Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 11, November 2018, hlm. 5326-5332 http://j-ptiik.ub.ac.id

Identifikasi *Tweet Cyberbullying* pada Aplikasi Twitter menggunakan Metode *Support Vector Machine* (SVM) dan *Information Gain* (IG) sebagai Seleksi Fitur

Ni Made Gita Dwi Purnamasari[1], M. Ali Fauzi[2], Indriati[3], Liana Shinta Dewi[4]

[1,2,3]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
[4]Program Studi Pendidikan Bahasa dan Sastra Indonesia, Fakultas Ilmu Budaya, Universitas Brawijaya
Email: [1]workpurnamasari@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]indriati.tif@ub.ac.id, [4]lianashinta@gmail.com

Abstrak

Cyberbullying merupakan salah satu tindakan yang melanggar UU ITE dimana kejahatan ini dilakukan di media sosial salah satunya aplikasi Twitter. Tindakan ini sulit terdeteksi jika tidak ada yang me-*report tweet* tersebut. Identifikasi *tweet cyberbullying* bertujuan untuk mengklasifikasikan *tweet* yang mengandung konten *bullying*. Klasifikasi dilakukan dengan menggunakan metode *Support Vector Machine* dimana metode bertujuan mencari *hyperplane* pemisah antara kelas negatif dan positif. Penelitian ini merupakan klasifikasi teks dimana semakin banyak datanya semakin banyak fitur yang dihasilkan, oleh karena itu penelitian ini juga menggunakan seleksi fitur *Information Gain* untuk menyeleksi fitur yang tidak relevan terhadap klasifikasi. Proses sistem dimulai dari *text preprocessing* dengan tahapan tokenisasi, *filtering*, *stemming* dan pembobotan kata. Kemudian melakukan seleksi fitur *information gain* dengan menghitung nilai *entropy* tiap kata. Setelah itu melakukan proses klasifikasi berdasarkan fitur yang telah diseleksi dan hasil keluaran sistem berupa identifikasi apakah *tweet* termasuk *bully* atau bukan *bully*. Hasil yang didapatkan dengan metode SVM adalah *accuracy* 75%, *precision* 70,27%, *recall* 86,66% dan *f-measur*e 77,61% pada percobaan nilai *iterMax* = 20, λ = 0,5, γ = 0,001, ε = 0,000001, dan C = 1. Nilai *threshold* terbaik seleksi fitur *information gain* adalah 90%, dengan nilai *accuracy* 76,66%, *precision* 72,22%, *recall* 86,66% dan *f-measure* 78,78%.

Kata kunci: *Cyberbullying, Klasifikasi, Support Vector Machine, Information Gain.*

Abstract

Cyberbullying is one of the actions that violate the ITE Law where the crime is committed on social media applications such as Twitter. This action is difficult to detect if no one is reporting the tweet. Cyberbullying tweet identification aims to classify tweets that contain bullying. Classification is done using Support Vector Machine method where this method aims to find the dividing hyperplane between negative and positive class. This study is a text classification where more data is used, the more features are produced, therefore this research also uses Information Gain as feature selection to select features that are not relevant to the classification. The process of the system starts from text preprocessing with tokenizing, filtering, stemming and term weighting. Then perform the information gain feature selection by calculating the entropy value of each term. After that perform the classification process based on the terms that have been selected, and the output of the system is identification whether the tweet is bullying or not. The result of using SVM method is accuracy 75%, precision 70.27%, recall 86.66% and f-measure 77.61% on experiment iterMax value = 20, λ = 0.5, γ = 0.001, ε = 0.000001, and C = 1. The best threshold of information gain is 90%, with accuracy 76.66%, precision 72.22%, recall 86.66% and f-measure 78.78%

Keywords: *Cyberbullying, Classification, Support Vector Machine, Information Gain.*

