untuk data berdimensi tinggi karena secara implisit melakukan pemilihan fitur. Metode *Penalized Logistic Regression* diperoleh dengan menambahkan penalti pada fungsi log likelihood negatif. Bentuk penalti yang umum digunakan adalah *L1 regularization* yang diperkenalkan oleh Tibshirani pada tahun 1996 (Khoerunisa, 2021). Metode tersebut bekerja dengan menyusutkan dan menyeleksi regresi dengan mereduksi koefisien menjadi nol atau mendekati nol, sehingga diperoleh parameter yang lebih kecil dan mudah untuk diinterpretasikan. Fungsi log likelihood pada metode *Penalized Logistic Regression* sebagai berikut (Khoerunisa, 2021):

$$PLR = -\sum_i \{y_i \log \sigma(\boldsymbol{w}^T \boldsymbol{x}_i) + (1 - y_i) \log[1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)]\} + \lambda P(\boldsymbol{w}) \quad (5)$$

dengan $P(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_{j=1}^{p} |w_j|$ yang merupakan penalti *L1 regularization.* Nilai λ dipilih secara simetris dan meningkat pada $[0, +\infty]$ (Pekhimenko, 2006).

Support Vector Machine (SVM) adalah metode pembelajaran *supervised learning* yang dikembangkan oleh Boser, Guyon dan Vapnik yang pertama kali dipresentasikan pada tahun 1992 dalam acara *Annual Workshop on Computational Learning Theory* (Testiana dan Erlina, 2022). Menurut Nugroho (2007) dalam Prasetyo (2014) dijelaskan bahwa secara sederhana ide klasifikasi menggunakan metode SVM adalah menemukan *hyperplane* terbaik yang berperan sebagai pemisah antara dua kelas data dalam ruang input.

Algoritma *Support Vector Machine* merupakan *hyperplane* linear sehingga hanya bekerja pada data yang dapat diklasifikasikan secara linear (Prasetyo, 2014). Pengklasifikasian tersebut kemungkinan terlalu terbatas untuk digunakan dalam praktiknya, karena pada beberapa kasus sering dijumpai suatu set data dengan hubungan yang nonlinear. Permasalahan tersebut dapat diatasi dengan menerapkan *trick kernel* untuk fitur data awal pada suatu set data, sehingga algoritma SVM bisa beradaptasi dengan hubungan nonlinear. Kernel adalah sebuah fungsi yang memetakan fitur dalam data dari dimensi awal yang rendah ke fitur baru dengan dimensi yang lebih tinggi.

$$\begin{matrix} \Phi : D^r \rightarrow D^q \\ \boldsymbol{w} \rightarrow \Phi(\boldsymbol{w}) \end{matrix} \text{ dengan } r < q \qquad (6)$$

Φ adalah sebuah fungsi kernel untuk pemetaan, *D* adalah data latih, *r* adalah suatu set fitur data yang lama, *q* adalah suatu set fitur data yang baru dan merupakan hasil dari pemetaan pada setiap data latih, sedangkan $\boldsymbol{w}$ adalah data latih yang akan dipetakan ke fitur berdimensi *q* dengan $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_m \in D^r$.

Proses pemetaan dalam proses pelatihan memerlukan perhitungan *dot-product* dua buah vektor $\boldsymbol{w_i}$ dan $\boldsymbol{z}$ yang diberi notasi $\Phi(\boldsymbol{w_i}).\Phi(\boldsymbol{z})$ pada sebuah ruang fitur baru. Nilai *dot-product* pada kedua vektor tersebut dapat diperoleh dengan menghitung secara tidak langsung tanpa diketahui fungsi transformasi kernel Φ. Prediksi untuk set data berdimensi baru menggunakan fungsi Kernel selanjutnya diformulasikan dalam persamaan berikut (Prasetyo, 2014):

$$f\big(\Phi(z)\big) = sign(\boldsymbol{v}.\Phi(\boldsymbol{z}) + b) = sign(\textstyle\sum_{i\in S} \alpha_i y_i \Phi(\boldsymbol{w_i}).\Phi(\boldsymbol{z}) + b) \qquad (7)$$

