Deteksi *Cyberbullying* pada Cuitan Media Sosial *Twitter*

Nassharih Abdulloh
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM. 14.5, Sleman, Yogyakarta, Indonesia
16523112@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM. 14.5, Sleman, Yogyakarta, Indonesia
Fathan@uii.ac.id

***Abstract*— Cyberbullying becomes a big problem that everyone must pay more attention to this bad habit. Cyberbullying has a dangerous effect such as psychological disorder to suicide. The purpose of this study to identify cyberbullying content on the social media. In this case, writers use tweets from Twitter as a study object. At least there are 1971 rows data collected, that contain both cyberbullying and non-cyberbullying tweets. To perform the identification/classification process, writers do five steps to get the result of the study. Those are data collection, preprocessing, feature extraction, classification, and evaluation. At least four Machine Learning algorithm applied in this study, those are K-Nearest Neighbor (KNN), Support Vector Machine with linear kernel (SVM), Logistic Regression, and Multinomial Naïve Bayes. The result of this study, those algorithm doesn't has a significant different of the performance although SVM reach the highest value. The accuracy of Multinomial Naïve Bayes, Logistic Regression, SVM, and KNN respectively 0.9619; 0.9949; 0.9975; 0.9188.**

Keywords—Twitter, Cyberbullying, Machine Learning, Classification

Abstrak— *Cyberbullying* menjadi sebuah masalah yang harus mendapat perhatian serius oleh semua pihak. Di samping tindakan ini merupakan kebiasaan yang buruk, *cyberbullying* juga memberikan dampak yang mengerikan, mulai dari gangguan psikis korban, hingga munculnya kasus bunuh diri. Tujuan dari penelitian ini adalah mengidentifikasi konten yang mengandung makna perundungan secara daring (*Cyberbullying*) pada media sosial. Dalam kasus ini, penulis memilih media sosial *Twitter* sebagai obyek penelitian. Setidaknya, ada 1971 baris data yang telah dikumpulkan. Data - data tersebut berisi dua jenis cuitan baik cuitan yang memiliki kecenderungan *Cyberbullying* dan yang tidak. Untuk mencapai tujuan penelitian, peneliti menggunakan lima langkah penelitian, yaitu pengumpulan data, *preprocessing*, ekstraksi fitur, klasifikasi, dan evaluasi. Empat algoritma *Machine Learning* diimplementasikan dalam penelitian ini, yaitu *K-Nearest Neighbor (KNN)*, *Support Vector Machine with linear kernel (SVM)*, *Logistic Regression*, dan *Multinomial Naïve Bayes*. Dapat disimpulkan bahwa keempat algoritma tersebut memiliki performa yang relatif sama. Akurasi dari masing masing algoritma dituliskan sebagai berikut *Multinomial Naïve Bayes* 0.9616, *Logistic Regression* 0.9949, *Support Vector Machine with linear kernel* 0.9975, *and K-Nearest Neighbor (KNN)* 0.9188.

Kata kunci—Twitter, Cyberbullying, Machine Learning, Classification

I. PENDAHULUAN

Media sosial adalah sebuah wadah berbasis daring yang memungkinkan penggunanya berinteraksi dengan orang lain tanpa adanya batasan waktu, bahkan wilayah. Indonesia merupakan salah satu negara dengan angka pengguna media sosial tertinggi di dunia. Kementerian Komunikasi dan Informatika (Kemenkominfo) menyatakan 95% dari sekitar 63 juta pengguna internet adalah pengguna media sosial [1]. *Twitter* menjadi salah satu media sosial yang sering digunakan oleh masyarakat Indonesia. *Country Industry Head Twitter* Indonesia mengklaim bahwa Indonesia merupakan negara dengan pertumbuhan pengguna aktif harian *Twitter*-nya paling besar [2].

Pengguna *Twitter* di Indonesia tidak hanya merupakan pengguna perorangan. *Twitter* juga digunakan oleh lembaga - lembaga negara, komunitas, hingga toko berbasis daring. Tujuan penggunaannya juga bermacam - macam mulai dari promosi, hingga berbagi informasi tentang kinerja/hal yang telah dilakukan.

