Penyaring Komentar Cyberbullying Pada Konten Blog

Danar Dono [1], Eka Rahayu Setyaningsih [1], dan C. Pickerling [1]
[1]Departemen Teknologi Informasi, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia

Corresponding author: Eka Rahayu Setyaningsih (e-mail: eka@stts.edu).

ABSTRACT Cyberbullying is a real threat to the interaction between blog content writers and blog readers. This research discusses the development of a cyberbullying filter feature on blog content to minimize cyberbullying on a blog site. The system development method uses an iterative waterfall including system analysis, system design, implementation, and testing. Based on testing with data training mode using 7755 comment datasets with a proportion of 3984 cyberbullying and 3771 non-cyberbullying results in an accuracy of 85.25% and an error of 14.75%. Testing with data testing mode using 1936 dataset comments with the proportion of 583 cyberbullying and 1353 non-cyberbullying resulted in 80% accuracy and 20% error. Based on the test, it can be concluded that the development of the cyberbullying comment filter feature using the Naive Bayes classifier produces an average accuracy of 80% and an average error of 20%.

KEYWORDS *Cyberbullying*, Comment Filtering, Classifier

ABSTRAK *Cyberbullying* merupakan ancaman nyata dalam interaksi di antara penulis konten blog dan pembaca blog. Penelitian ini membahas tentang pengembangan fitur penyaring *cyberbullying* pada konten blog untuk meminimalisir *cyberbullying* dalam suatu situs blog. Adapun metode pengembangan sistem menggunakan *iterative waterfall* meliputi analisis sistem, desain sistem, implementasi dan pengujian. Berdasarkan pengujian dengan mode pelatihan data menggunakan 7755 dataset komentar dengan proporsi 3984 *cyberbullying* 3771 non-*cyberbullying* menghasilkan akurasi 85,25% dan error 14,75%. Pengujian dengan mode testing data menggunakan 1936 dataset komentar dengan proporsi 583 *cyberbullying* dan 1353 non-*cyberbullying* menghasilkan akurasi 80% dan error 20%. Dari hasil pengujian disimpulkan bahwa pengembangan fitur penyaring komentar *cyberbullying* dengan menggunakan naive bayes classifier menghasilkan rata-rata akurasi sebesar 80% dan rata-rata error sebesar 20%.

KATA KUNCI *Cyberbullying*, Penyaringan Komentar, Classifier

I. PENDAHULUAN

Dewasa ini ada banyak sistem manajemen konten (CMS) yang digunakan untuk mengelola konten suatu situs web. CMS biasanya dilengkapi dengan fitur komentar bagi penggunanya, memungkinkan terjadinya interaksi antara pembuat konten dan pembaca artikel [1]. Komentar dapat berisi cyberbullying dalam konten blog. Cyberbullying adalah perlakuan kejam yang disengaja kepada orang lain dengan mengirimkan atau mengedarkan materi berbahaya atau terlibat dalam bentuk agresi sosial menggunakan internet atau teknologi digital lainnya [2]. Cyberbullying berdampak negatif pada korban trauma psikologis, emosional dan sosial. Fitur filter komentar cyberbullying diperlukan untuk menyaring berbagai komentar berpotensi cyberbullying dalam konten blog. Dalam penelitian ini akan membahas bagaimana mengembangkan sistem filter komentar cyberbullying pada konten blog yang nantinya dapat digunakan untuk meminimalisir penyalahgunaan komentar, khususnya terkait bullying oleh pengguna melalui fitur komentar pada konten blog. Cyberbullying merupakan ancaman nyata dalam interaksi di antara penulis konten blog dan pembaca artikel blog. Cyberbullying berdampak negatif terhadap korbannya seperti gangguan mental, tekanan emosional, hingga trauma sosial. Penelitian ini membahas bagaimana pengembangan fitur penyaring komentar cyberbullying pada konten blog untuk meminimalisir cyberbullying dalam suatu situs blog.