$$f(z) = sign(\textstyle\sum_{i\in S} \alpha_i y_i K(\boldsymbol{w_i}, \boldsymbol{z}) + b) \qquad (8)$$

dengan,

$$\boldsymbol{v} = \textstyle\sum_{i\in S} \alpha_i y_i \Phi(\boldsymbol{w_i}) \qquad (9)$$

$$b = -\frac{1}{2}(\boldsymbol{v}.\boldsymbol{w_{-1}} + \boldsymbol{v}.\boldsymbol{w_{+1}}) \qquad (10)$$

Banyak data yang merupakan *support vector* disimbolkan dengan *S*, *support vector* disimbolkan dengan $\boldsymbol{w_i}$ dan data uji yang akan diprediksi disimbolkan dengan $\boldsymbol{z}$.

Evaluasi kinerja klasifikasi bertujuan untuk melihat performa klasifikasi yang telah dilakukan sebelumnya. *Confusion matrix* adalah hasil dari klasifikasi yang berupa jumlah data yang terklasifikasi dengan benar atau salah dan bertujuan untuk memudahkan dalam perhitungan akurasi. Nilai *accuracy* dapat dihitung menggunakan persamaan berikut (Reviantika *et al.*, 2021):

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\% \qquad (11)$$

Tabel 1. *Confusion Matrix*

Kelas aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Visualisasi data adalah proses ekstraksi informasi dari data berupa topik yang sering diperbincangkan, sehingga dari keseluruhan teks yang ada dapat mengambil informasi yang

penting (Novantika, 2022). Visualisasi data dapat dilakukan dengan menggunakan grafik atau plot seperti *word cloud* dan *barplot*.

3. METODE PENELITIAN

Penelitian ini menggunakan data primer berupa data kualitatif yang diperoleh melalui pengambilan *tweet* pengguna media sosial *Twitter*. Pengambilan data *tweet* dilaksanakan selama periode 17 Juni 2022 sampai dengan 17 Agustus 2022 menggunakan kata kunci "#BlokirKominfo" dalam kategori *tweet* berbahasa Indonesia sebanyak 5128 *tweets* Pada proses klasifikasi sentimen proporsi pembagian data latih dan data uji adalah 90:10 yang dibagi berdasarkan urutan data, yaitu 90 persen *tweets* pertama sebagai data latih dan sisanya sebagai data uji. Langkah-langkah analisisnya adalah sebagai berikut:

1. Persiapan data meliputi pengambilan *tweets* pada *Twitter, filtering* data dan *text preprocessing* meliputi *case folding*, *cleansing*, *normalize* dan *remove duplicate*
2. Pelabelan data menggunakan kamus InSet, *Sentistrength*, negasi dan *boosterwords* kedalam dua kategori, yaitu positif dan negatif
3. *Feature Selection* meliputi *stopword removal*, *stemming* dan *tokenizing*
4. Pembobotan kata menggunakan TF-IDF
5. Klasifikasi menggunakan *Penalized Logistic Regression* dengan *L1 regularization* dan *Support Vector Machine* dengan kernel RBF
6. Evaluasi kinerja klasifikasi dengan *confusion matrix*
7. Visualisasi Data

4. HASIL DAN PEMBAHASAN

Persiapan data dilakukan dengan pengambilan data *tweet* dan *filtering* data untuk menghapus *tweet* yang membicarakan diluar topik #BlokirKominfo sehingga data yang digunakan menjadi 1692 *tweets*. Proses dilanjutkan dengan *Text preprocessing* untuk transformasi data yang tidak terstruktur menjadi lebih terstruktur sehingga memudahkan proses analisis selanjutnya.

