Aplikasi Analisa Sentimen Bilingual dan Emoji pada Komentar Media Sosial Instagram Menggunakan Metode Support Vector Machine

Satria Adi Nugraha, Henry Novianus Palit, Hans Juwiantho
Program Studi Informatika Fakultas Teknologi Industri Universitas Kristen Petra
Jalan Siwalankerto 121 – 131 Surabaya 60236
Telp. (031) – 2983455, Fax. (031) - 8417658
E-mail: satriaadin45@gmail.com, hnpalit@petra.ac.id, hans.juwiantho@petra.ac.id

ABSTRAK

Indonesia menduduki peringkat 4 sebagai pengguna Instagram terbanyak di dunia. Hal ini membuat pelaku bisnis terpicu untuk mempromosikan produk maupun jasanya kepada content creator untuk dibuat ulasannya dan diunggah di Instagram. Pelaku bisnis perlu melakukan evaluasi terhadap unggahan untuk menilai apakah promosi yang dilakukan mendapatkan respon positif atau negatif dari warganet. Evaluasi dapat dilakukan dengan melakukan pengecekan pada kolom komentar. Komentar Instagram tidak hanya berisi komentar berbahasa Indonesia, namun berbahasa Inggris beserta emoji. Namun, pengecekan secara manual tentu akan memakan banyak waktu. Oleh karena itu, perlu dibangun sistem aplikasi yang mampu mendeteksi sentimen bilingual dan emoji pada komentar Instagram. Sistem ini dibangun menggunakan metode Support Vector Machine untuk mengklasifikasikan bahasa, sentimen Indonesia, serta sentimen Inggris dan kemudian dievaluasi menggunakan nilai accuracy. Data yang dipakai merupakan sampel dari komentar unggahan berupa post, reels, maupun IGTV. Kombinasi preprocessing cleansing, normalisasi, stopwords removal, dan stemming serta parameter tuning menggunakan GridSearchCV juga diuji untuk menemukan model terbaik. Model dibagi menjadi model klasifikasi bahasa dengan label Indonesia, Inggris, dan Campuran, klasifikasi sentimen Indonesia dan klasifikasi sentimen Inggris dengan label positif, netral, dan negative. Akurasi terbaik yang didapatkan model untuk klasifikasi bahasa, sentimen Indonesia, dan sentimen Inggris masing-masing 88.77%, 73.10%, 71.56%. Selain itu, emoji perlu dianalisa karena model yang menganalisa emoji memiliki akurasi 3.875% lebih baik daripada model yang mengabaikan emoji.

Kata Kunci: analisa sentimen, Support Vector Machine, komentar Instagram, analisa sentimen bilingual

ABSTRACT

Indonesia is ranked 4th as the most Instagram user in the world. This makes business people triggered to promote their products and services to content creators to make reviews and upload them on Instagram. Business people need to evaluate uploads to assess whether the promotions carried out get a positive or negative response from netizens. Evaluation can be done by checking the comments column. Instagram comments not only contain comments in Indonesian but in English along with emojis. However, checking manually will certainly take a lot of time. Therefore, it is necessary to build an application system that can detect bilingual sentiments and emojis in Instagram comments. This system was built using the Support Vector Machine method to classify language, Indonesian sentiment, and English sentiment and then evaluated using the accuracy value. The data used is a sample of uploaded comments in the form of posts, reels, and IGTV. The combination of preprocessing cleansing, normalization, stopwords removal, and stemming as well as parameter tuning using GridSearchCV was also tested to find the best model. The model is divided into language classification models with Indonesia, Inggris, and Campuran labels, Indonesian sentiment classifications, and English sentiment classifications with positive, neutral, and negative labels. The best accuracy obtained by the model for language classification, Indonesian sentiment, and English sentiment is 88.77%, 73.10%, and 71.56%, respectively. In addition, emojis need to be analyzed because the model that analyzes emojis has 3.875% better accuracy than the model that ignores emoji..

Keywords: *sentiment analysis, Support Vector Machine, Instagram comments, bilingual sentiment analysis*