II. TINJAUAN PUSTAKA

Pada bagian ini dijelaskan tentang teori penunjang yang digunakan dalam pengembangan sistem ini meliputi:

A. CYBERBULLYING

Menurut Williard (2005), *cyberbullying* adalah perlakuan kejam yang dilakukan dengan sengaja kepada orang lain dengan mengirimkan atau mengedarkan materi berbahaya atau terlibat dalam bentuk agresi sosial menggunakan internet atau teknologi digital lainnya. Adapun aspek-aspek *cyberbullying* meliputi [2]:

1. *Flamming* yakni perilaku pengiriman pesan teks dengan kata-kata kasar dan frontal.
2. *Harassment* yakni perilaku pengiriman pesan tidak sopan kepada seseorang berupa gangguan yang dikirimkan melalui email, sms, atau pesan singkat di jaringan media sosial secara terus menerus.
3. *Denigration* yakni perilaku mengubar keburukan seorang di internet dengan maksud merusak reputasi dan nama baik dari orang yang dituju..
4. *Impersonation* yakni perilaku berpura-pura menjadi orang lain dan mengirimkan pesan yang tidak baik.
5. *Outing and Trickery* yakni perilaku menyebarkan rahasia orang lain atau foto pribadi orang lain.
6. *Exclusion* yakni perilaku yang dengan sengaja dan kejam menghilangkan orang dari grup online
7. *Cyberstalking* yakni perilaku berulang kali mengirimkan ancaman berbahaya atau pesan yang mengintimidasi menggunakan komunikasi elektronik.

Dalam penelitian ini lebih difokuskan pada aspek-aspek *cyberbullying* meliputi *flamming, harassment, denigration, impersonation, dan cyberstalking*.

B. NAÏVE BAYES CLASSIFIER

Naive Bayes Classifier (NBC) merupakan metode pembelajaran dengan konsep probabilitas sederhana [3]-[5]. NBC menggunakan teorema kuno, warisan abad ke-18, yang ditemukan oleh Thomas Bayes. NBC menyertakan dokumen klasifikasi terbimbing, metode pembelajaran yang menghasilkan fungsi untuk memetakan masukan ke keluaran yang diinginkan. NBC menganggap kemunculan satu kata tidak mempengaruhi kemunculan kata lainnya. NBC mampu memberikan kinerja yang cukup baik untuk banyak kasus modern dengan data yang besar. Adapun untuk menghitung probabilitas fitur kata menggunakan persamaan (1):

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|} \quad (1)$$

Kemudian untuk menghitung probabilitas prior menggunakan persamaan (2):

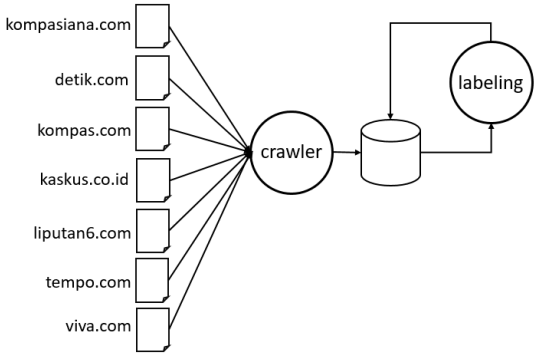
$$\hat{P}(c) = \frac{N_c}{N} \quad (2)$$

Terakhir untuk menentukan sentimen menggunakan persamaan (3):

$$c = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(w_i \mid c) \quad (3)$$

C. DATASET

Dataset yang digunakan dalam penelitian ini adalah pasangan data berita dan komentar-komentarnya yang diperoleh melalui proses crawling sejumlah website berita di Indonesia, yaitu kompasiana, detik, dan kompas, viva, tempo, liputan6, serta sebuah website forum, yaitu kaskus, seperti yang ditunjukkan pada gambar 1 berikut ini.