1. *Case folding*
   Pada tahapan *case folding* keseluruhan bentuk huruf diseragamkan menjadi huruf kecil atau *lowercase* agar lebih memudahkan pencarian.
2. *Cleansing*
   Tahapan *cleansing adalah* proses pembersihan data yang dilakukan dengan menghapus serangkaian tagar, simbol atau karakter, *emoticon*, tautan atau URL, nama pengguna dan tanda baca pada kalimat.
3. *Normalize*
   Proses *normalize* bertujuan untuk mengoreksi kesalahan eja pada kata dalam dokumen teks atau menggantikan kata-kata yang ditulis dengan singkat.
4. *Remove Duplicate*
   Proses *remove duplicate* dilakukan untuk menghilangkan data *tweets* yang sama dalam dokumen teks.

Proses pemberian label pada data *tweet* bertujuan untuk mengkategorikan *tweet* ke dalam kelas sentimen positif ataupun sentimen negatif dengan menghitung *sentiment score* berdasarkan kamus yang digunakan. Proses pelabelan data menggunakan *sentiment scoring* menghasilkan sejumlah 149 data *tweets* terklasifikasi kedalam sentimen positif dan 1543 data *tweets* terklasifikasi kedalam sentimen negatif. Proses analisis secara manual perlu dilakukan untuk mengoreksi hasil klasifikasi pada *sentiment scoring*. Setelah proses analisis ulang pada 1692 *tweets* ditemukan sejumlah 149 data mengalami kesalahan dalam pelabelan,

sehingga jumlah data *tweets* yang menghasilkan sentimen positif sejumlah 300 *tweets* dan sentimen negatif sejumlah 1392 *tweets*.

Tabel 2. Hasil Bobot *term* TF-IDF

***Tweet* ke-**	**atur**	**becus**	**bobrok**	**bonus**	**bukti**	**dasar**	**demografi**	...	**zona**
1	0	0	1,38922	3,06415	0	0	1,5320734	...	0
2	0,1048	0	0	0	0	0,22739	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
1692	0	1,72929	0	0	1,47929	0	0	...	0

Proses *feature selection* meliputi *stopword removal*, *stemming*, dan *tokenizing*. Proses *stopword removal* dilakukan dengan menggunakan kamus *stopword* untuk menghilangkan kata-kata yang kurang penting. Pada proses *stemming* kata berimbuhan pada *tweet* diubah menjadi kata dasarnya. Sedangkan pada proses *tokenizing* dilakukan pemisahan dokumen teks yang bergantung pada karakter spasi untuk melakukan pemisahan. TF-IDF (*Term Frequency- Inverse Document Frequency*) adalah pembobotan yang memberi setiap *term* nilai bobot didasarkan pada pentingnya *term* tersebut di dalam dokumen teks. Tabel 2 menunjukkan hasil proses perhitungan bobot setiap *term* dengan TF-IDF.

Data *tweets* yang digunakan adalah data dengan label sentimen yang dibedakan menjadi data latih dan data uji dengan proporsi 90:10. Tabel 3 menunjukkan pembagian *tweets* menjadi data latih dan data uji.

Tabel 3. Pembagian Data Latih dan Data Uji

Klasifikasi	**Data Latih**	**Data Uji**	**Jumlah**
Positif	273	27	300
Negatif	1249	143	1392
Jumlah	1522	170	1692

Klasifikasi sentimen menggunakan model *Logistic Regression* menggunakan fungsi logistik atau fungsi sigmoid yang melakukan prediksi dua nilai maksimum antara 0 dan 1. Pada penelitian ini klasifikasi sentimen dilakukan dengan model yang terpenalti yaitu *Penalized Logistic Regression* yang mana fungsi *logloss error function* diberikan sebuah kendala *L1 regularization.* Fungsi kendala ini akan digunakan untuk melakukan regularisasi dengan penyesuaian bobot sesuai dengan jumlah nilai bobot absolut. Pada proses regularisasi bobot fitur yang tidak relevan atau hampir tidak relevan akan diubah menjadi tepat 0 dengan menghapus fitur tersebut dari model. Proses penalti pada klasifikasi sentimen dengan *Penalized Logistic Regression* dipengaruhi oleh parameter regularisasi λ. Tabel 4 menunjukkan nilai λ yang dihasilkan pada proses komputasi *software* RStudio.