1. PENDAHULUAN

Penggunaan media sosial terus berkembang dengan pesat tak terkecuali di Indonesia. Media sosial menjadi salah satu sarana komunikasi yang paling *up to date* dan sering digunakan pada saat ini. Instagram menjadi satu dari sekian media sosial yang sering dipakai oleh para content creator di Indonesia. Hal itu didukung dengan adanya 93 juta pengguna Instagram di Indonesia per Juli 2021 [15]. Angka tersebut membuat Indonesia menduduki peringkat 4 terbesar pengguna Instagram di dunia. Dari 270,2 juta penduduk Indonesia [2], artinya 34,4% masyarakat sudah menggunakan Instagram sebagai salah satu media sosial pilihannya. Tingginya minat masyarakat Indonesia akan penggunaan Instagram membuat pelaku bisnis terpicu untuk mempromosikan produknya kepada para content creator agar dibuat ulasannya dan diunggah di Instagram. Pelaku bisnis perlu menilai apakah promosi yang dilakukan mendapatkan respon positif atau negatif dari masyarakat. Upaya yang biasanya dilakukan adalah dengan melihat berapa jumlah likes dan share. Namun, likes tidak bisa merepresentasikan apakah unggahan tersebut benar-benar disukai oleh masyarakat [13]. Hal lain yang dapat dievaluasi adalah komentar warganet. Melalui kolom komentar, pelaku bisnis dan content creator dapat melihat pandangan masyarakat terhadap unggahan tersebut. Sayangnya, pengecekan komentar secara manual tentu kurang efektif dan memakan banyak waktu karena komentar yang masuk jumlahnya cukup banyak. Oleh karena itu, diperlukan aplikasi yang dapat mengkategorikan komentar secara otomatis. Penelitian sebelumnya [5] mengkategorikan komentar berbahasa Indonesia pada komentar di video Youtube menjadi sentimen positif dan negatif menggunakan metode Naïve Bayes Classifier. Namun pada kenyataannya, media sosial sekarang, termasuk Instagram dan Youtube, tidak hanya di isi oleh komentar berbahasa Indonesia, namun juga bahasa Inggris. Selain itu, tak jarang ditemukan emoji

dalam komentar pada unggahan Instagram. Emoji perlu di analisa karena berpengaruh terhadap hasil sentimen komentar yang ada [1]. Oleh karena itu, pada skripsi ini akan menggunakan sentiment analysis bilingual dan emoji untuk memecahkan masalah tersebut. Metode SVM digunakan untuk mengklasifikasi sentimen komentar. SVM terbukti menjadi satu dari sekian banyak metode yang mampu melakukan kategorisasi teks secara optimal[10][17]. SVM juga lebih unggul dari metode Naïve Bayes [12].

2. PENELITIAN SEBELUMNYA

Beberapa penelitian mengenai sentimen analisis pernah dilakukan sebelumnya menggunakan metode SVM maupun Naïve Bayes. Penelitian sebelumnya menggunakan Naïve Bayes untuk analisa sentimen komentar Youtube dan analisis *cyberbullying* pada Instagram hanya dilakukan pada komentar berbahasa Indonesia[5][9]. Metode SVM juga dilakukan pada penelitian analisa sentimen Twitter mengenai calon presiden dan *review* produk *smartphone*[17][8]. Penelitian tersebut juga hanya menggunakan satu bahasa. Metode SVM lebih unggul dibanding Naïve Bayes ditunjukkan pada penelitian yang membandingkan dua metode tersebut untuk membandingkan review film[12]. Dari semua penelitian sebelumnya emoji juga dihapus. Oleh karena itu penelitian ini akan menggunakan metode SVM untuk menganalisa sentimen komentar bilingual (Indonesia dan Inggris) beserta emojinya.

3. DATASET

Dataset yang digunakan dalam penelitian ini diambil dari 82 unggahan Instagram baik berupa *post*, *reels*, maupun *IGTV*. Dari unggahan tersebut kemudian diambil sampel komentar sebanyak 3874 komentar yang dikemudian diberi label sesuai bahasa ataupun sentimen. Label bahasa terdiri dari bahasa Indonesia, Inggris, dan Campuran. Sedangkan, label sentimen terdiri dari *Positive, Neutral,* dan *Negative*. Pelabelan dilakukan oleh pengguna Instagram. Setelah dilakukan pelabelan, dataset dengan label bahasa terdiri dari 392 bahasa Campuran, 2429 bahasa Indonesia, dan 1053 bahasa Inggris. Untuk label sentimen, dataset terdiri dari 1381 *negative*, 1027 *neutral*, dan 1466 data *positive*. Secara rinci, dataset terbagi seperti pada Tabel 1.

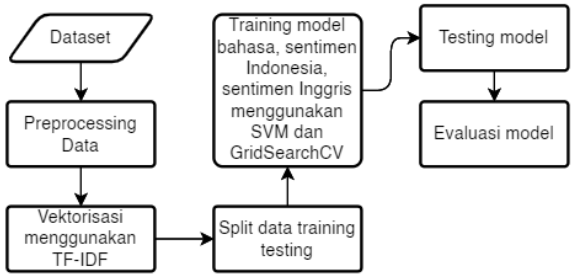
Tabel 1. Jumlah Data berdasar Label Bahasa dan Sentimen

Sentimen	Bahasa	Jumlah
Negative	Campuran	68
Negative	Indonesia	939
Negative	Inggris	374
Neutral	Campuran	162
Neutral	Indonesia	649
Neutral	Inggris	216
Positive	Campuran	162
Positive	Indonesia	841
Positive	Inggris	463

Data kemudian dibagi menjadi 2, sebanyak 80% untuk data training dan 20% untuk data testing yang kemudian digunakan untuk membangun model klasifikasi bahasa maupun sentimen.

