Penerapan Metode *Information Gain* dan *Support Vector Machine* untuk Identifikasi Konten Kasar pada *Tweets*

Kevin Aprilion[1], Ken Ratri Retno Wardani, S.Kom, M.T.[2]

Program Studi Informatika, Institut Teknologi Harapan Bangsa Bandung, Indonesia

[1]kevinaprilion@gmail.com

[2]ken_ratri@ithb.ac.id

Abstract*— In the use of social networks, especially Twitter, they often find posts that contain harsh and offensive words in Indonesian. This study will test the accuracy of Information Gain, the TF-IDF method and test the parameters on the SVM method to find out the best combination and parameters to find out. This study has several main indicators, namely using stemming or not using stemming when preprocessing the type of feature capture, parameters C and gamma, and the type of kernel. The data used is obtained from the website of the dataset provider, in the form of a Twitter data collection containing harsh words. Based on the test, the highest accuracy result was obtained with a value of 93.14% with f-measure 90.78% the results obtained using the Twitter dataset that were not manually labeled the method used was Information Gain with stemming . From this research, it can be said that using stemming and feature retrieval can affect the accuracy value in detecting the abusive words. Factors such as kernel and SVM parameters are also one of the things that affect the results of the detection of these abusive words.

Keywords*— Keyword: Identification of Abusive Word, Word Weight, TF-IDF, Information Gain (IG), Support Vector Machine (SVM)

Abstrak*— Dalam penggunaan jejaring sosial apalagi Twitter sering sekali menemukan tulisan yang mengandung kata-kata kasar dan bersifat ofensif dalam Bahasa Indonesia. Penelitian ini akan menguji akurasi metode *Information Gain*, metode TF-IDF dan menguji parameter pada metode SVM untuk mengetahui kombinasi metode dan parameter terbaik untuk identifikasi. Pada penelitian ini memiliki beberapa indikator utama yaitu menggunakan stemming atau tidak menggunakan stemming pada saat preprocessing jenis pengambilan fitur, parameter C dan *gamma*, dan jenis *kernel*. Data yang digunakan didapat dari website penyedia dataset, berupa kumpulan data Twitter yang terdapat kata kasar. Berdasarkan pengujian, hasil akurasi tertinggi didapatkan dengan nilai 93.1402% dengan *f-measure* 90.7787% hasil ini didapat dengan menggunakan dataset Twitter yang tidak dilakukan pelabelan manual metode yang gunakan adalah *Information Gain* dengan *stemming*. Dari penelitian ini, dapat disimpulkan bahwa menggunakan stemming dan pengambilan fitur dapat mempengaruhi nilai akurasi dalam pendeteksian kata kasar tersebut. Faktor seperti jenis kernel dan parameter SVM juga menjadi salah satu hal yang mempengaruhi hasil pendeteksian kata kasar tersebut.

Kata kunci: Identifikasi Kata Kasar, Pembobotan Kata, TF-IDF, *Information Gain* (IG), *Support Vector Machine* (SVM).

I. PENDAHULUAN

Seiring dengan pertumbuhan teknologi yang berkembang sangat pesat dan kemudahan menggunakan internet menimbulkan pertumbuhan pengguna di internet khususnya di indonesia. Menurut survei pengguna aktif di indonesia berkembang hingga penggunanya mencapai 170 juta pengguna [1]. Namun demikian, tidak semua individu pengguna Twitter bersikap bijak dalam memilih kata-kata dalam cuitannya. Tidak sedikit netizen menuliskan kata-kata yang berbau SARA (Suku, Agama, Ras, dan Antar golongan) atau bahkan mengungkapkan ekspresi dengan menuliskan kata-kata kasar dan bersifat ofensif. Kata-kata kasar biasanya diucapkan atau dituliskan untuk menyerang pihak tertentu seperti ke perorangan atau sebuah instansi, untuk mengungkapkan kekesalan, kekecewaan, atau untuk meluapkan emosi terhadap peristiwa tertentu [2].

Menurut laporan terbaru dari *We Are Social*, perusahaan asal Inggris bekerja sama dengan Hootsuite yang melakukan survei per Januari 2021, mendapatkan kesimpulan bahwa pengguna aktif media sosial bertambah 6.3 % atau mengalami kenaikan dari januari 2020 mencapai 10 juta pengguna [3]. Jika dilihat dari jumlah populasi di dunia ada di sekitar 274.9 juta jiwa, maka 61.8 % di antaranya adalah pengguna aktif di media sosial. Sementara itu, diketahui, pengguna internet Indonesia terkini mencapai 202.6 juta dan 170 juta diantaranya pengguna aktif. *We Are Social* dan Hootsuite mengungkapkan jumlah pengguna Twitter di Indonesia adalah 14.05 juta \[3].