GAMBAR 1. Pembentukan Dataset

Kepada semua laman website yang menjadi sumber data, dilakukan pembatasan topik yang diambil. Adapun topik yang diolah hanyalah topik sosial budaya, politik, dan olahraga. Ketiga topik tersebut dipilih karena biasanya merupakan diskusi sensitif dan rawan *cyberbullying* di kalangan pembacanya. Total artikel yang diperoleh untuk dataset ini adalah 686 artikel dari berbagai sumber. Sedangkan untuk komentar terdiri dari 9803 komentar dari 686 artikel. Setiap komentar tersebut selanjutnya akan diberi label secara manual. Proses pelabelan dilakukan untuk membedakan komentar menjadi dua sentimen yaitu *cyberbullying* dan non-*cyberbullying*. Pada akhir proses pembentukan dataset, diperoleh 4.598 komentar yang termasuk dalam kategori *cyberbullying* dan 5205 komentar yang termasuk dalam kategori non-*cyberbullying*. Pasangan data artikel dan komentar-komentarnya yang diperoleh dari proses crawling tersebut selanjutnya akan disimpan ke dalam database [6] untuk kemudian dilanjutkan dengan preprocessing. Tahap preprocessing itu sendiri akan dijelaskan pada subbab yang terpisah.

D. PREPROCESSING

Preprocessing merupakan memproses data uji sebelum digunakan dalam program bertujuan untuk mengurangi jumlah kosa kata, menyeragamkan kata dan menghilangkan *noise* [7]. Setiap tahapan yang dilakukan dalam *preprocessing* adalah sebagai berikut:

1. *Case folding* adalah proses mengubah huruf dalam dokumen menjadi huruf kecil. Dokumen mengandung beragam variasi bentuk huruf sampai tanda baca. Variasi huruf ini harus diseragamkan dan tanda bacanya harus dihilangkan. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter lainnya dihilangkan dan dianggap sebagai delimiter.
2. *Cleaning* yakni proses menghilangkan dokumen dari mention, hastag, link, emoticon, dan karakter lainnya yang tak berguna.
3. *Stopword* yakni proses menghapus kata-kata yang tidak perlu dari dokumen. Pada tahap *stopword-list*, kata-kata yang tidak penting dibuang dari daftar kata, misalnya kata "yang", "dimana", "mengapa", "yaitu", "yakni", dan sebagainya.
4. Normalisasi bahasa gaul adalah proses mengubah kata-kata tak lazim berupa kata-kata gaul menjadi kata-kata formal berbahasa Indonesia. Umumnya, tidak semua komentar pada suatu artikel menggunakan bahasa formal. Pembaca biasanya juga memakai bahasa gaul, seperti nggak, gue, loe, dll. Kata-kata yang tak lazim ini perlu diseragamkan melalui normalisasi bahasa menjadi bahasa formal berbahasa Indonesia.
5. *Stemming* adalah proses mengubah kata menjadi kata dasar. Pada umumnya kata dalam dokumen memiliki variasi kombinasi imbuhan kata yang beragam, seperti imbuhan awalan, akhiran, sisipan, dan kombinasi. Kata-kata tersebut perlu diseragamkan menjadi kata dasar supaya seragam dan mengurangi komplesitas kata. Adapun algoritma stemming yang terkenal ialah algoritma Nazief dan Adriani. Algoritma ini dikembangkan berdasarkan aturan morfologi Bahasa Indonesia yang mengelompokkan imbuhan menjadi awalan (prefiks), sisipan (infiks), dan akhiran (suffiks) dan gabungan awalan-akhiran (confiks). Algoritma ini menggunakan kamus kata dasar dan mendukung *recording*, yakni penyusunan kembali kata-kata yang mengalami proses *stemming* berlebih.
6. *Tokenizing* adalah proses pemotongan dokumen menjadi kata-kata setelah menjadi proses *filtering*. Hasil pemotongan kata-kata tersebut dijadikan kumpulan kata dan membentuk daftar kata. Potongan tersebut dikenal dengan istilah token

III. METODE

Perancangan sistem ini menggunakan metodologi *iterative waterfall* yang terdiri atas empat tahapan meliputi:

1. Analisis sistem yakni tahapan analisis terhadap kebutuhan sistem. Untuk itu, diperlukan sejumlah literatur terkait berupa jurnal dan buku-buku yang relavan terkait pengembangan sistem manajemen konten dan analisis sentimen *cyberbullying* guna mendapat informasi terkait kebutuhan, fitur dan batasan dalam pengembangan sistem.
2. Desain sistem yakni tahapan desain sistem berupa perancangan arsitektur, database, interface, dan desain prosedural sesuai dengan kebutuhan dan fitur-fitur yang akan dikembangkan.
3. Implementasi yakni tahapan implementasi sistem dalam terhadap modul-modul yang dikembangkan dengan cara pemograman. Pengembangan CMS dalam penelitian ini menggunakan bahasa PHP menggunakan *framework* CodeIgniter.
4. Pengujian yakni tahapan pengujian terhadap sistem baik berupa pengujian fungsional maupun non fungsional, guna mengetahui bahwa sistem yang dikembangkan dapat berjalan secara baik. Adapun pengujian fungsional sistem menggunakan *black box testing.* Sedangkan untuk mengetahui tingkat keakuratan sistem dalam mengklasifikasikan sentimen *cyberbullying* digunakan uji *hold-out* dimana dataset dibagi menjadi dataset latih dan dataset testing. Adapun model akan dievaluasi menggunakan parameter *accuracy and error rate*

IV. HASIL EKSPERIMEN DAN PENELITIAN

Pada bagian ini akan dijelaskan tentang hasil eksperimen dan penelitian. Adapun penjelasannya meliputi:

A. FITUR-FITUR SISTEM

Sistem CMS dalam penelitian ini memiliki beberapa fitur yang terbagi atas fitur front end dan fitur back end. Sistem ini memiliki 3 level hak akses meliputi pembaca, member, dan admin dimana masing-masing user memiliki hak akses berbeda. Selengkapnya ditunjukan pada tabel I

TABEL I
FITUR FRONT END

	Fitur	Pembaca	Member	Admin
Front End				
1.	Registrasi	V	X	X
2.	Login	X	V	V
3.	Artikel	V	V	V
4.	Komentar artikel	X	V	V
5.	Rating artikel	X	V	V
6.	Statistik artikel	X	V	V
7.	Kategori artikel	V	V	V
8.	Following	X	V	V
9.	Peringatan *cyberbullying*	X	V	V
10.	Pelaporan Komentar *cyberbullying*	X	V	V
11.	Tag Terpopuler	V	V	V
12.	Artikel Terpopuler	V	V	V
13.	Berbagi ke sosial media	V	V	V
14.	Tentang web	V	V	V
15.	Kebijakan web	V	V	V
16.	Evaluasi web	X	V	V
17.	Profil	V	V	V

. Selanjutnya CMS juga memiliki fitur *backend*. Pada bagian *backend* terdapat 2 level hak akses yakni member dan admin. Pada bagian *backend*, pembaca tidak bisa

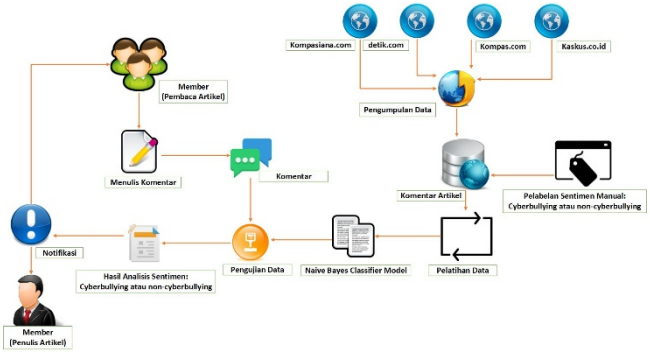
mengaksesnya. Selengkapnya ditunjukan pada tabel II di bawah ini:.