4. METODE

Metode pada penelitian ini terbagi menjadi beberapa bagian seperti *preprocessing* data menggunakan NLTK maupun Sastrawi [14], vektorisasi menggunakan TF-IDF, pembentukan model menggunakan SVM dan pemilihan parameter terbaik menggunakan GridSearchCV. Gambaran besar alur pembentukan model dan evaluasi ditunjukkan pada Gambar 1. Data yang masuk dilakukan *preprocessing* untuk membersihkan data dan menghapus karakter-karakter yang tidak diperlukan. Kemudian data ditranformasi menjadi matriks vector menggunakan TF-IDF agar dapat diproses oleh metode SVM.



Gambar 1. Alur Pembentukan dan Evaluasi Model

Data kemudian dibagi menjadi data training dan data testing. Data training digunakan untuk membentuk model menggunakan SVM serta GridSearchCV sebagai hyperparameter tuning. Parameter yang diuji antara lain kernel, C, degree, dan gamma. Model yang menghasilkan skor terbaik digunakan untuk memprediksi data test dan kemudian dilakukan evaluasi untuk mengukur performa model.

4.1 Support Vector Machine

Support Vector Machine (SVM) adalah salah satu metode dalam supervised learning yang umumnya digunakan untuk regresi dan klasifikasi. SVM dipopulerkan oleh Vapnik, Boser, dan Guyon pada tahun 1992 berdasarkan prinsip Structural Risk Minimization [7]. Tujuan dari model ini adalah menemukan hyperplane terbaik yang memisahkan dua buah kelas pada sebuah tempat sehingga terbentuk classifier. Hyperplane adalah sebuah fungsi yang digunakan untuk menjadi pemisah antar kelas. Metode SVM pada awalnya dibuat hanya untuk melakukan klasifikasi anatara dua buah kelas (binary classfication), namun seiring dengan perkembangan zaman SVM telah dimodifikasi agar mampu menyelesaikan masalah lebih dari dua kelas (multiclass classification). Hyperplane terbaik dapat dicari dengan cara mengukur nilai margin dari hyperplane tersebut dan menemukan titik maksimalnya. Margin merupakan jarak yang dibuat antara objek atau pattern terluar dari masing-masing class dengan hyperplane yang terbuat. Objek atau pattern terluar atau yang paling dekat dengan hyperplane ini disebut sebagai support vector. Untuk menyelesaikan masalah multiclass classification, metode ini menyederhanakan dan memecah masalah multiclass menjadi multiple binary classification. Metode yang dikenal adalah One-to-One dan One-to-Rest [3].

4.2 TF-IDF Vectorizer

Term Frequency – Inverse Document Frequencty (TF-IDF) adalah metode yang sering digunakan dalam Natural Language Processing (NLP) yang bertujuan untuk mengubah teks dokumen menjadi matrix vector. TF-IDF merupakan teknik vektorisasi yang digunakan untuk mengukur jumlah kata dalam dokumen dan kemudian bobot setiap kata tersebut dihitung untuk menandakan pentingnya kata dalam sebuah dokumen [11]. Tujuan penggunaan vektorisasi TF-IDF dibanding metode lain seperti Bag of Words Model dan Word Counts adalah karena dalam sebuah dokumen bisa terdapat kata yang muncul berulang namun tidak punya makna yang signifikan. TF-IDF mampu mengesampikan kata dengan frekuensi kemunculan tinggi namun tidak ada makna signifikan dan mampu memperhitungkan kata yang punya makna signifikan walau tidak terlalu sering muncul. Terdapat beberapa cara dalam menghitung bobot skema untuk TF-IDF secara umum.

4.3 Grid Search CV

Grid search merupakan salah satu metode yang sering kali digunakan untuk melakukan pencarian nilai paling optimal dari sebuah parameter dalam proses pembentukan model. Proses ini biasa disebut dengan hyper parameter tuning. Tujuan proses GridSearch beserta Cross-Validation adalah untuk mengidentifikasi kombinasi hyper-parameter terbaik sehingga model dapat memprediksi data dengan akurat dan optimal [16] Proses pemilihan parameter dalam SVM dapat terbantu dengan menggunakan GridSearchCV. SVM memiliki beberapa parameter seperti nilai C, gamma, kernel (linear, poly, rbf, sigmoid), dan degree. Setiap parameter diatas memiliki rentang nilai yang bervariasi yang sebaiknya dicoba untuk membentuk model terbaik berdasar data yang ada. Proses hyper-parameter tuning menggunakan GridSearchCV dapat mempersingkat waktu tersebut. Penelitian [6] juga membutikkan bahwa model SVM menggunakan Grid Search lebih baik daripada menggunakan model SVM normal (default).