1. PENDAHULUAN

Penyalahgunaan tentang sosial media diatur pada Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) nomor 11 tahun 2008 pasal 27 ayat 3 yang menyatakan tentang penyebaran nama baik atau penghinaan. *Cyberbullying* termasuk ke dalam kategori tersebut dimana terdapat penghinaan. Saat ini KOMINFO tengah bekerja sama dengan Google dan Twiiter dalam pemberantasan konten-konten negatif di internet seperti pornografi, *hoax*, *cyberbullying*, dan lain-lain. KOMINFO berharap agar pencegahan penyebaran konten negatif tersebut cepat diselesaikan. Pihak Twitter pun sudah menyediakan fungsi dimana report tentang konten negatif tersebut secara tersendiri agar lebih cepat ditanggapi (Devaga, 2017). Oleh karena itu dengan adanya sistem otomatis untuk identifikasi *tweet cyberbullying* dapat membantu dalam penanganan *tweet* pengujar kebencian lebih cepat diatasi dan lebih efisien.

Penelitian oleh Miftah Andriansyah tahun 2017 tentang *cyberbullying* yang berjudul *Cyberbullying Comment Classification on Indonesian Selebgram Using Support Vector Machine (SVM) Method* dimana *cyberbullying* di fokuskan pada *comment* sebuah akun seorang selebgram (selebritis Instagram) di aplikasi Instagram. Data yang digunakan diambil dari *comment section* salah satu foto yang diunggah oleh akun selebgram Awkarin, dan hasil akurasi yang didapatkan cukup tinggi. Masalah dalam penelitian klasifikasi ini adalah ketika bahasa dari *comment* yang digunakan tidak termasuk dalam *bullying*, tetapi dalam artinya *comment* tersebut termasuk dalam bullying (Adriansyah, et al., 2017).

Penelitian tentang *cyberbullying* (Noviantho, 2017) yang berjudul *Cyberbullying Classification using Text Mining*, peneliti menggunakan metode *Support Vector Machine* (SVM) dan *Naive Bayes* untuk proses klasifikasi. Data yang digunakan didapatkan dari aplikasi Kaggle (www.kaggle.com), dimana aplikasi tersebut menyediakan total percakapan 1600 dari website Formspring.me yang menggunakan Bahasa Indonesia. Peneliti mengolah data dengan teknik *Preprocessing*, *Extraction*, *Classification* dan *Evaluation*. Peneliti juga membandingan hasil dari penelitian sebelumnya oleh Kelly Reynolds tahun 2012 yang menggunakan *Decision Tree* dan *K-Nearest Neighbor* (K-NN). Dari penelitian tersebut didapatkan hasil bahwa metode SVM lebih baik dalam pengklasifikasian *cyberbullying* dibanding metode K-NN dan *Decision Tree* yang digunakan oleh Kelly Reynolds.

Dalam komputasinya penggunaan metode SVM memakan waktu yang cukup lama tergantung pada banyak fitur yang digunakan. Dalam pengklasifikasian teks, setiap kata dari hasil pemrosesan teks akan dijadikan fitur, dengan begitu jika menggunakan dokumen yang banyak akan menghasilkan fitur yang banyak pula. Oleh karena itu seleksi fitur digunakan dalam penelitian ini untuk menyeleksi fitur-fitur yang relevan saja agar dapat mengurangi waktu dari komputasi metode SVM.

Pada penelitian tentang Komparasi Algoritma Klasifikasi *Machine Learning* dan *Feature Selection* pada Analisis Sentimen Review Film (Chandani, et al., 2015) menggunakan metode *Artificial Neural Network* (ANN), SVM dan *Naive Bayes* (NB), serta menggunakan seleksi fitur *Information Gain* (IG), *Chi Square, Forward Selection* dan *Backward Selection*. Data yang digunakan diambil dari situs *Internet Movie Database* (IMDb) berupa *review* film dari *user*. Dari penelitian tersebut didapatkan hasil bahwa metode SVM memiliki akurasi paling tinggi dibanding metode klasifikasi yang lainnya. Serta fitur seleksi IG menghasilkan akurasi yang lebih tinggi dibanding fitur seleksi yang lain yang digunakan pada metode SVM.