TABEL II
FITUR BACK END

	Fitur	Pembaca	Member	Admin
Back End				
1.	Konfigurasi profil	X	V	V
2.	Manajemen user	X	X	V
3.	Notifikasi	X	V	V
4.	Mengelola artikel	X	V	V
5.	Mengelola follower-following	X	V	V
6.	Mengelola rating	X	V	V
7.	Mengelola komentar	X	V	V
8.	Manajemen penyaring komentar *cyberbullying*	X	V	V
9.	Mengelola kategori	X	X	V
10.	Manajemen galeri	X	V	V
11.	Manajemen tag	X	X	V
12.	Manajemen tentang web	X	X	V
13.	Manajemen kebijakan	X	X	V
14.	Manajamen evaluasi web	X	X	V

B. ARSITEKTUR APLIKASI PENYARING KOMENTAR CYBERBULLYING

Pada gambar 2, ditunjukkan arsitektur penyaring komentar *cyberbullying* beserta cara kerjanya. Pertama, dataset komentar diambil secara manual dari berbagai sumber data seperti detik.com, kompas.com, dll. Dataset komentar selanjutnya disimpan ke database. Setelah setiap dataset yang ada dilabeli secara manual, dataset komentar dibagi menjadi dataset training dan dataset test.

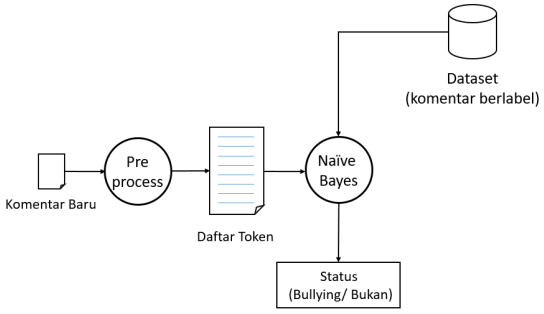


GAMBAR 2. **Arsitektur Aplikasi Penyaring Komentar Cyberbullying**

Dataset pelatihan dilatih untuk menghasilkan *Naive Bayes Classifier Model* [8]. Model tersebut digunakan untuk menyaring komentar dari pembaca yang mengirimkan komentar terkait konten blog. Kemudian, model akan mendeteksi komentar yang mengandung *cyberbullying*, kemudian sistem mengirimkan notifikasi kepada pembuat artikel dan pembaca artikel tentang peringatan *cyberbullying*[9][10].

C. ARSITEKTUR PENDETEKSI CYBERBULLYING

Seperti yang ditunjukkan pada gambar 3 mengenai arsitektur pendeteksi *cyberbullying,* pada bagian ini akan dijelaskan arsitektur beserta cara kerjanya pendeteksian *cyberbullying*. Pertama, kepada setiap komentar baru akan dilakukan preprocessing yang sama seperti pada tahap pembentukan dataset. Setelah diperoleh setiap token yang merupakan fitur dari komentar tersebut, akan dilakukan perhitungan probabilitas prior dan posterior dengan menggunakan Naïve Bayes Classifier. Dengan menggunakan model Naïve Bayes Classifier yang telah ditraining sebelumnya dengan menggunakan dataset yang ada, maka komentar yang baru akan diuji sentimennya oleh sistem dengan menghitung Vmap dan menentukan sentimen komentar tersebut [6][7]. Dalam penghitungan Vmap, semua kata komentar akan diberi bobot untuk menghasilkan sentimen *cyberbullying* dan non-*cyberbullying* sentimen. Jika bobot sentimen *cyberbullying* lebih besar dari bobot sentimen non *cyberbullying*, maka dapat disimpulkan bahwa komentar tersebut termasuk dalam *cyberbullying*. Jika bobot sentimen *cyberbullying* tidak lebih besar dari bobot sentimen non *cyberbullying* maka dapat disimpulkan bahwa komentar tersebut adalah non *cyberbullying*. Terakhir, hasilnya adalah sentimen *cyberbullying* atau sentimen non-*cyberbullying* dari komentar anggota.