4.4 Preprocessing Data

Data akan melalui beberapa tahap seperti cleansing termasuk case folding, normalisasi, stopwords removal maupun stemming sebelum masuk ke dalam vektorisasi dan pembentukan model. Cleansing adalah proses menghapus elemen pada kalimat/kata yang tidak perlukan agar model dapat berjalan lebih cepat. Cleansing merupakan salah satu tahap pre-processing dalam data berbentuk teks. Beberapa hal yang biasanya dilakukan pada teks media sosial adalah menghapus tanda baca, menghapus link dan menghapus mention pengguna. Cleansing juga mengubah kata dengan huruf yang duplikat menjadi kata yang normal. Normalisasi digunakan untuk mengganti kata-kata gaul atau singkatan menjadi kata normalnya. Stopwords removal adalah proses menghapus kata-kata yang tidak mempunyai makna [4]. Tujuan dari proses ini adalah meningkatkan kecepatan dan performa pemrosesan model. Umumnya pemilihan stopword adalah kata yang mempunyai frekuensi kemunculan yang tinggi seperti kata penghubung. Stemming adalah proses menyederhanakan atau memetakan kata-kata berbeda atau bervarian ke dalam basis atau kata umum. Proses ini dapat meningkatkan kinerja pengambilan informasi serta juga akan mengurangi ukuran file indeks saat melakukan pengindeksan. Algoritma stemming untuk bahasa satu dengan yang lainnya tentu berbeda karena karakteritik bahasa yang dimiliki. Emoji pada komentar juga diubah menjadi *unicode point hexadecimal* seperti pada Tabel 2.

Tabel 2. Contoh Emoji beserta Unicode Name dan Codepoint

Emoji (*char*)	*Unicode Name*	*Unicode Codepoint*
😂	FACE WITH TEARS OF JOY	0x1f602
👍	THUMBS UP SIGN	0x1f44d
😭	LOUDLY CRYING FACE	0x1f62d
🙏	PERSON WITH FOLDED HANDS	0x1f64f
👏	CLAPPING HANDS SIGN	0x1f44f

5. PENGUJIAN

Model klasifikasi bahasa, klasifikasi sentimen Indonesia, dan klasifikasi sentimen Inggris diuji untuk menemukan performa model yang terbaik menggunakan akurasi. Pengujian dilakukan dengan kombinasi preprocessing (*cleansing*, normalisasi, *stopwords removal*, dan *stemming*) serta percobaan parameter SVM yang terdiri kernel (rbf, linear, poly, sigmoid), C (0.1, 1, 10, 100), gamma (1, 0.1, 0.01, 0.001, 0.0001) dan degree (1, 2, 3, 4, 5, 6) menggunakan GridSearchCV.

5.1 Model Klasifikasi Bahasa

Pengujian pertama dilakukan pada model klasifikasi bahasa dengan label Indonesia, Inggris, dan Campuran. Tabel 3 menunjukkan 5 parameter terbaik untuk model klasifikasi bahasa dengan preprocessing *cleansing* saja. Parameter terbaik untuk model ini adalah kernel RBF, dengan nilai C adalah 10, dan gamma sebesar 0.1.

Tabel 3. Lima Besar Tuning Parameter Preprocessing Cleansing Klasifikasi Bahasa

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
10	1	0.1	rbf	0.856	0.010	1
1	1	1	poly	0.856	0.008	7
1	1	1	linear	0.856	0.008	13
10	1	0.1	poly	0.856	0.008	13
10	1	0.1	sigmoid	0.856	0.008	13

Pada proses preprocessing *cleansing* dan normalisasi, parameter terbaik untuk klasifikasi bahasa adalah kernel RBF, dengan nilai C sebesar 10 dan gamma 0.1 yang ditunjukkan pada Tabel 4.

Tabel 4. Lima Besar Tuning Parameter Preprocessing Cleansing dan Normalisasi Klasifikasi Bahasa

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
10	1	0.1	rbf	0.865	0.009	1
100	1	0.01	rbf	0.863	0.007	7
1	1	1	poly	0.860	0.005	13
1	1	1	linear	0.860	0.005	13
10	1	0.1	poly	0.860	0.005	13

Sedangkan, parameter terbaik untuk klasifikasi bahasa dengan proses preprocessing *cleansing*, normalisasi, dan *stopwords removal* adalah kernel RBF, dengan nilai C dan gamma masing-masing 1 seperti pada Tabel 5.

Tabel 5. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi, dan Stopwords Removal Klasifikasi Bahasa

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.831	0.009	1
1	2	1	poly	0.827	0.010	7
100	2	0.1	poly	0.827	0.010	7
1	1	1	poly	0.824	0.011	9
1	1	1	linear	0.824	0.011	9

Pengujian terakhir klasifikasi bahasa adalah dengan semua proses preprocessing (*cleansing*, normalisasi, *stopwords removal*, dan stemming) dengan parameter terbaik adalah kernel poly, dengan nilai C, gamma, dan degree masing-masing 1 seperti pada Tabel 6.