Twitter Merupakan sebuah situs *mikroblog* yang mana situs *mikroblog* berevolusi menjadi sebuah sumber informasi karena *mikroblog* tempat orang bisa memposting pesan secara *real time* tentang pendapat mereka tentang berbagai topik. Maka dari itu para pengguna Twitter *men-twee*t informasi dengan berbagai topik secara *real time*, para pengguna Twitter sering berdiskusi tentang permasalahan yang baru atau yang sedang hangat dan mengungkapkan opini dengan sentimen positif atau negatif tentang topik tersebut yang dekat dengan aktivitas mereka [4].

Seleksi fitur sangat penting untuk klasifikasi text dan bagus untuk mempengaruhi performa klasifikasi. *Support Vector Machine* akan bermasalah ketika memilih parameter atau fitur yang sesuai untuk memilih fitur dalam SVM sangat berpengaruh hasil akurasi klasifikasi maka dari itu seleksi fitur itu sangat penting untuk text *classification* dan mempengaruhi performa dan seleksi fitur yang cocok untuk kasus ini adalah *Information Gain* [5]. Hasil penelitian yang dilakukan mendapatkan akurasi SVM sebesar 83.05% dan ketika menggunakan *Information Gain* menghasilkan akurasi sebesar 85.65% mengalami peningkatan dari SVM [5]. Pada Jurnal penelitian lain juga membuktikan *Information Gain* merupakan model yang lebih baik dibandingkan *Chi Squared Statistic* dalam meningkatkan tingkat akurasi SVM, yaitu dengan menghasilkan rata-rata kenaikan tingkat akurasi sebesar 2.514% dengan tingkat akurasi terbaik sebesar 72.45% [6]..

Berdasarkan penelitian terdahulu yang sudah dibahas secara singkat sebelumnya, pada penelitian ini akan mencoba untuk memaksimalkan hasil pendeteksian. Peneliti akan membandingkan pengambilan fitur dengan menggunakan seleksi fitur *Information Gain* (IG) dan ekstraksi fitur TF-IDF untuk meningkatkan akurasi dari pendeteksian dengan *Support Vector Machine* (SVM) untuk mendeteksi suatu konten yang menggunakan kata kasar di media Twitter dengan bahasa Indonesia.

II. Metodologi

A. Information Gain

Information Gain (IG) atau dikenal juga *Mutual Information* (MI) merupakan salah satu metode untuk melakukan seleksi fitur, untuk menentukan batas dari kepentingan sebuah atribut [8]. Nilai IG diperoleh dari nilai *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan [8]. Nilai *entropy* digunakan untuk menghitung sebagai penentuan atribut mana yang akan dibuang dan digunakan. Atribut yang memenuhi kriteria pembobotan nantinya akan digunakan untuk proses klasifikasi [9]..

Untuk memudahkan klasifikasi bisa menggunakan *confusion matrix* untuk menempatkan nilai jumlah kemunculan terms(kata/istilah) seperti:

Tabel 1 *Confusion matrix* untuk *information gain*

	ec = 1	ec = 0
et = 1	bb	bs
et = 2	sb	ss

Entropy berfungsi untuk menunjukkan nilai kemiripan varian pada suatu atribut. Nilai *gain* diperoleh dari total *entropy* mengurangkan atribut loyal *entropy*, sehingga diperoleh hasil perhitungan berupa nilai variasi pada masing-masing atribut. Nilai *gain* itu sendiri dihitung dengan menggunakan persamaan seperti Persamaan 1.

$$I(U,C) = \frac{N_{11}}{N}\log_2\left(\frac{N \cdot N_{11}}{N_1 \cdot N_1}\right) + \frac{N_{01}}{N}\log_2\left(\frac{N \cdot N_{01}}{N_0 \cdot N_1}\right) + \frac{N_{10}}{N}\log_2\left(\frac{N \cdot N_{10}}{N_1 \cdot N_0}\right) + \frac{N_{00}}{N}\log_2\left(\frac{N \cdot N_{00}}{N_0 \cdot N_0}\right) \quad (1)$$

B. TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) merupakan teknik pembobotan berbasis statistik yang memberikan bobot pada tiap kata dalam teks, bobot ini mencerminkan ciri khas dari kata tersebut di dalam teks [10]. Persamaan TF-IDF memiliki persamaan sebagai berikut:

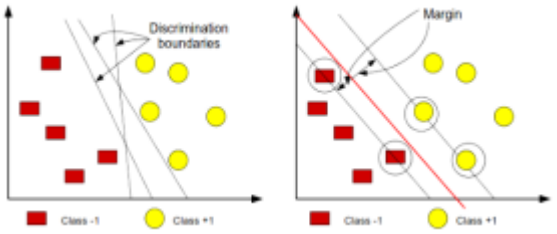
$$TF = \frac{KemunculanKataDalamDokumen}{JumlahKataDalamDokumen} \quad (2)$$

$$IDF = log\frac{JumlahDokumen}{JumlahDokumenYangMemuatKata} \quad (3)$$

$$TF - IDF = TFXIDF \quad (4)$$

C. Support Vector Machine

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992.
% sebagai rangkaian konsep – konsep unggulan dalam bidang *pattern recognition*.
SVM masih terhitung salah satu metode yang dibilang baru dan masih termasuk dari metode dari *machine learning*. *Hyperplane* disebut optimal jika memiliki margin yang maksimal. Untuk mendapatkan margin yang maksimal dibutuhkan titik kelas yang paling dekat dengan garis *hyperplane*, titik itu disebut titik *support vector* [11].



Gambar 1 *Hyperlane Class* -1 dan +1

Pada Gambar 1 memperlihatkan terdapat pola-pola sebagai bagian dari anggota pola lain yang terdiri dari dua buah *class* yang mempunyai bobot nilai +1 dan -1. Dalam menentukan suatu nilai
pembobotan kelas positif dan negatif atau sebaliknya,
dalam SVM ditentukan berdasarkan jika nilai bobot
lebih dari 0 maka diklasifikasikan ke dalam positif dan
sebaliknya jika nilai bobot kurang dari 0 maka

diklasifikasikan kedalam kelas negatif.

$$w^T \times x + b = 0 \quad (5)$$
$$w^T \times x + b = 1 \quad (6)$$
$$w^T \times x + b = -1 \quad (7)$$

Untuk mendapatkan nilai w, digunakan persamaan berikut:

$$W = \sum_{i=1}^{n} \alpha_i Y_i X_i \quad (8)$$

III. PERANCANGAN SISTEM

Sistem identifikasi konten kasar di Twitter menggunakan dua dataset yaitu dataset asli dan dataset yang dilakukan pelabelan manual dan terbagi tiga proses yaitu proses *preprocessing, training, dan testing*. Proses *preprocessing* adalah proses untuk membersihkan data dari *noise.* Proses *training* dilakukan untuk mendapatkan fitur yang berpengaruh terhadap konten kasar dan model prediksi. Proses *testing* dilakukan untuk menghitung hasil prediksi yang berupa akurasi dan *f-measure.*

A. Data sampling

Dataset diambil dari Kaggle akan menggunakan dan dibagi menjadi dua jenis dataset yaitu dataset asli dan dataset yang sudah diubah dengan melakukan pelabelan secara manual. Dataset hanya menggunakan 2 kolom yaitu *Tweet* dan *Abusive*.

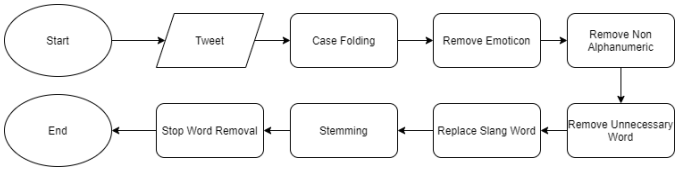
Dataset Asli adalah dataset yang diambil dari Kaggle dan tidak diubah sama sekali dalam penelitian ini untuk menjadi perbandingan dengan dataset pelabelan manual.

Dataset pelabelan manual adalah dataset yang dibuat

t dari dataset asli yang akan diperiksa ulang dan diubah jika tidak sesuai.

B. Proses *Preprocessing*

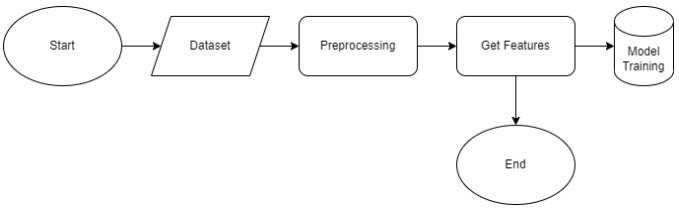
Proses *preprocessing* adalah proses dimana semua data akan dibersihkan dari *noise* agar dataset yang dipakai lebih mudah untuk dilakukan klasifikasi.