Pada penelitian ini metode yang digunakan adalah SVM. Tetapi dalam komputasi metode SVM tergolong memakan waktu yang cukup lama, maka dari itu peneliti menggunakan seleksi fitur untuk mengurangi fitur-fitur agar memperingan kerja komputasi dari metode SVM. Fitur seleksi yang digunakan adalah *Information Gain* (IG), dimana IG digunakan untuk menentukan batas dari sebuah fitur. Pada penelitian ini juga menganalisis pengaruh parameter metode SVM dan pengaruh penggunaan seleksi fitur IG, dengan tujuan dapat membangun sebuah system yang dapat mengidentifikasi *tweet cyberbullying*. Data yang digunakan diambil menggunakan R-Studio dari API Twitter dengan sejumlah 300

tweets yang me-*mentioned* Wakil Ketua DPR RI Bapak Fadli Zon.

2. DASAR TEORI

2.1 Twitter

Twitter merupakan salah satu layanan jejaring sosial yang masuk kedalam *Microblogging* atau ngeblog secara singkat dalam satu paragraf dengan maksimal 280 huruf (Motivasee, 2017).

2.2 Cyberbullying

Cyberbullying merupakan salah satu tindak kekerasan yang dilakukan oleh seseorang terhadap korbannya di internet, dimana korban dihina, diejek, dipermalukan dan diintimidasi oleh pelaku. *Cyberbullying* bisa berdampak pada mental korban, bahkan banyak dari korban *bullying* berakhir dengan bunuh diri karena tidak tahan dengan banyak tekanan (Oktaviani, 2013).

2.3 Text Preprocessing

2.3.1 Tokenisasi

Tokenisasi merupakan suatu proses pemotongan *string*/kata pada suatu kalimat dan semua tanda baca dan tanda hubung akan dihilangkan. Proses ini bertujuan untuk memisahkan tiap kata agar dapat membedakan karakter-karakter tertentu yang diperlakukan sebagai pemisah kata atau bukan. Proses tokenisasi mengandalkan karakter spasi pada dokumen sebagai pemisah kata (Garcia, 2005).

2.3.2 Filtering

Tahap *filtering* merupakan tahap mengambil kata-kata penting dari sebuah hasil token. Banyak kata yang paling sering digunakan dalam Bahasa Indonesia tidak berguna dalam *Information Retrieval* (IR) dan *text mining*, kata-kata ini disebut *Stopwords* (Gaigole, et al., 2013). Kamus stopwords yang digunakan didapatkan dari Tala. Misalnya adalah kata-kata "yang", "di", dan yang lainnya.

2.3.3 Stemming

Teknik *stemming* digunakan untuk mengetahui akar sebuah kata. *Stemming* dilakukan selain untuk memperkecil jumlah indeks berbeda dari suatu dokumen, juga dilakukan untuk mengempokkan kata-kata yang memiliki kata dasar dan arti yang serupa tetapi memiliki bentuk yang berbeda karena terdapat imbuhan yang berbeda (Gaigole, et al., 2013). Misalnya pada kata "bersama", "kebersamaan", "menyamai" akan di *stem* ke akar katanya yaitu "sama".

2.3.4 Term Weighting

Proses penghitungan TF-IDF (*Term frequency-Inverse Document Frequency*) dimana disini dilakukan proses penghitungan jumlah bobot tiap token hasil stemming yang nantinya akan dibandingkan dengan dokumen lain untuk membandingkan frequency dari jumlah kemunculan token tersebut. Rumus TF-IDF dapat dilihat dalam persamaan (1) dan persamaan (2).

$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}, & if\ tf_{t,d} > 0 \\ 0, & if\ tf_{t,d} = 0 \end{cases} \quad (1)$$

Dimana *tf* adalah *term frequency* yang menyatakan berapa banyak jumlah suatu term dalam sebuah dokumen.

$$idf_t = log_{10}(\frac{N}{df_t}) \quad (2)$$

Dimana N merupakan jumlah banyak dokumen, karena terkadang suatu term muncul di beberapa dokumen sehingga proses pencarian *term* unik akan terganggu. IDF befungsi untuk mengurangi bobot suatu term jika kemunculannya dari term tersebut banyak dan tersebar diseluruh dokumen (Gaigole, et al., 2013).