Tabel 6. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi, Stopwords Removal, dan Stemming Klasifikasi Bahasa

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	poly	0.840	0.007	1
1	1	1	linear	0.840	0.007	1
10	1	0.1	poly	0.840	0.007	1
100	1	0.01	poly	0.840	0.007	1
100	1	0.01	sigmoid	0.840	0.007	1

5.2 Model Klasifikasi Sentimen Indonesia

Pengujian kedua dilakukan pada model klasifikasi sentimen berbahasa Indonesia dengan kombinasi preprocessing serta parameter tuning seperti pada klasifikasi bahasa. Tabel 7 menunjukkan parameter terbaik untuk klasifikasi sentimen Indonesia dengan preprocessing *cleansing* saja. Parameter terbaiknya adalah kernel RBF dengan nilai C dan gamma masing-masing 1.

Tabel 7. Lima Besar Tuning Parameter Preprocessing Cleansing Klasifikasi Sentimen Indonesia

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.721	0.020	1
1	2	1	poly	0.714	0.020	7
100	2	0.1	poly	0.714	0.020	7
1	1	1	poly	0.711	0.012	9
1	1	1	linear	0.711	0.012	9

Pada proses preprocessing cleansing dan normalisasi, parameter terbaik untuk klasifikasi sentimen Indonesia adalah kernel RBF, dengan nilai C sebesar 1 dan gamma 1 yang ditunjukkan pada Tabel 8.

Tabel 8. Lima Besar Tuning Parameter Preprocessing Cleansing dan Normalisasi Klasifikasi Sentimen Indonesia

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.727	0.015	1
1	2	1	poly	0.721	0.013	7
100	2	0.1	poly	0.721	0.013	7
1	1	1	poly	0.715	0.019	9
1	1	1	linear	0.715	0.019	9

Sedangkan, parameter terbaik untuk klasifikasi sentimen Indonesia dengan proses preprocessing *cleansing*, normalisasi, dan *stopwords removal* adalah kernel RBF, dengan nilai C dan gamma masing-masing 1 seperti pada Tabel 9.

Tabel 9. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi dan Stopwords Removal Klasifikasi Sentimen Indonesia

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.703	0.020	1
1	1	1	poly	0.697	0.013	7
1	1	1	linear	0.697	0.013	7
10	1	0.1	poly	0.697	0.013	7
10	1	0.1	sigmoid	0.697	0.013	7

Pengujian terakhir klasifikasi sentimen Indonesia adalah dengan semua proses preprocessing (*cleansing*, normalisasi, *stopwords removal*, dan stemming) dengan parameter terbaik adalah kernel RBF, dengan nilai C dan gamma masing-masing 1 seperti pada Tabel 10.

Tabel 10. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi, Stopwords Removal dan Stemming Klasifikasi Sentimen Indonesia

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.697	0.019	1
10	1	0.1	rbf	0.691	0.013	7
10	1	1	rbf	0.691	0.016	13
1	1	1	sigmoid	0.691	0.012	19
10	1	0.1	sigmoid	0.691	0.012	19

5.3 Model Klasifikasi Sentimen Inggris

Pengujian terakhir dilakukan apda model klasifikasi sentimen berbahasa Inggris dengan kombinasi preprocessing serta parameter tuning. Parameter terbaik untuk klasifikasi sentimen Inggris dengan preprocessing *cleansing* adalah kernel RBF dengan nilai C dan gamma masing-masing 1 seperti yang ditunjukkan pada Tabel 11.

Tabel 11. Lima Besar Tuning Parameter Preprocessing Cleansing Klasifikasi Sentimen Inggris

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.694	0.018	1
10	1	0.1	rbf	0.690	0.026	7
10	1	1	rbf	0.685	0.023	13
100	1	1	rbf	0.685	0.023	13
100	1	0.01	rbf	0.679	0.020	25

Pada proses preprocessing cleansing dan normalisasi, parameter terbaik untuk klasifikasi sentimen Inggris adalah kernel RBF,

dengan nilai C sebesar 1 dan gamma 1 yang ditunjukkan pada Tabel 12.

Tabel 12. Lima Besar Tuning Parameter Preprocessing Cleansing dan Normalisasi Klasifikasi Sentimen Inggris

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.704	0.031	1
1	2	1	poly	0.698	0.035	7
100	2	0.1	poly	0.698	0.035	7
10	1	0.1	rbf	0.691	0.025	9
10	1	1	rbf	0.690	0.025	15

Sedangkan, parameter terbaik untuk klasifikasi sentimen Inggris dengan proses preprocessing *cleansing*, normalisasi, dan *stopwords removal* adalah kernel RBF, dengan nilai C dan gamma masing-masing 10 dan 1 seperti pada Tabel 13.