Gambar 2 *Flowchart preprocessing*

C. Proses *Training*

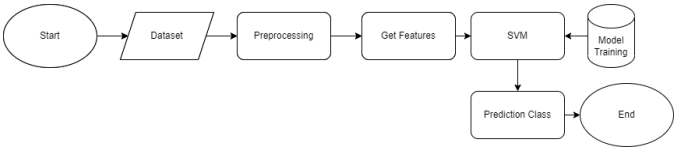
Proses *training* adalah proses pelatihan data sebelum proses uji untuk mendapatkan model yang nantinya model tersebut akan digunakan pada proses pengujian menggunakan SVM. diikuti oleh sebuah tanda kurung kanan. *Heading* level 3 harus diakhiri dengan titik dua. Isi dari bagian level-3 langsung mengikuti (bersambung) *heading* level-3 pada paragraf yang sama. Contohnya, paragraf ini mulai dengan *heading* level 3.



Gambar 3 *Flowchart training*

D. Proses *Testing*

Proses *Testing* akan dilakukan untuk mengetahui hasil dari pengujian.



Gambar 3 *Flowchart Testing*

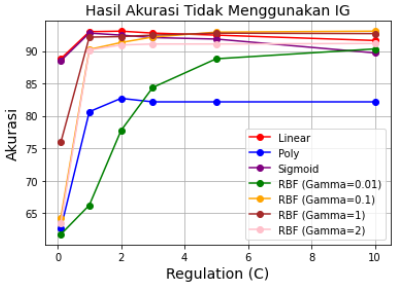
IV. PERANCANGAN SISTEM

Bagian ini akan menjelaskan mengenai hasil dari nilai akurasi dan *f-measure* mengenai pengujian yang akan dilakukan oleh sistem pendeteksian kata kasar untuk membanding penggunaan SVM dan *Information Gain*. Skenario pengujian akan membandingkan *kernel* metode SVM saat klasifikasi. Data *training* juga akan dibagi menjadi dua jenis yaitu menggunakan TF-IDF dan yang *Information Gain* untuk sebagai perbandingan metode mana yang terbaik saat mengolah data. Data *training* tersebut akan dikombinasikan dengan skenario pengujian pada Tabel 1.

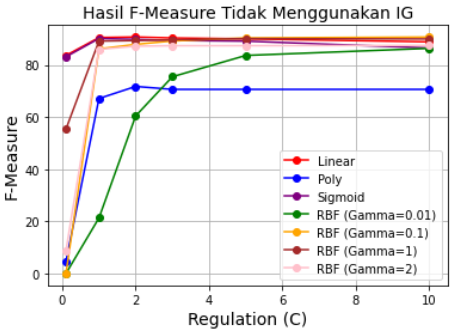
Tabel 4 Skenario pengujian parameter dan kernel SVM

Kernel	C	Gamma
Linear	0.1, 1, 2, 3, 5 dan 10	
Sigmoid	0.1, 1, 2, 3, 5 dan 10	
Polynomial	0.1, 1, 2, 3, 5 dan 10	
RBF	0.1, 1, 2, 3, 5 dan 10	0.01, 0.1, 1 dan 2

A. Dataset Asli Menggunakan *Stemming* (TF-IDF)



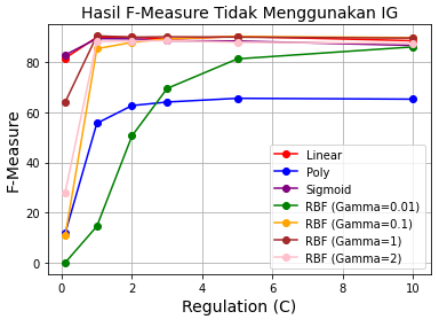
Gambar 4 Hasil akurasi dataset asli menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*



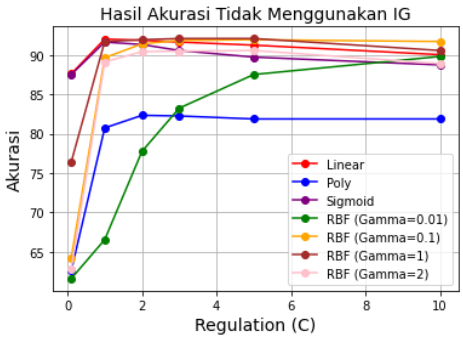
Gambar 5 Hasil *f-measure* dataset asli menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*