2.4 Support Vector Machine

Metode SVM dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan tahun 1992 pada Annual Workshop on Computational Learning Theory (Nugroho, et al., 2003). Konsep klasifikasi dengan SVM adalah mencari *hyperplane* (garis) terbaik yang berfungsi sebagai pemisah dua kelas data. SVM memaksimalkan *margin*, yang merupakan jarak pemisah antara kelas data. SVM juga mampu bekerja pada dataset yang berdimensi tinggi dengan menggunakan *kernel trick*. Ada beberapa macam fungsi *kernel* SVM, yaitu : Linear, *Polynomial*, Gaussian RBF, Sigmoid, Invers Multi Kuadratik, dan *Additive*.

Pada penelitian ini fungsi kernel yang digunakan adalah SVM *Polynomial*. SVM linear digunakan ketika data yang akan diklasifikasi dapat terpisah dengan sebuah *hyperplane*, sedangkan SVM non-linear digunakan ketika data hanya dapat dipisahkan dengan garis lengkung. SVM *Polynomial* memiliki definisi fungsi dengan Persamaan (3).

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d \quad (3)$$

Dimana $K(\vec{x_i}, \vec{x_j})$ merupakan fungsi *kernel*, x merupakan fitur dan d merupakan ordo.

Hyperplane dalam SVM yang optimal didapatkan dengan merumuskannya ke dalam QP problem dan diselesaikan menggunakan library yang banyak tersedia dalam analisa numeric. Namun ada alternatif lain yang cukup sederhana yaitu metode sequential. Metode ini dikembangkan oleh Vijayakumar untuk mencari nilai α, yang diuraikan dalam tahapan:

1. Inisialisasi $\alpha_i = 0$

   Menghitung nilai matriks *Hessian* dengan menggunakan persamaan (4).

   $$D_{ij} = y_i y_j (K(x_i . x_j) + \lambda^2) \quad (4)$$

   Dimana y merupakan kelas dari data ke-i dan ke-j, $K(x_i . x_j)$ merupakan fungsi kernel *polynomial* yang digunakan.

2. Menghitung setiap level dengan tahapan menggunakan persamaan (5) sampai (7).

a) $E_i = \sum_{j=1}^{n} \alpha_i D_{ij}$ (5)

b) $\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\}$ (6)

c) $\alpha_i = \alpha_i + \delta\alpha_i$ (7)

3. Melakukan perulangan ke tahap 2 sampai nilai α mencapai konvergen.

Dimana ϒ merupakan parameter untuk mengontrol kecepatan proses learning. Konvergensi dapat didefinisikan dari perubahan nilai α (Nugroho, et al., 2003).

2.5 Feature Selection

Feature Selection atau dalam Bahasa Indonesia seleksi fitur merupakan salah satu proses pemilihan fitur yang relevan dalam hal masalah pembelajaran target. Tujuan dari seleksi fitur adalah untuk menghapus fitur yang berlebihan dan yang tidak relevan (Bangsheng, 2013).

Information Gain (IG) merupakan salah satu metode untuk seleksi fitur yang banyak digunakan oleh peneliti untuk menentukan batas dari kepentingan sebuah atribut (Deng & Runger, 2012). Nilai IG diperoleh dari nilai *entropy* sebelum pemisahan dikurang dengan nilai *entropy* setelah pemisahan. Nilai ini digunakan untuk penentuan atribut mana yang akan dibuang atau digunakan. Atribut yang memenuhi kriteria pembobotan nantinya akan digunakan untuk proses klasifikasi

Dalam pemilihan fitur dengan IG dilakukan dengan 3 tahapan, yaitu:

1. Menghitung nilai IG untuk setiap atribut.
2. Menentukan *threshold* (batas). Hal ini digunakan untuk menentukan atribut yang bobotnya lebih kecil dari *treshold* akan dibuang.
3. Memperbaiki dataset dengan pengurangan atribut.