Tabel 13. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi dan Stopwords Removal Klasifikasi Sentimen Inggris

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
10	1	1	rbf	0.670	0.043	1
1	1	1	rbf	0.665	0.031	7
100	1	1	rbf	0.663	0.039	13
1	2	1	poly	0.657	0.031	19
100	2	0.1	poly	0.657	0.031	19

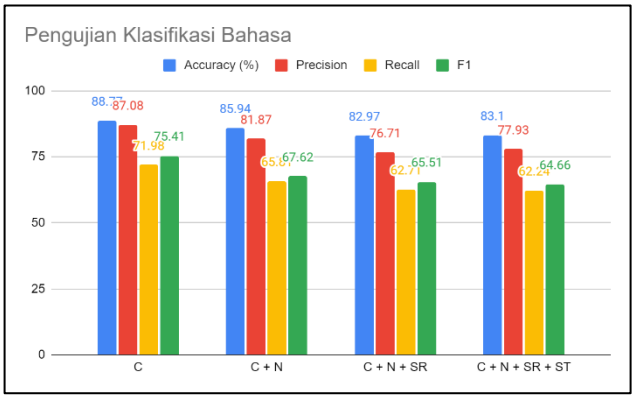
Pengujian terakhir klasifikasi sentimen Inggris adalah dengan semua proses preprocessing (*cleansing*, normalisasi, *stopwords removal*, dan stemming) dengan parameter terbaik adalah kernel RBF, dengan nilai C dan gamma masing-masing 1 seperti pada Tabel 14.

Tabel 14. Lima Besar Tuning Parameter Preprocessing Cleansing Normalisasi, Stopwords Removal dan Stemming Klasifikasi Sentimen Inggris

param_C	param_degree	param_gamma	param_kernel	mean_test_score	std_test_score	rank_test_score
1	1	1	rbf	0.692	0.035	1
1	1	1	poly	0.687	0.032	7
1	1	1	linear	0.687	0.032	7
10	1	0.1	poly	0.687	0.032	7
10	1	0.1	sigmoid	0.687	0.032	7

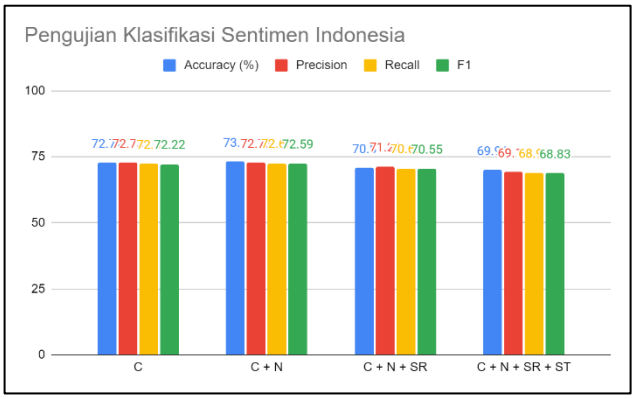
5.4 Perbandingan Hasil Pengujian

Setiap model yang dihasilkan dari pengujian kombinasi preprocessing dan parameter tuning kemudian dievaluasi menggunakan *accuracy* dan F1-score dengan data test yang telah dipisahkan sebelumnya. Pada model klasifikasi bahasa, *preprocessing cleansing* memiliki performa model terbaik dengan akurasi 88.77%. Setelah data telah melewati proses normalisasi dan stopwords removal justru performa model untuk klasifikasi bahasa turun dan berkurang antara 2.83% hingga 5.8% dan sedikit meningkat ketika data telah melewati proses *stemming*. Data evaluasi model klasifkasi bahasa ditunjukkan pada Gambar 2.



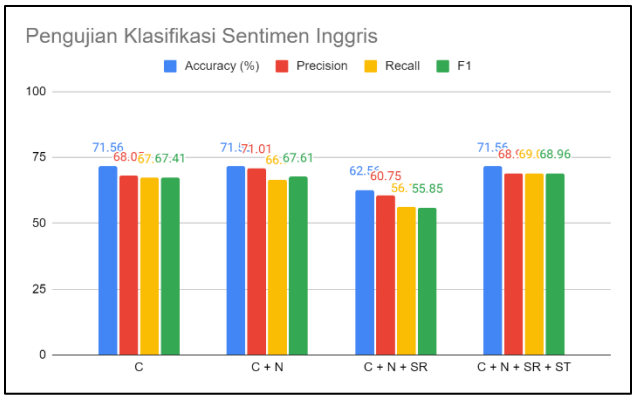
Gambar 2. Evaluasi Model Klasifikasi Bahasa

Untuk model klasifikasi sentimen bahasa Indonesia evaluasi ditunjukkan pada Gambar 3. Pada gambar tersebut terlihat bahwa proses *preprocessing cleansing* dan normalisasi menghasilkan performa model terbaik dengan akurasi sebesar 73.1%. Sedikit lebih tinggi dari proses *cleansing* saja. Performa model juga menurun ketika telah melalui tahap *stopwords removal* dan *stemming*.