Dilihat pada Gambar 4 dan 5 untuk ekstraksi fitur dataset asli dengan menggunakan *Stemming* tanpa menggunakan IG tetapi menggunakan TF-IDF didapat bahwa akurasi terbaik yang didapat linear dengan parameter C=2 yaitu sebesar 93.0640% dan *f-measure* yang di dapat 90.6858%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kernel sigmoid* , RBF dan linear mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

B. Dataset Asli Tidak Menggunakan *Stemming* (TF-IDF)



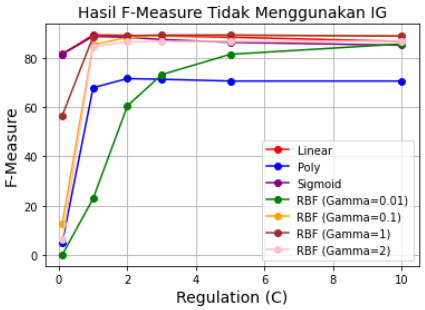
Gambar 6 Hasil akurasi dataset asli tidak menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*



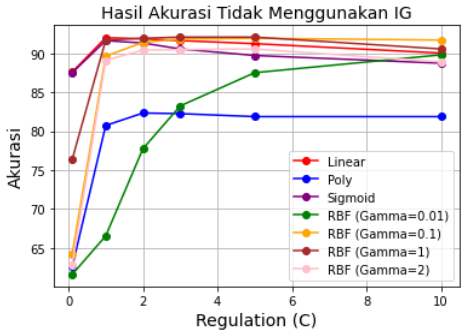
Gambar 7 Hasil *f-measure* dataset asli tidak menggunakan Stemming pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*

Dilihat pada Gambar 6 dan 7 untuk ekstraksi fitur dataset asli dengan tidak menggunakan *Stemming* tanpa menggunakan IG tetapi menggunakan TF-IDF didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=1 dan *gamma*=1 yaitu sebesar 92.9116% dan *f-measure* yang di dapat 90.4811%. Pada k*ernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kernel* sigmoid dan linear mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C=10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

C. Dataset Pelabelan Manual Menggunakan *Stemming* (TF-IDF)



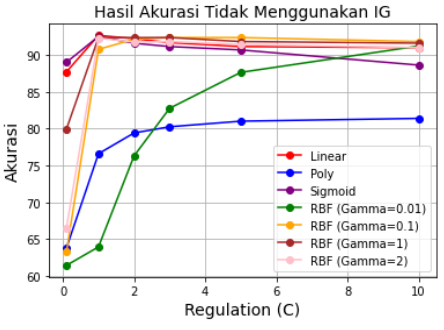
Gambar 8 Hasil akurasi dataset pelabelan manual menggunakan Stemming Pengujian *kernel* dan parameter tanpa menggunakan Information Gain



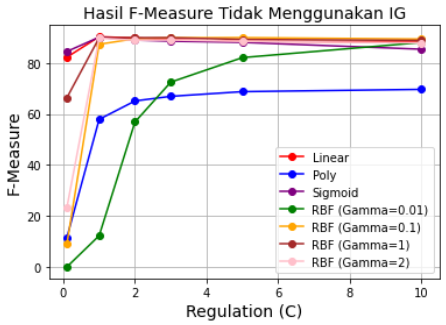
Gambar 9 Hasil *f-measure* dataset asli tidak menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*

Dilihat pada Gambar 8 dan 9 untuk ekstraksi fitur dataset pelabelan manual dengan menggunakan *Stemming* tanpa menggunakan IG tetapi menggunakan TF-IDF didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=3 dan 4 dengan *gamma* = 1 yaitu sebesar 92.0732% dan f-measure yang di dapat 89.3004%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kerne} sigmoid* dan RBF mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

D. Dataset Pelabelan Manual Tidak Menggunakan *Stemming* *(TF-IDF)*



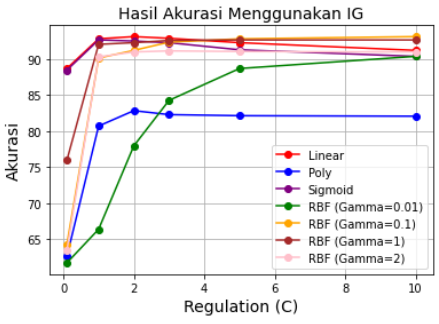
Gambar 10 Hasil akurasi dataset pelabelan manual tidak menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*