Namun tidak semua pengguna media sosial, menggunakan teknologi ini dengan bijak. Tak sedikit pengguna menggunakan *Twitter* untuk melakukan tindakan negatif seperti penipuan, penyebaran berita bohong, menulis hal - hal yang cenderung mengandung ujaran kebencian, hingga perundungan secara daring (*cyberbullying*). Hal ini tentu menjadi dampak negatif adanya media sosial di dalam masyarakat. Hal - hal negatif seperti itu cukup meresahkan. Dampak - dampak yang dapat ditimbulkan seperti kerusuhan karena menyebarnya berita bohong, kerugian materiil karena penipuan, hingga kasus karena *cyberbullying* [3]. Pemerintah menyatakan 84% remaja berusia 12 sampai 17 tahun di Indonesia menjadi korban tindakan perundungan (*bullying)* dan kebanyakan kasus *bullying* yang ditemukan merupakan *cyberbullying* [4]. Perundungan di media sosial dilakukan dengan menulis cuitan yang mengandung kata - kata hinaan, seperti tol*l, go*lok, kampung*n, bahkan kata kata yang menjurus kepada penghinaan terhadap suatu ras, suku, hingga agama.

Berdasarkan pada penelitian - penelitian serupa sebelumnya, belum ditemukan model yang dibuat khusus untuk mendeteksi / mengklasifikasi cuitan yang bermakna *cyberbullying* berbahasa Indonesia. Penelitian ini bertujuan untuk mewujudkan hal tersebut, dengan harapan bisa berkontribusi dalam pembuatan korpus berbahasa Indonesia.

Penelitian ini dilakukan dengan melakukan lima tahapan dalam *Text Mining* yaitu pengumpulan data, *preprocessing*, ekstraksi fitur, klasifikasi, dan evaluasi. Data yang dikumpulkan berupa cuitan - cuitan yang mengandung kata - kata perundungan dan yang tidak mengandung kata - kata

perundungan. Algoritma yang digunakan adalah Multinomial NBC, Linear SVM, Logistic Regression, dan KNN.

II. PENELITIAN TERKAIT

Penelitian tentang klasifikasi teks secara umum sudah banyak dilakukan oleh peneliti - peneliti sebelumnya. [5] Xiang et al. melakukan penelitian tentang ujaran kebencian *(Hate Speech)* pada media social *Twitter* dengan membagi *dataset* menjadi dua kelas, yaitu kelas *hate speech* dan kelas *non-hatespeech* kemudian menguji data dengan mengklasifikasikannya dengan menggunakan algoritma *Logistic Regression* dengan persentase nilai *True Positive* (TP) 75.1% dari total 4029 baris data cuitan.

Klasifikasi konten kasar pada cuitan media sosial *Twitter* dengan korpus berbahasa Indonesia juga pernah dilakukan. Hidayatullah et al. [6] mengklasifikasi cuitan menjadi dua kelas dan melakukan komparasi terhadap performa algoritma *Multinomial Naïve Bayes* dan *SVM* dalam melakukan klasifikasi.

Penelitian tentang deteksi *cyberbullying* pada media sosial sudah pernah dilakukan sebelumnya. Homa Hosseinmardi et al. melakukan penelitian tentang *cyberbullying* pada media sosial Instagram dengan memanfaatkan foto dan komentarnya, kemudian diklasifikasikan menggunakan NBC dan Linear SVM [7]. Penelitian tentang deteksi *cyberbullying* yang dipadukan dengan pendekatan psikologi juga pernah dilakukan sebelumnya. Model dibangun dengan menggunakan model Big Five and Triad dan algoritma *Random Forest* [8].