Gambar 3. Evaluasi Model Klasifikasi Sentimen Indonesia

Pengujian pada model klasifikasi sentimen bahasa Inggris memiliki akurasi yang mirip untuk setiap kombinasi proses preprocessingnya. Hanya pada kombinasi preprocessing cleansing, normalisasi, dan stopwords removal yang memiliki akurasi terendah pada nilai 62.56%. Untuk kombinasi preprocessing 1,2, dan 4 sama-sama memiliki akurasi sebesar 71.56% dalam memprediksi sentimen bahasa Inggris. Untuk itu pemilihan model terbaik dalam klasifikasi sentimen Inggris akan berdasarkan macro average F1 score. Data yang melalui tahap preprocessing cleansing mendapat skor F1 sebesar 67.41%. Setelah melalui tahap normalisasi skor F1 yang didapat sebesar 67.61%. Data yang telah melalui semua tahap preprocessing cleansing, normalisasi, stopwords removal, dan stemming mendapat skor F1 terbesar yaitu sebesar 68.96%, lebih tinggi 1.55% dari data yang hanya melalui tahap cleansing. Evaluasi model sentimen bahasa Inggris ditunjukkan pada Gambar 4.

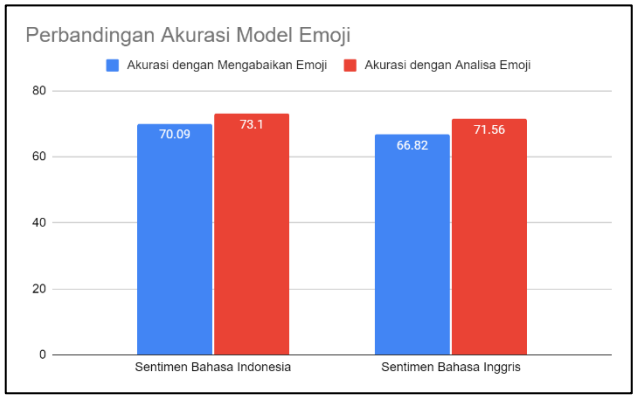


Gambar 4. Evaluasi Model Klasifikasi Sentimen Inggris

Perbedaan hasil performa model ini disebabkan oleh karateristik data yang berbeda-beda. Selain itu daftar kata yang terbatas dalam normalisasi baik bahasa Indonesia maupun bahasa Inggris juga berpengaruh dalam pembentukan model. Proses stopwords removal dan stemming justru juga membuat performa model menurun dalam klasifikasi bahasa maupun klasifikasi sentimen Indonesia. Hal ini dapat disebabkan proses filtering yang dilakukan stopwords removal dan stemming tidak tepat dan dapat menghapus kata-kata yang justru mempunyai makna penting dalam proses klasifikasi bahasa dan klasifikasi sentimen Indonesia. Namun, dalam klasifikasi sentimen Inggris proses stopwords removal dan stemming membantu meningkatkan performa model. Proses filtering dengan stopwords removal dan stemming cocok dengan data yang ada untuk mengklasifikan sentimen berbahasa Inggris.

5.5 Pengujian Model Sentimen dengan Mengabaikan Emoji pada Kalimat

Pengujian dilakukan dengan membandingkan akurasi pada model klasifikasi sentimen Indonesia dan sentimen Inggris dengan meggunakan emoji dan tanpa emoji (emoji diabaikan). Parameter SVM yang digunakan dalam pengujian menggunakan parameter terbaik untuk tiap model klasifikasi sentimen bahasa Indonesia dan klasifikasi sentimen bahasa Inggris. Klasifikasi sentimen bahasa Indonesia dan sentimen bahasa Inggris sama-sama menggunakan kernel RBF dengan nilai C dan gamma masing-masing 1. Perbandingan performa model dengan dan tanpa analisa emoji dapat dilihat pada Gambar 5.



Gambar 5. Perbaningan Akurasi Model Sentimen Emoji

Terlihat bahwa akurasi model untuk sentimen Bahasa Indonesia dan sentimen Bahasa Inggris dengan mengabaikan emoji masing-masing sebesar 70.09% dan 66.82%. Sedangkan untuk model dengan menganalisa emoji mendapat akurasi sebesar 73.10% untuk sentimen bahasa Indonesia dan 71.56% untuk sentimen bahasa Inggris. Terdapat perbedaan sebesar 3.01% untuk model sentimen Bahasa Indonesia dan 4.74% untuk model sentimen Bahasa Inggris. Terlihat bahwa model dengan analisa emoji memiliki performa lebih baik dibanding dengan model yang mengabaikan emoji. Rata-rata model dengan yang menganalisa emoji memiliki akurasi lebih tinggi sebesar 3.875% dibandingkan dengan model yang mengabaikan emoji.