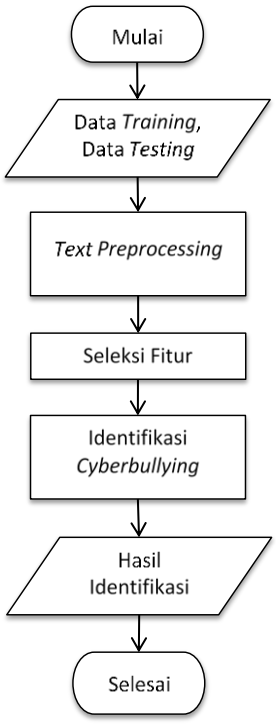
Seleksi fitur *Information Gain* $IG(t)$ dirumuskan pada persamaan (8).

$$IG(t) = -\sum_{i=1}^{|C|} P(C_i) \log P(C_i) + P(t) \sum_{i=1}^{|C|} P(C_i|t) \log P(C_i|t) + P(\bar{t}) \sum_{i=1}^{|C|} P(C_i|\bar{t}) \log P(C_i|\bar{t}) \quad (8)$$

Dimana C_i merupakan kelas data, $P(C_i)$ merupakan peluang dari kelas data, $P(t)$ dan $P(\bar{t})$ merupakan peluang *term* t yang muncul atau tidak muncul dalam dokumen. Dalam *machine learning*, perolehan informasi dapat digunakan untuk membantu menentukan peringkat fitur (Bangsheng, 2013).

3. PERANCANGAN

Perancangan pada sistem ini dimulai dari melakukan pemrosesan teks kemudian hasil *term weighting* digunakan untuk seleksi fitur IG setelah itu melakukan perhitungan dengan metode SVM yang keluarannya adalah hasil identifikasi *tweet* yang ditunjukan Gambar 1.



Gambar 1 Diagram Alir Sistem

Sistem dimulai dengan memasukan data *training* dan data *testing*, kemudian melakukan proses *text preprocessing*. *Text preprocessing* dimulai dari melakukan proses tokenisasi, *filtering*, *stemming* dan kemudian *term weighting*. Dalam *term weighting* menghitung nilai *term frequency*, *document frequency*, *inverse document frequency*, *weight term frequency* dan *weight term document.* Nilai *weight term document* digunakan untuk masukan dari proses SVM. Sebelum melakukan perhitungan dengan metode SVM, dilakukan seleksi fitur terlebih dahulu dengan menghitung *term presence*, kemudian hitung nilai *entropy*, setelah itu mengurutkan nilai secara *ascending*. *Threshold* digunakan untuk menggunakan fitur sesuai nilai *threshold* yang ditentukan. Setelah fitur telah diseleksi, kemudian mengambil nilai *weight term document* dari fitur tersebut untuk dilakukan proses klasifikasi menggunakan SVM. Setelah proses klasifikasi SVM didapatkan hasil identifikasi *tweet* yang mengandung konten *cyberbullying*.

4. HASIL DAN PEMBAHASAN

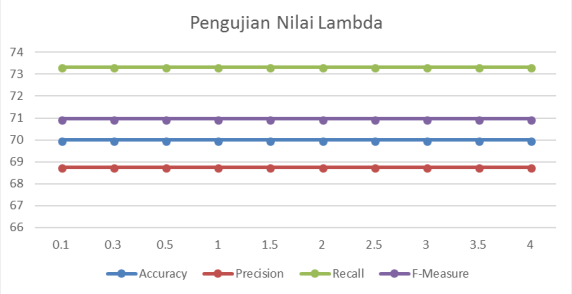
Pada pengujian ini menggunakan *Accuracy*, *Precision*, *Recall* dan *F-measure*. Hal ini dikarenakan pada penelitian ini hanya mengidentifikasi *tweet* yang mengandung *cyberbullying*, sehingga *retrieve* hasil identifikasi dibutuhkan untuk mengetahui apakah informasi yang diminta oleh pengguna sudah sesuai dengan informasi yang diberikan oleh sistem.

Jumlah data yang digunakan adalah sebanyak 300 *tweet*, dimana 150 *tweet bully* dan 150 *tweet* bukan *bully*. Data divalidasi oleh Ibu Liana Shinta Dewi, M.A yang merupakan Dosen Bahasa Indonesia Universitas Brawijaya. Pada pengujian skenario perbandingan yang digunakan adalah 240 data *training* dan 60 data *testing*.