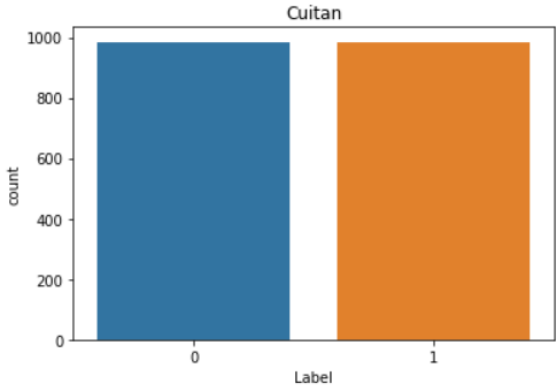
III. METODOLOGI

Bagian ini menjelaskan metodologi yang digunakan dalam penelitian. Identifikasi cuitan dilakukan dalam lima langkah penyelesaian, yaitu pengumpulan data, *preprocessing*, ekstraksi fitur, klasifikasi, dan evaluasi seperti yang terlihat pada gambar 1 dan akan dijabarkan pada sub bab berikutnya.

Gambar 1 Alur Penelitian

A. *Pengumpulan Data*

Twitter menjadi sumber data pada penelitian ini. Proses pengumpulan data dilakukan dengan memanfaatkan *Twitter* API. Dalam proses penggunaan *Twitter* API, penulis mendaftar sebagai *developer* untuk mendapatkan izin akses berupa *Consumer Key*, *Consumer Secret*, *Access Token*, dan *Access Secret*. Data yang diambil merupakan data cuitan berbahasa Indonesia. Cuitan yang terkumpul untuk membangun model pada penelitian ini adalah 1971 cuitan seperti yang terlihat pada gambar 2 dengan cacah masing masing yaitu 985 cuitan *non-cyberbullying* dan 986 cuitan *cyberbullying*.



Gambar 2 Cacah Data

Cuitan *non-cyberbullying* dihimpun dengan mengambil cuitan positif dari akun berita, tokoh agama, dan politisi. Contoh data yang merupakan cuitan *non-cyberbullying* dapat dilihat pada table 1.

Tabel 1
Cuitan *non-cyberbullying*

No	Cuitan
1	@pengguna1 Terima kasih. Semoga Allah SWT memberikan selalu memberikan kemudahan kepada kita semua dalam berbuat baik. Aamiin.
2	@pengguna2 Kematian, jodoh, Allah SWT yang menentukan. Kita manusia hanya berdoa dan berusaha semoga agar diberikan yang terbaik .
3	@pengguna3 Terima kasih atas sarannya. Semoga ibu Chrisna Ida selalu dalam keadaan baik dan berbahagia.
4	@pengguna4 Aamiin. Jangan lupa untuk bangun Indonesia dengan karyamu. Terima kasih.

Sedangkan cuitan *cyberbullying* dihimpun dengan menggunakan kata kunci yang merupakan kata - kata yang bermakna merundungan seperti bel*gu, d*ngu, g*blok, iq jongk*k, jel*k, jij*k, kampungan, konte*, ngart*s, n*rak, sombong, s*ngong, tol*l, ud*k. kata - kata tersebut diperoleh berdasarkan hasil survei mengguanakan *Google Form* pada beberapa pengguna media sosial. Tabel 2 berisi contoh cuitan yang mengandung makna *cyberbullying*.

Tabel 2
Cuitan *cyberbullying*

No	Cuitan
1	@pengguna10 Duh yang bentukannya gini emang biasanya udik plus otaknya kosong. Maap ya saya stereotyping but tru :)

2	@pengguna11 ya elu emang tolol, udah tau President panglima tertinggi abri, kalau jagoan situ kan pecatan... jgn disamain dong
3	@pengguna12 Justru kalian lah yg merusak kontestan demokrasi, tidak taat aturan, lha kok mlah nuduh balik, kan goblok sampean itu..

B. Preprocessing

Pada bagian ini, dijelaskan bagaimana data berupa cuitan - cuitan dibersihkan sehingga menjadi data yang baik dan terstruktur. Langkah - langkah *preprocessing* yang digunakan merujuk pada langkah - langkah *preprocessing* yang dilakukan [9] Hidayatullah et al. Adapun langkah - langkah *preprocessing* yang dilakukan antara lain:

- Menghilangkan URL
- Menghapus karakter NON-ASCII
- Menghapus angka, simbol dan tanda baca
- Menghapus *hashtag*, *username*, dan RT
- Penyeragaman huruf ke dalam bentuk *lowercase*
- Menghapus *stopwords*
- Menghapus kata yang terdiri dari satu huruf.