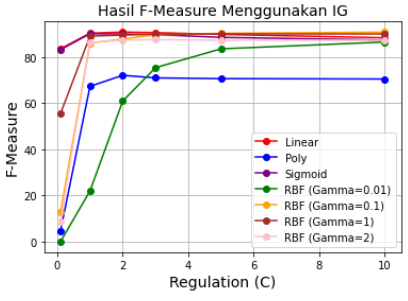
Gambar 11 Hasil *f-measure* dataset pelabelan manual tidak menggunakan *Stemming* pengujian *kernel* dan parameter tanpa menggunakan *Information Gain*

Dilihat pada Gambar 10 dan 11 untuk ekstraksi fitur tanpa menggunakan IG tetapi menggunakan TF-IDF didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=3 dengan *gamma* = 0.01, 0.1, dan 1 yaitu sebesar 92.3780% dan *f-measure* yang di dapat 89.8990%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kernel sigmoid* dan RBF mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

E. Dataset Asli Menggunakan *Stemming* (IG)



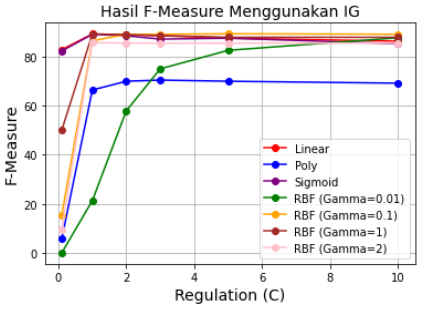
Gambar 12 Hasil akurasi dataset asli menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*



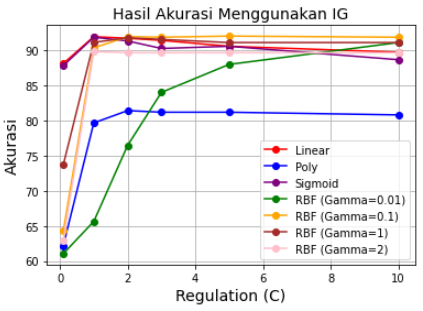
Gambar 12 Hasil *f-measure* dataset asli menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*

Dilihat pada Gambar 12 dan 13 untuk seleksi fitur dataset asli dengan menggunakan *Stemming* menggunakan IG didapat bahwa akurasi terbaik yang didapat oleh *kernel* linear dengan parameter C=2 dan RBF dengan parameter C=10 dengan *gamma*=0.1 yaitu sebesar 93.1402% dan *f-measure* yang di dapat 90.7787%. Pada *kernel* Linear dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* RBF. Dari hasil parameter C=3 didapatkan *kernel* RBF dan linear mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measur*e pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

F. Dataset Asli Tidak Menggunakan *Stemming* (IG)



Gambar 14 Akurasi dataset asli tidak menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*

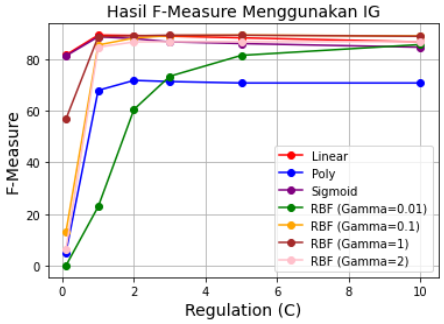


Gambar 15 Hasil *f-measure* dataset asli tidak menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*

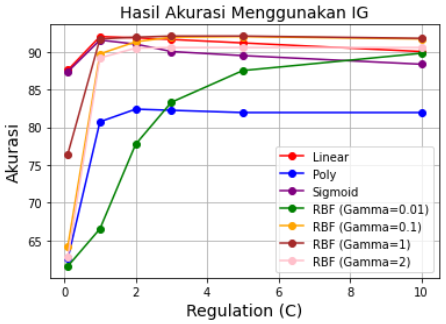
Dilihat pada Gambar 14 dan 15 untuk seleksi fitur dataset asli dengan tidak menggunakan *Stemming* menggunakan IG didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=5 dan *gamma* = 0.1 yaitu sebesar 91.9970% dan *f-measure* yang di dapat 89.3617%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kernel*