Tabel 4. Parameter Regularisasi λ

No.	**Df**	**%Dev**	**Lambda**
1	0	0	0,04926
2	2	0,24	0,04702
3	3	0,64	0,04489
4	3	1	0,04285
⋮	⋮	⋮	⋮
100	654	96,4	0,00049

Optimalisasi parameter $\boldsymbol{w}$ dilanjutkan dengan melakukan iterasi dengan menentukan nilai awal $t = 0$ dan *learning rate* (η) yang nilainya $\eta > 0$. Model dengan parameter $\boldsymbol{w}$ yang optimal selanjutnya dapat digunakan untuk prediksi sentimen data uji dengan menghitung fungsi sigmoid. Pada proses prediksi dilakukan dengan menambahkan nilai parameter

penalti λ sesuai Persamaan (5). Jika nilai fungsi sigmoid lebih dari *decision boundary* yang nilainya 0,5 maka data *tweet* akan diklasifikasikan kedalam sentimen positif. Sebaliknya jika nilai fungsi sigmoid kurang dari *decision boundary* yang nilainya 0,5 maka data *tweet* akan diklasifikasikan kedalam sentimen negatif.

Klasifikasi sentimen pada algoritma *Support Vector Machine* (SVM) menggunakan fungsi kernel *Radial Basis Function* (RBF) dengan parameter C (*Cost*) dan γ (*Gamma*). Nilai parameter *Cost* yang diujikan pada data latih untuk membangun model adalah 1, 100 dan 1000. Nilai parameter *Gamma* yang digunakan adalah nilai yang diperoleh dari perhitungan nilai *default*, yaitu:

$$\gamma = \frac{1}{ncol} = \frac{1}{2890} = 0{,}0003460208$$

dengan nilai *ncol* adalah jumlah dimensi data.

Konsep pembelajaran SVM adalah untuk menentukan nilai *dot product* pada dua data dalam suatu ruang *fitur* yang baru dengan dimensi lebih tinggi. Proses tersebut melibatkan *kernel trick* untuk penentuan *dot product* dengan cara mensubstitusikan fungsi kernel pada setiap data. Tabel 5 menunjukkan hasil perhitungan Fungsi Kernel RBF pada Data Latih.

Tabel 5. Hasil Perhitungan Fungsi Kernel RBF pada Data Latih

***Tweet* ke-**	1	2	3	4	...	1522
1	1	0,99361	0,99313	0,99341	...	0,99181
2	0,99361	1	0,99846	0,99858	...	0,99692
3	0,99313	0,99846	1	0,99813	...	0,99652
4	0,99341	0,99858	0,99813	1		0,99681
⋮	⋮	⋮	⋮	⋮	⋱	⋮
1522	0,99181	0,99692	0,99652	0,99681	...	1

Setelah dilakukan pemetaan data latih menggunakan fungsi *kernel*, selanjutnya dihitung nilai α dan b yang optimum dengan menggunakan *Quadratic Programming* (QP) *problem* pada setiap nilai parameter *Cost* yang digunakan, yaitu 1, 100 dan 1000. Pada proses perhitungan *Quadratic Programming* (QP) *problem* dilakukan dengan bantuan *software* RStudio karena ukuran data yang digunakan cukup besar. Tabel 6 merupakan hasil *output* dari *software* RStudio.