4.1 Pengujian Parameter Sequential Training SVM

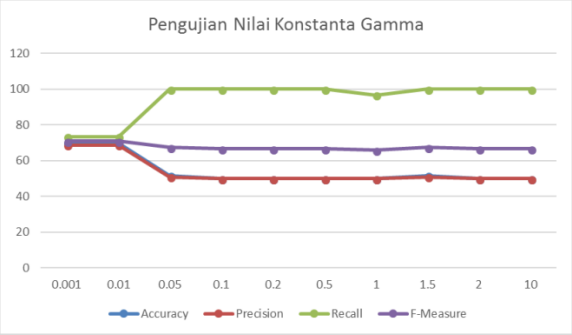
Ada 6 parameter yang diuji pada *sequential training* SVM dengan 10 nilai percobaan yang berbeda yaitu variabel *lamda*, konstanta *gamma*, *epsilon*, maksimum iterasi dan nilai *complexity* (C).

Nilai parameter *sequential training* SVM yang digunakan dalam pengujian adalah λ = 0,5, Υ = 0,001, ε = 0,0001, C = 1, dan iterMax = 100.



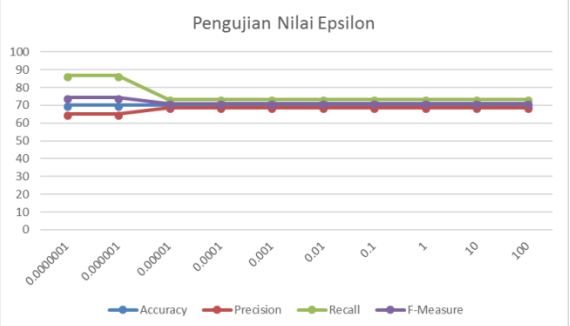
Gambar 2 Grafik Hasil Pengujian Lambda

Pada grafik Gambar 2 didapatkan bahwa hasil yang didapatkan adalah konstan pada semua nilai lambda yang diujikan yaitu *accuracy* 70%, *precision* 68,75%, *recall* 73,33% dan *f-measure* 70,96%. Hal ini terjadi karena nilai *lambda* hanya digunakan untuk melakukan perhitungan matriks *hessian*. Dan matriks *hessian* digunakan ketika menghitung nilai *Ei* untuk *sequential training* SVM. Sehingga hasil yang didapatkan tidak terlalu mempengaruhi nilai *bias*.



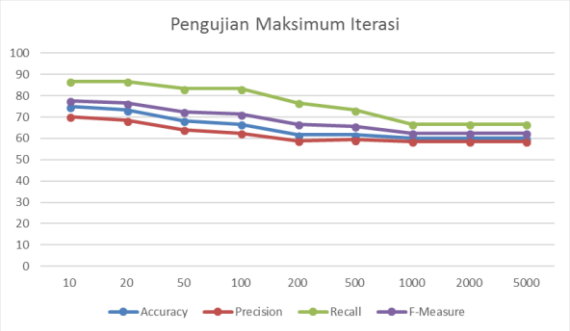
Gambar 3 Grafik Hasil Pengujian Konstanta Gamma

Pada grafik Gambar 3 didapatkan bahwa hasil terbaik didapatkan pada nilai *gamma* 0,001 dan 0,01, kemudian nilai menurun ketika nilai *gamma* semakin tinggi. Hal in disebabkan karena nilai *gamma* digunakan untuk menghitung nilai *delta alpha*, dimana nilai *delta* alpha merupakan nilai yang menentukan apakah hasil konvergen atau belum. Nilai *gamma* yang diambil adalah 0,001 dengan *accuracy* 70%, *precision* 68,75%, *recall* 73,33% dan *f-measure* 70,96%.