sigmoid dan RBF mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

G. Dataset Pelabelan Manual Menggunakan Stemming (IG)



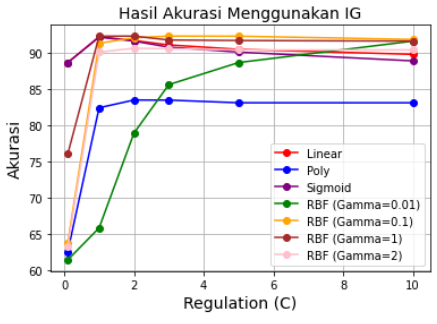
Gambar 16 Hasil akurasi dataset pelabelan manual menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*



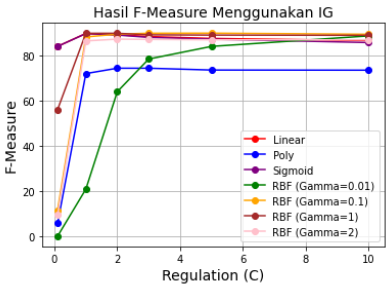
Gambar 17 Hasil *f-measure* dataset pelabelan manual menggunakan stemming pengujian *kernel* dan parameter menggunakan *Information Gain*

Dilihat pada Gambar 16 dan 17 untuk seleksi fitur dataset pelabelan manual dengan menggunakan *Stemming* menggunakan IG didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=3 dengan *gamma*=1 yaitu sebesar 92.0732% dan *f-measure* yang di dapat 89.3004%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh *kernel* linear. Dari hasil parameter C=3 didapatkan *kernel* sigmoid dan RBF mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

H. Dataset Pelabelan Manual Tidak Menggunakan *Stemming* (IG)



Gambar 18 Hasil Akurasi dataset pelabelan manual tidak menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*



Gambar 19 Hasil *f-measure* dataset pelabelan manual tidak menggunakan *Stemming* pengujian *kernel* dan parameter menggunakan *Information Gain*

Dilihat pada Gambar 18 dan 19 untuk seleksi fitur menggunakan IG didapat bahwa akurasi terbaik yang didapat RBF dengan parameter C=1 dengan *gamma* = 1 yaitu sebesar 92.3780% dan f-*measure* yang di dapat 89.8990%. Pada *kernel* RBF dapat dilihat memiliki akurasi rata-rata terbaik lalu disusul oleh k*ernel* linear. Dari hasil parameter C=3 didapatkan *kernel sigmoid* dan RBF mengalami akurasi tertinggi sedangkan. Hampir semua metode mengalami penurunan akurasi dan *f-measure* pada C = 10. Untuk skenario pengujian pendeteksian kata kasar metode *polynomial* mendapatkan rata-rata akurasi yang paling kecil.

V. KESIMPULAN

Kesimpulan dari pengujian untuk sistem identifikasi konten kasar di Twitter adalah:

1. Stemming dan tidak menggunakan *Stemming* sangat mempengaruhi akurasi data tetapi saat menggunakan *Stemming* menghasilkan akurasi paling tinggi 93.1402% lebih baik dibanding tidak menggunakan *Stemming* dengan akurasi paling tinggi 92.3780%.
2. Dari hasil pengujian yang diuji, menggunakan *Information Gain* mendapatkan hasil akurasi dan nilai *f-measure* paling tinggi dengan menggunakan *kernel* linear . Nilai akurasi dan nilai *f-measure* paling tinggi yang didapat adalah sebesar 93.1402% dan f-measure 90.7787% sedangkan menggunakan TF - IDF mendapatkan akurasi sebesar 93.0640% dan *f-measure* 90.6858% linear perbedaan nya hanya 0.0762% saja. Tetapi karena akurasi keduanya sudah mencapai 90% kenaikan 0.0762% saja sudah sangat besar.

3. Penggunaan *kernel polynomial* menjadi *kernel* yang memiliki nilai akurasi dan *f-measure* yang paling kecil untuk penelitian ini. Nilai akurasi dan nilai *f-measur*e paling kecil di dapat oleh kernel *polynomial* bernilai 61.0518% dan 0% pada model Dataset Original No *Stemming* (TF-IDF) dan Dataset Original No *Stemming* (IG).
4. Kombinasi Parameter C terbaik adalah 3 karena jarak *hyperlane* tidak terlalu jauh dan tidak terlalu dekat. *gamma* terbaik berada di 1 karena radius kesamaanya tidak terlalu kecil dan tidak terlalu besar.