Tabel 6. Nilai α dan b pada Setiap Nilai *Cost*

Nilai *Cost*	**1**	**100**	**1000**
α	$\begin{bmatrix} 0{,}158354 \\ 0{,}231599 \\ 0{,}182676 \\ \vdots \end{bmatrix}_{796\times1}$	$\begin{bmatrix} 24{,}11849 \\ 10{,}26432 \\ 15{,}08666 \\ \vdots \end{bmatrix}_{870\times1}$	$\begin{bmatrix} 44{,}96175 \\ 166{,}6897 \\ 78{,}58487 \\ \vdots \end{bmatrix}_{855\times1}$
b	-0,8509	12,3907	23,1261

Nilai α dan b selanjutnya digunakan dalam persamaaan SVM sesuai Persamaan (8) untuk melakukan prediksi klasifikasi sentimen pada data uji. Data uji dimasukkan ke dalam data latih pada fungsi kernel untuk mendapatkan nilai *dot product* dari kedua pasangan data. Jika angka yang dihasilkan y_i negatif, maka hal tersebut menunjukkan bahwa data *tweets* bersentimen negatif. Sebaliknya, jika angka yang dihasilkan y_i positif, maka hal tersebut menunjukkan bahwa data *tweets* bersentimen positif.

Evaluasi kinerja klasifikasi pada penelitian ini menggunakan nilai akurasi yang dihasilkan melalui perhitungan tabel *Confusion matrix* dengan Persamaan (11). *Confusion matrix* merupakan hasil dari klasifikasi sentimen yang berupa jumlah data yang terklasifikasi dengan benar atau salah. Tabel 7 menunjukkan nilai akurasi pada algoritma *Penalized*

Logistic Regression dan Tabel 8 menunjukkan nilai akurasi pada algoritma *Support Vector Machine* yang dihasilkan melalui proses komputasi dengan bantuan *software* RStudio.

Berdasarkan pada Tabel 7, klasifikasi sentimen dengan menggunakan algoritma *Penalized Logistic Regression* menunjukkan hasil terbaik pada parameter lambda 0,04926 yang merupakan lambda dengan selang terkecil. Nilai akurasi tertinggi pada klasifikasi sentimen menggunakan algoritma *Penalized Logistic Regression* dengan lambda 0,04926 yaitu sebesar 0,8412 atau 84,12%.

Tabel 7. Nilai Akurasi *Penalized Logistic Regression*

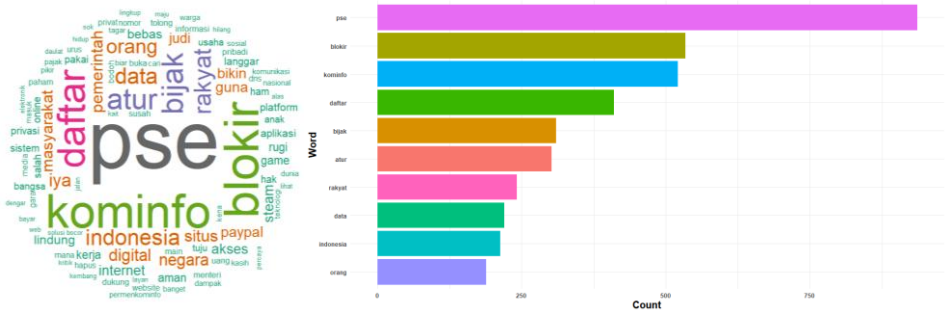
Algoritma Analisis Sentimen	Parameter *Lambda*	Nilai Akurasi
Penalized Logistic Regression	0,01	0,8235
	0,04926	0,8412
	0,00049	0,8

Berdasarkan pada Tabel 8, klasifikasi sentimen menggunakan algoritma *Support Vector Machine* hasil terbaik pada parameter *Cost* 1000. Nilai akurasi tertinggi pada klasifikasi sentimen menggunakan algoritma *Support Vector Machine* dengan *Cost* 1000 yaitu sebesar 0,8353 atau 83,53%. Hal tersebut menunjukkan bahwa kedua algoritma memiliki nilai akurasi yang cukup baik dengan akurasi terbaik untuk klasifikasi sentimen data *tweets* terkait kebijakan PSE adalah menggunakan algoritma *Penalized Logistic Regression*.