Hasil dari tahap *preprocessing* dapat dilihat pada tabel di bawah ini.

Tabel 3
Preprocessing

No	Sebelum	Sesudah
1	vivanewscom: Truk Hantam Kendaraan di Cianjur, Salah Satunya Rombongan Pengantin https://t.co/wjwvdPTRBa #vivanews'"	truk hantam kendaraan cianjur salah satunya rombongan pengantin
2	MUI Kalbar: Larangan Cadar-Celana Cingkrang Khawatir Timbulkan Gejolak https://t.co/SOouK1vBlP	mui kalbar larangan cadar celana cingkrang khawatir timbulkan gejolak
3	@pengguna13 @pengguna00 Biasa gaya hidup sok...ibarat org udik baru melek liat metropolitan jd jumpalitan dan gayanya gak karu karuan....	biasa gaya hidup sok org udik melek liat metropolitan jd jumpalitan dan gayanya gak karu karuan

C. Ekstraksi Fitur

Ekstraksi fitur adalah proses untuk menggambarkan karakteristik dari sebuah objek [10]. Ekstraksi fitur dilakukan menggunakan metode TF-IDF *(Term Frequency - Inverse Document Frequency)*. TF-IDF adalah perhitungan untuk menentukan bobot setiap token pada sekumpulan data [6]. Konsep dari metode TF-IDF adalah melihat seberapa penting suatu token data (kata yang sudah berbentuk vektor) dalam sebuah korpus. Persamaan untuk menghitung bobot tiap - tiap kata menggunakan TF-IDF didapatkan dari persamaan (1):

$$w_{t,d} = tf_{t,d} \; x \; log \frac{N}{df_t} \qquad (1)$$

di mana:

$tf_{t,d}$ = jumlah kemunculan token t pada dokumen d

df_t = jumlah dokumen yang memuat token t

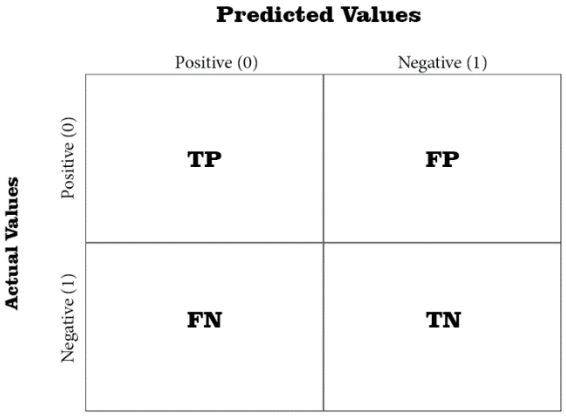
N = total dokumen

D. Klasifikasi

Bagian ini menjabarkan tentang proses klasifikasi cuitan dengan menggunakan beberapa algoritma *Machine Learning* yaitu *Multinomial Naïve Bayes*, *Linear SVM*, *Logistic Regression*, dan *KNN*. Algoritma tersebut dipilih karena terbukti memiliki akurasi yang sangat baik dalam klasifikasi data dalam bentuk teks, hal ini terbukti dalam penelitian penelitian sebelumnya [6] [7] [8].

E. Evaluasi

Confusion matrix merupakan matriks yang informatif untuk mengetahui performa model yang digunakan [6] dan digunakan sebagai tolok ukur performa klasifikasi dari algoritma yang digunakan pada tahap evaluasi.



Gambar 3 *Confusion Matrix*

Confusion Matrix menjadi sumber informasi apakah model yang digunakan memiliki performa yang baik atau tidak. Hal itu dapat dilihat pada angka - angka yang terdapat di dalamnya. Angka pada variabel TP *(True Positif)* dan variabel TN *(True Negative)* merepresentasikan total prediksi benar yang dilakukan oleh model. Sedangkan angka pada variabel FP *(False Positive)* dan variabel FN *(False Negative)* merepresentasikan total prediksi salah yang dihasilkan. Penghitungan performa model didapatkan dengan menghitung nilai *accuracy*, *precission*, *recall*, dan *F1-Score*, dengan rumus yang bisa dilihat pada persamaan (2), (3), (4) dan (5):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precission = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-Score = \frac{2\ x\ Precission\ x\ Recall}{Precission+Recall} \quad (5)$$