VI. Saran

Saran untuk sistem identifikasi konten kasar di Twitter adalah:

1. Perbedaan jumlah dataset menggunakan kata kasar dan tidak menggunakan kata kasar pada penelitian selanjutnya disarankan menggunakan dataset yang memiliki jumlah kelas yang setara.
2. Untuk mengatasi kata dengan banyak arti. Menggunakan jenis ekstraksi fitur yang lain seperti Algoritma *decision tree*, karena dapat merepresentasikan kata dengan lebih terstruktur dan multilabel [12].
3. Media sosial akan makin berkembang terus dan akan muncul kata atau istilah baru begitu juga penggunaan kalimat yang menggunakan satir dan sarkas yang bermaksud untuk mengumpat maka dari itu untuk penelitian selanjutnya harus memperbarui kata atau istilah dan mengklasifikasikan kalimat supaya bisa dimengerti sistem.
4. Untuk parameter *gamm*a tidak ada aturan khusus untuk nilai parameter yang akan digunakan tetapi penggunaan nilai parameter disarankan untuk tidak terlalu berdekatan supaya tidak menyebabkan *overfitting* karena untuk nilai *gamma* yang tinggi akan membuat titik-titik *support vector* sangat dekat satu sama lain lebih baik seperti 0.001, 0.01, 0.1, 1, 10, 100.

Daftar Referensi

[1] Agus Tri Haryanto, "Pengguna Aktif Medsos RI 170 Juta, Bisa Main 3 Jam Sehari", Detik ,2021. [Online]. Available: https://inet.detik.com/cyberlife/d-5407834/pengguna-aktif-medsos-ri-170-juta-bisa-main-3-jam-sehari. [Accessed: 05 April 2021].

[2] R F Hidayattullah, A A F Yusuf, K P Juwairi, R A N Nayoan, "Identifikasi Konten Kasar pada *Tweet* Bahasa Indonesia", Journal Komputasional (JLK), Vol. 2, No. 1, Maret 2019..

[3] R Kusumawati, A D 'arofah and P A Pramana, "Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services", Journal of Physics: Conference Series, 1320 (2019) 012016 doi:10.1088/1742-6596/1320/1/012016.

[4] Xiaoamei Zou, Jing Yang. 2017 "Microblog Sentiment Analysis Using Social and Topic Context". PLoS One 13(2):e0191163, https://doi.org/10.1371/nejournal.po.0191163.

[5] Reza Maulana, Panny Agustia R., Windi Irmayani, Dedi Saputra, Wanty Eka Jayanti, "*Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain*". *Journal of Physics: Conference Series*, 1641 (2020) 012060 doi:10.1088/1742-6596/1641/1/012060.

[6] Oman Somantril, Dyah Apriliani , "*Support Vector Machine* Berbasis *Feature Selection* untuk *Sentiment Analysis* Kepuasan Pelanggan Terhadap Pelayanan Warung Dan Restoran Kuliner Kota Tegal". Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), Vol.5, No.5 Oktober 2018, hlm.537-548 p-ISSN: 2355-7699.

[7] Christevan Destitus, Wella, Suryasari, "*Support Vector Machine* VS *Information Gain*: Analisis Sentimen *Cyberbullying* Twitter Indonesia". ULTIMA InfoSys, Vol. XI, No. 2 Desember 2020, ISSN 2085-4579.

[8] A. Paul. (Okt 1987). “Electrical properties of flying machines.” *Flying Machines*. [Daring]. 38(1), hlm. 778-998. Tersedia: www.flyingmachjourn/properties/fly.edu [1 Des 2003]

[9] M. O. Ibrohim and I. Budi, ”*Multi-label Hate Speech and Abusive Language Detection in Indonesia Twitter*”, ALW3: 3rd *Workshop on Abusive Language Online,* 46-57. 2019.

[10] L. Derczynski, "*Complementarity, F-score, and NLP Evaluation", Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC' 16), pp. 261-266, 2016.

[11] C. Stocan and R. Stoean, *Support vector machines and evolutionary algorithms for classification*. Cham: Springer, 2014.

[12] Fauzi Ihsan, Iwan Iskandar, N. S. Harahap, Surya Agustian, ”Algoritme *decision tree* untuk mendeteksi ujaran kebencian dan bahasa kasar multilabel pada Twitter berbahasa Indonesia ”, Jurnal Teknologi dan Sistem Komputer, ISSN: 2338-0403.