Tabel 8. Nilai Akurasi *Support Vector Machine*

Algoritma Analisis Sentimen	Parameter *Cost*	Nilai Akurasi
Support Vector Machine	1	0,8235
	100	0,8235
	1000	0,8353

Word cloud bertujuan untuk menggambarkan kata-kata yang paling banyak digunakan dalam dokumen teks dengan memanfaatkan frekuensi kemunculan kata pada dokumen. Visualisasi menggunakan *barplot* memudahkan pembaca mengetahui jumlah frekuensi munculnya kata tertentu dalam dokumen teks. Gambar 1 merupakan bentuk visualisasi data menggunakan *word cloud* dan *barplot*. Berdasarkan Gambar 1, kata "pse", "blokir", dan "kominfo" menjadi kata yang banyak diperbincangkan dalam dokumen *tweets*.



Gambar 1. *Word cloud* dan *Barplot*

Melalui visualisasi data tersebut dapat diketahui bahwa komentar masyarakat terhadap implementasi peraturan terkait Penyelenggara Sistem Elektronik (PSE) menunjukkan sentimen negatif akibat penerapan sanksi blokir terhadap beberapa PSE yang tidak melakukan pendaftaran yang ditentukan oleh Kominfo dalam Peraturan Menteri Komunikasi dan Informatika Republik Indonesia Nomor 5 Tahun 2020 Tentang Penyelenggara Sistem Elektronik Lingkup Privat.

5. KESIMPULAN

Opini masyarakat pengguna media sosial *Twitter* cenderung menuliskan *tweet* yang menunjukkan sentimen negatif terhadap implementasi kebijakan Penyelenggara Sistem Elektronik (PSE). Hal ini dibuktikan dari sejumlah 1692 *tweets* yang dianalisis, diperoleh sebanyak 1543 *tweets* bersentimen negatif dan hanya 149 yang bersentimen positif. Klasifikasi sentimen dengan 1522 *tweets* sebagai data latih dan 170 sebagai data uji menggunakan algoritma *Penalized Logistic Regression* dengan parameter $\lambda = 0{,}04926$ menghasilkan akurasi terbesar yaitu 84,12%. Pada klasifikasi menggunakan algoritma *Support Vector Machine* dengan parameter $Cost = 1000$ menghasilkan akurasi terbesar yaitu 83,53%. Hal tersebut mengartikan bahwa algoritma *Penalized Logistic Regression* bekerja lebih baik jika dibandingkan dengan algoritma *Support Vector Machine* dalam mengklasifikasikan sentimen masyarakat terhadap implementasi kebijakan Penyelenggara Sistem Elektronik (PSE) di media sosial *Twitter*.

Sentimen masyarakat terhadap implementasi kebijakan Penyelenggara Sistem Elektronik (PSE) yang menerapkan pemblokiran terhadap sejumlah PSE yang tidak melakukan pendaftaran ke KemKominfo menunjukkan sentimen negatif. Sehingga disarankan bagi lembaga pemerintah, khususnya KemKominfo dalam mengimplementasikan kebijakan-kebijakan lainnya dapat mempertimbangkan kembali terkait beberapa aspek yang berkaitan dan berdampak terhadap masyarakat supaya implementasi kebijakan lebih matang dan tereksekusi dengan tepat.

DAFTAR PUSTAKA

Alsaeedi, A. dan Khan, M. Z. (2019) 'A study on sentiment analysis techniques of Twitter data', *International Journal of Advanced Computer Science and Applications*, 10(2), pp. 361–374. doi: 10.14569/ijacsa.2019.0100248.

Andrew. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 615–622. https://doi.org/10.1145/1015330.1015435

Darwis, D., Pratiwi, E. S. dan Pasaribu, A. F. O. (2020) 'Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia', *Jurnal Ilmiah Edutic*, 7(1), pp. 1–11.

Ihsan, I., Nurjanah, D. dan Nurrahmi, H. (2021) 'Sentiment Analysis RKUHP Pada Twitter Menggunakan Metode Support Vector Machine', *e-Proceeding of Engineering*, 8(2), pp. 3521–3536.

Kassambara, A. (2019). *Machine Learning Essentials Practical Guide in R, First Edition*. www.sthda.com/english

Kementerian Komunikasi dan Informatika (2020) 'Peraturan Kementerian Kominfo No. 5 Tahun 2020', *Kementerian Komunikasi dan Informatika*.