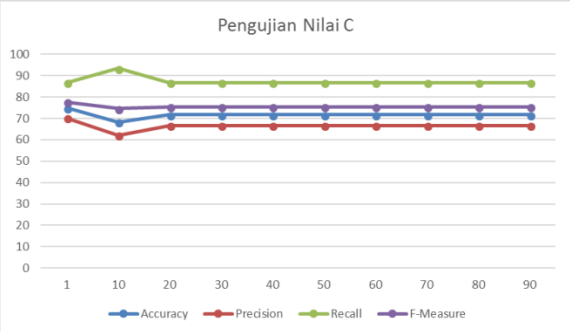
Gambar 4 Grafik Hasil Pengujian Epsilon

Pada grafik Gambar 4 didapatkan bahwa hasil terbaik pada nilai *epsilon* kurang dari sama dengan 0,000001 dan 0,0000001. Hal ini disebabkan karena nilai *epsilon* digunakan sebagai batas maksimal untuk hasil konvergen. Ketika nilai *epsilon* semakin tinggi, maka hasil akan semakin cepat konvergen. Pemilihan nilai epsilon 0,000001 dengan *accuracy* 68,33%, *precision* 64,10%, *recall* 83,33% dan *f-measur*e 72,46%.



Gambar 5 Grafik Hasil Pengujian Maksimum Iterasi

Pada grafik Gambar 5 bahwa hasil terbaik yang didapatkan terdapat pada iterasi sebanyak 20 dengan nilai *accuracy* 75%, *precision* 70,27%, *recall* 86,66% dan *f-measur*e 77,61%. Kemudian pada iterasi lebih dari 1000 hasil yang dihasilkan sama, hal ini dikarenakan perhitungan telah memasuki keadaan konvergen pada iterasi 1404.

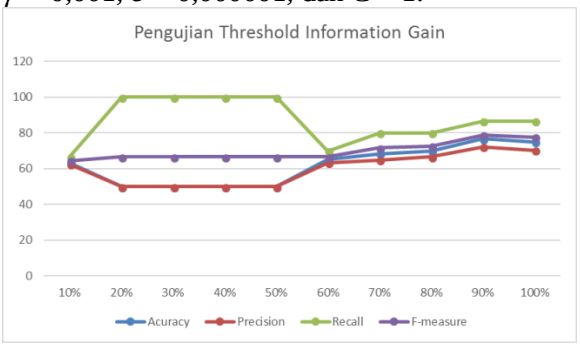


Gambar 6 Grafik Hasil Pengujian Nilai C

Grafik Gambar 6 didapatkan bahwa hasil terbaik pada pengujian nilai C sama dengan 1 dengan nilai *accuracy* 75%, *precision* 70,27%, *recall* 86,66% dan *f-measur*e 77,61%. Hal ini disebabkan karena semakin tinggi nilai C maka nilai kernel polynomial akan semakin tinggi pula. Ketika nilai *kernel polynomial* semakin tinggi maka nilai matriks *Hessian* semakin tinggi pula, sehingga menyebabkan pada perhitungan nilai konstanta *gamma* yang hasilnya didapatkan semakin kecil kemudian dapat menyebabkan iterasi menjadi semakin cepat konvergen.

4.2 Pengujian Threshold Feature Selection IG

Setelah melakukan pengujian *sequential training* SVM maka didapatkan hasil untuk melakukan pengujian selanjutnya yaitu *threshold* IG dengan nilai *iterMax* = 20, λ = 0,5, γ = 0,001, ε = 0,000001, dan C = 1.



Gambar 7 Grafik Hasil Pengujian Threshold IG

Pada grafik Gambar 7 hasil terbaik didapatkan pada nilai *threshold* 90%. Ketika semua fitur digunakan hasil yang didapatkan *accuracy* sebesar 76,66%, *precision* 72,22%, *recall* 86,66% dan *f-measure* 78,78%. Pada saat pemilihan *term* menggunakan seleksi fitur *information gain* proses menghasilkan nilai yang tinggi pada *term* yang unik dimana *term* tersebut muncul sekali pada satu kelas. Nilai *information gain* akan mempengaruhi *term* yang digunakan pada proses identifikasi, dimana *term* dengan nilai *information gain* yang tinggi yang akan digunakan