6. KESIMPULAN

Tahapan kombinasi preprocessing untuk klasifikasi bahasa, sentimen Indonesia, maupun sentimen Inggris mempengaruhi peforma model menggunakan SVM. Akurasi terbaik yang didapatkan untuk klasifikasi model bahasa sebesar 88.77%. Model terbaik untuk klasifikasi sentimen bahasa Indonesia dan sentimen bahasa Inggris masing-masing memiliki akurasi sebesar 73.10% dan 71.56%. Proses *preprocessing* yang cocok untuk model klasifikasi bahasa adalah proses *cleansing*. Tahap preprocessing cleansing dan normalisasi menghasilkan model dengan performa terbaik untuk klasifikasi sentimen Indonesia. Model klasifikasi sentimen Inggris mendapat performa terbaik ketika data telah melalui tahap *preprocessing cleansing*, normalisasi, *stopwords removal*, dan *stemming*. Emoji sebaiknya tidak diabaikan karena modedl dengan melakukan analisa emoji mendapatkan performa lebih baik sebesar 3.875% dibandingkan dengan model yang mengabaikan emoji.

7. REFERENSI

[1] Ayvaz, S., & Shiha, M. O. 2017. The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering*, 9(1), 360–369. DOI= https://doi.org/10.17706/ijcee.2017.9.1.360-369.

[2] Badan Pusat Statistik. *Badan Pusat Statistik*. URI= https://www.bps.go.id/pressrelease/2021/01/21/1854/hasil-sensus-penduduk-2020.html.

[3] Baeldug. 2021. *Multiclass Classification Using Support Vector Machines*. URI= https://www.baeldung.com/cs/svm-multiclass-classification.

[4] Bird, S., Klein, E., & Loper, E. 2009. *Natural Language Processing with Python* (J. Steele (ed.)). O'Reilly Media, Inc.

[5] Christianto, M., Andjarwirawan, J., & Tjondrowiguno, A. (2020). Aplikasi analisa sentimen pada komentar berbahasa Indonesia dalam objek video di website YouTube menggunakan metode Naïve Bayes classifier. *Jurnal Infra*, 8.1, 255–259.

[6] Deshwal, V., & Sharma, M. 2019. Breast Cancer Detection using SVM Classifier with Grid Search Technique. *International Journal of Computer Applications*, *178*(31), 18–23. DOI= https://doi.org/10.5120/ijca2019919157.

[7] Joachims, T. 2001. *Learning To Classify Text Using Support Vector Machine*. Library of Congress Cataloging-in-Publication Data. DOI= https://doi.org/10.1007/978-1-4615-0907-3.

[8] Kumari, U., Sharma, D. A. K. S., & Soni, D. 2017. Sentiment Analysis of Smart Phone Product Review using SVM Classification Technique. *International Conference on Energy, Comunication, Data Analytics and Soft Computing (ICECDS)*, 1469–1474. DOI= https://doi.org/10.1109/ICECDS.2017.8389689.

[9] Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., & Nugraha, N. A. S. 2019. Sentiment Analysis of Cyberbullying on Instagram User Comments. *Journal of Data Science and Its Applications*, *2*(1), 88–98. DOI= https://doi.org/10.21108/jdsa.2019.2.20.

[10] Rahat, A. M., Kahir, A., & Masum, A. K. M. 2020. Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, *June 2020*, 266–270. DOI= https://doi.org/10.1109/SMART46866.2019.9117512.

[11] Rakhmanov, O. 2020. A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. *Procedia Computer Science*, *178*, 194–204. DOI= https://doi.org/10.1016/j.procs.2020.11.021.

[12] Rana, S., & Singh, A. 2017. Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques. *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016*, *October*, 106–111. DOI= https://doi.org/10.1109/NGCT.2016.7877399.

[13] Ross, S. (019. Being Real on Fake Instagram: Likes, Images, and Media Ideologies of Value. *Journal of Linguistic Anthropology*, *29*(3), 359–374. DOI= https://doi.org/10.1111/jola.12224.

[14] Sastrawi. 2015. *sastrawi: High quality stemmer library for Indonesian Language (Bahasa)*. URI= https://github.com/sastrawi/sastrawi.

[15] Statista. 2021. *Instagram: users by country*. URI= https://www.statista.com/statistics/578364/countries-with-most-instagram-users.

[16] Syarif, I., Prugel-Bennett, A., & Wills, G. 2016. SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(4), 1502. DOI= https://doi.org/10.12928/telkomnika.v14i4.3956.

[17] Tane, O. Z. A., Lhaksmana, K. M., & Nhita, F. 2019. Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM (Support Vector Machine). *Seminar Nasional Teknologi Fakultas Teknik Universitas Krisnadwipayana*, *1*(1), 739–742.