KemKominfo (2022) *Siaran Pers No. 308/HM/KOMINFO/07/2022*. Available at: https://www.kominfo.go.id/content/detail/43385/siaran-pers-no-308hmkominfo072022-tentang-pendaftaran-penyelenggara-sistem-elektronik-pse-lingkup-privat/0/siaran_pers (Accessed: 22 September 2022).

Khoerunisa. (2021). *Kajian Regresi Logistik Terpenalti Pada Data Dimensi Tinggi*. IPB University.

Koto, F., dan Rahmaningtyas, G. Y. (2018). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, *2018-January* (December), 391–394. https://doi.org/10.1109/IALP.2017.8300625

Liu, B. (2012) *Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies*. Morgan dan Claypool Publishers. doi: 10.2200/S00416ED1V01Y201204HLT016.

Molnar, C. (2019) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Second. Morisville, North Carolina: Lulu. Available at: https://christophm.github.io/interpretable-ml-book/index.html.

Pekhimenko, G. G. (2006) 'Penalized Logistic Regression for Classification', *Department of Computer Science University of Toronto*, (2).

Prasetyo, E. (2014) *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Edited by A. Sahala. Yogyakarta: ANDI PUBLISHER.

Rahmawati, F. (2022) *Dirjen Aptika: Kominfo akan Blokir PSE Lingkup Privat yang Tidak Terdaftar, Kominfo*. Available at: https://aptika.kominfo.go.id/2022/06/dirjen-aptika-kominfo-akan-blokir-pse-lingkup-privat-yang-tidak-terdaftar/ (Accessed: 22 September 2022).

Rahutomo, F., Saputra, P. Y. dan Fidyawan, M. A. (2018) 'Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine', *Jurnal Informatika Polinema*, 4(2), p. 93. doi: 10.33795/jip.v4i2.152.

Reviantika, F., Azhar, Y. dan Marthasari, G. I. (2021) 'Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression', *Jurnal Sistem Cerdas*, 4(2), pp. 37–43. Available at: https://www.cnbcindonesia.com.

Rezeki, R. I., Restiviani, Y. dan Zahara, R. (2020) 'PENGGUNAAN SOSIAL MEDIA TWITTER DALAM KOMUNIKASI ORGANISASI (Studi Kasus Pemerintah Provinsi DKI Jakarta Dalam Penanganan Covid-19)', 4(2), pp. 63–78.

Sabrila, T. S., Azhar, Y. dan Aditya, C. S. K. (2022) 'Analisis Sentimen Tweet Tentang UU Cipta Kerja Menggunakan Algoritma SVM Berbasis PSO', *JISKA (Jurnal Informatika Sunan Kalijaga)*, 7(1), pp. 10–19. doi: 10.14421/jiska.2022.7.1.10-19.

Testiana, G. dan Erlina, D. (2022) *Analisis Sentimen Pada Twitter Terhadap Uin Raden Fatah Menggunakan Support Vector Machine, JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*. doi: 10.35957/jatisi.v9i1.1433.

Tyagi, A. dan Sharma, N. (2018) 'Sentiment Analysis using logistic regression and effective word score heuristic', *International Journal of Engineering and Technology (UAE)*, 7(2), pp. 20–23. doi: 10.14419/ijet.v7i2.24.11991.

Wahid, D. H., dan Azhari. (2016). Peringkasan Sentimen Ekstraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *10*(2), 207. https://doi.org/10.22146/ijccs.16625

Wardhani, E. D., Areka, S. K., Nugroho, A. W., Zakaria, A. R., Prakasa, A. D. dan Nooraeni, R. (2020) 'Sentiment Analysis Using Twitter Data Regarding BPJS Cost Increase and Its Effect on Health Sector Stock Prices', *Indonesian Journal of Artificial Intelligence and Data Mining*, 3(1), p. 1. doi: 10.24014/ijaidm.v3i1.8245.