GAMBAR 3. **Arsitektur Penyaring Komentar Cyberbullying**

D. PENGUJIAN

Pengujian terhadap penyaring komentar *cyberbullying* dilakukan dengan dua mode yakni pengujian menggunakan data pelatihan dan pengujian menggunakan data testing. Berdasarkan pengujian model dengan mode data pelatihan menggunakan total 7755 dataset komentar, maka dapat disimpulkan bahwa model memiliki akurasi sebesar 85,25% dan kesalahan 14,75%. Selengkapnya hasil pengujian menggunakan data pelatihan dapat dilihat pada gambar 4.

GAMBAR 4. Pengujian Model Menggunakan Data Pelatihan

Sedangkan berdasarkan pengujian model dengan mode data testing menggunakan total 1936 komentar dataset, didapatkan hasil model memiliki akurasi sebesar 80,48% dan kesalahan 19,52%

V. KESIMPULAN

Berdasarkan pembahasan di atas, dapat disimpulkan:

1. Dengan adanya aplikasi penyaringan komentar pada aplikasi web, dapat membantu menghindarkan masyarakat dari potensi *cyberbullying*.
2. Pengembangan fitur penyaring komentar *cyberbullying* menggunakan naive bayes classifier menghasilkan rata-rata akurasi sebesar 80% dan rata-rata error sebesar 20%.
3. Penggunaan algoritma Naïve Bayes untuk proses klasifikasi tergolong sederhana dan menuntut jumlah dataset yang besar, serta pembagian jumlah dataset untuk setiap class yang seimbang untuk dapat menghasilkan tingkat akurasi yang cukup baik.

PERAN PENULIS

Setiap penulis memiliki kontribusi yang sama dalam Analisis Formal, Investigasi, Administrasi Proyek, Sumber Daya, Perangkat Lunak, Validasi, Visualisasi, Penulisan dan Penyusunan Draf Asli.

COPYRIGHT

DAFTAR PUSTAKA

[1] Subagia, Anton. 2018. *Kolaborasi Codeigneter dan Ajax dalam Perancangan CMS*. Jakarta: PT Elex Media Komputindo.

[2] S. Hinduja & J. Patchin. 2010. *Bullying, Cyberbullying, adn Suicide*. Archives of Suicides Research. Vol. 14.

[3] Sipayung, M. Evasaria, dkk. 2016. *Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classfier*. Jurnal Sistem Informasi (JSI). Vol. 8. No. 1. April 2016.

[4] Rahayu, Dwi Yeni Made Ni. 2018. *Rancangan Penerapan Metode Naive Bayes dalam Mendeteksi Hate Speech di Media Sosial*. Prosiding Seminar Nasional Pendidikan Teknik Informatika (SENTAPATI). Vol. 9. 8 September 2018.

[5] Kim Schouten , Onne van der Weijde, FlaviusFrasincar ,Rommert Dekker. 2018. *Supervised and Unsupervised Aspect Category detection for sentiment analysis with co-occurrence data.* IEEE Transactions on Cybernetics. Vol. 48, No. 4.

[6] Indrajani. 2014. *Pengantar Sistem Basis Data Case Study All In One*. Jakarta : PT. Elex Media Komputindo.

[7] Suyanto. 2018. *Machine Learning: Tingkat Dasar dan Lanjut.* Bandung: Informatika Bandung.

[8] Prabha, Surya, B Subbulakshmi. 2019. *Sentimental Analysis using Naive Bayes Classifier.* International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN).

[9] VandanaJha, Savitha.R ,P.DeepaShenoy, ArunKumarSangaiah, Venugopal KR. 2018. *A novel sentiment aware dictionary for multi domain sentiment classification*. Journal of Computers and Electrical Engineering. Vol. 69. Halm. 585-597.

[10] ShufengXiong , KuiyiWang, Donghong , Ji BingkunWang. 2018. *A short text sentiment topic model for product reviews*. Journal of Neurocomputing. Vol. 297. Halm. 94-102.