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, pengujian dan anilisis yang telah dilakukan dapat diambil kesimpulan bahwa hasil pengujian identifikasi *tweet cyberbullying* menggunakan metode SVM mendapatkan hasil terbaik berdasarkan pengujian *iterMax*, parameter λ (*lambda*), γ (konstanta *gamma*), ε (*epsilon*), dan C (*complexity*), pada *sequential training* SVM berpengaruh pada perubahan nilai bobot α_i (*alpha* ke-*i*) dan nilai *b* (*bias*). Hasil terbaik yang didapatkan dari seluruh pengujian parameter *sequential training* SVM yaitu *iterMax* = 20, λ = 0,5, γ = 0,001, ε = 0,000001, dan C = 1. Hasil akurasi yang diperoleh yaitu *accuracy* 75%, *precision* 70.27%, *recall* 86,66% dan *f-measur*e 77,61%. Dan pada hasil pengujian *threshold* seleksi fitur *information gain*, hasil terbaik didapatkan adalah pada nilai *threshold* 90%, dengan nilai *accuracy* 76,66%, *precision* 72,22%, *recall* 86,66% dan *f-measure* 78,78%. Hal ini disebabkan karena seleksi fitur

information gain mempunyai nilai yang tinggi jika fitur tersebut merepresentasikan suatu kelas tertentu, dan mempunyai nilai yang rendah jika fitur tersebut muncul didalam semua kelas. Jadi hasil identifikasi *tweet cyberbullying* dengan seleksi fitur mendapatkan akurasi lebih tinggi dibandingkan menggunakan seluruh fitur yang ada.

Sedangkan saran yang dapat diberikan berdasarkan penelitian yang telah dilakukan, yaitu diharapkan menambah jumlah data yang digunakan, karena akan berpengaruh pada proses identifkasi untuk mendapatkan hasil yang optimal. Serta pada penelitian selanjutnya perlu mempertimbangkan *dimensionality* dari data *training*. Karena jika memiliki data *training* yang *dimensionality* tinggi kemungkinan dapat menyebabkan keadaan *overfitting*.

6. DAFTAR PUSTAKA

Adriansyah, M. et al., 2017. Cyberbullying Comment Classification on Indonesian Selebgram Using Support Vector Machine Method. *Research Gate.*

Bangsheng, S., 2013. *Information Gain Feature Selection Based on Feature Interactions.* Houston: s.n.

Chandani, V., Wahono, R. S. & Purwanto, 2015. Komparasi ALgoritma Klasifikasi Machine Learning dan Feature Selection pada Analisis Sentimen Review Film. *Journal of Intellignet Systems,* Volume 1.

Devaga, Evita. 2017. RI Perangi Konten Negatif. Tersedia di: <https://www.kominfo.go.id/content/detail/11226/ri-perangi-konten-negatif/0/sorotan_media> [Diakses 26 Januari 2018].

Deng, H. & Runger, G., 2012. Feature Selection via Regularized Trees. *The 2012 International Joint Conference on Neural Networks (IJCNN),* Volume 3.

Gaigole, P. C., Patil, L. H. & Chaudhari, P. M., 2013. *Preprocessing Techniques in Text Categorization.* s.l., International Journal of Computer Application.

Garcia, E. 2005. *Document Indexing Tutorial*. Tersedia di: <http://www.miislita.com/information-retrieval-tutorial/indexing.html> [Diakses 29 September 2017].

Motivasee, 2017. Apa itu Twitter dan Tips Cara Menggunakannya. Motivasee. Tersedia di: <http://motivasee.com/twitter/> [Diakses 29 September 2017].

Noviantho, Isa, S. M. & Ashianti, L., 2017. Cyberbullying Classification using Text Mining. *IEEE.*

Nugroho, A. S., Witarto, A. B. & Handoko, D., 2003. Teori dan Aplikasinya dalam Bioinformatika. *IEEE.*

Oktaviani, Kirana. 2013. Apa itu *Cyber Bullying*. Kompasiana. Tersedia di: <https://www.kompasiana.com/kiranaoktaviani/apa-itu-cyber-bullying_552ff83a6ea8344b778b45d4> [Diakses 29 September 2017].