TP = Jumlah data kelas positif (0) diprediksi benar sebagai kelas positif (0)

FN = Jumlah data kelas positif (0) diprediksi salah sebagai kelas negatif (1)

TN = Jumlah data kelas negatif (1) diprediksi benar sebagai kelas negatif (1)

FP = Jumlah data kelas negatif (1) diprediksi salah sebagai kelas positif (0)

IV. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil dari penelitian yang telah dilakukan berdasarkan skenario dan langkah - langkah yang sudah dijabarkan sebelumnya. Data dibagi menjadi dua kelas, yaitu data *testing* dan data *training* dengan cacah masing - masing 0.2 data *testing* dan 0.8 data *training.*

Berikut disajikan tabel hasil dari masing - masing *confusion matrix* algoritma *Machine Learning*:

Tabel 4
confussion matrix Multinomial Naïve Bayes

Multinomial Naïve Bayes			
Actual Value	Predicted Value		
		0	**1**
	0	**189**	**15**
	1	**0**	**190**

Tabel 5
confussion matrix Logistic Regression

Logistic Regression			
Actual Value	Predicted Value		
		0	**1**
	0	**204**	**15**
	1	**2**	**188**

Tabel 6
confussion matrix Linear SVM

Linear SVM			
Actual Value	Predicted Value		
		0	**1**
	0	**203**	**1**
	1	**0**	**190**

Tabel 7
confussion matrix KNN

KNN			
Actual Value	Predicted Value		
		0	**1**
	0	**174**	**30**
	1	**2**	**188**

Berdasarkan *confussion matrix* yang sudah diperoleh, secara otomatis nilai dari *accuracy, precission, recall,* dan *F1-Score* dapat diketahui. Berikut disajikan hasil dari masing - masing variabel tersebut pada tabel 7

Tabel 8
dari accuracy, precission, recall, dan F1-Score

	Accuracy	*Precision*	*Recall*	*F1-Score*
Multinomial Naïve Bayes	0.961	0.96	0.96	0.96
Logistic Regression	0.994	0.99	0.99	0.99
Linear SVM	0.997	1.00	1.00	1.00
KNN	0.918	0.93	0.92	0.92

Berdasarkan hasil yang ditunjukkan pada tabel 8, algoritma *Linar SVM* terbukti lebih unggul dibanding dengan tiga algoritma lainnya baik dari segi dari *accuracy, precission, recall,* maupun *F1-Score*. *Logistic Regression* menjadi algoritma terbaik kedua dengan nilai *accuracy* hanya selisih 0.003 dari *Linear SVM*. *Multinomial Naïve Bayes* dan KNN secara berurutan menempati posisi terbaik ketiga dan keempat.

V. KESIMPULAN

Berdasarkan hasil yang telah dicapai, penelitian ini telah berhasil melakukan identifikasi cuitan bermakna *cyberbullying* pada media sosial *Twitter* dengan melakukan klasifikasi antara dua kelas cuitan yang tersedia pada *dataset*.

Hasil evaluasi dari masing - masing algoritma *Machine Learning* menempatkan *Linear SVM* sebagai algoritma terbaik dalam mengklasifikasi data cuitan dibanding dengan tiga algoritma lainnya. Hal ini dibuktikan dengan nilai dari *accuracy, precission, recall,* dan *F1-Score* dari *Linear SVM* paling tinggi dengan nilai masing - masing 0.997; 1.00; 1.00; 1.00.

Namun dapat disimpulkan bahwa tiga algoritma lainnya juga merupakan algoritma yang cukup baik dalam mengklasifikasi data teks. Ini dibuktikan dengan nilai dari *accuracy, precission, recall,* dan *F1-Score* masing - masing algoritma lebih dari 0.9.

Penelitian ini masih sebatas model dalam klasifikasi cuitan bermakna *cyberbullying* pada media sosial. Dengan demikian penulis berharap penelitian ini bisa dikembangkan dengan melakukan penambahan data *training*, dan membuat antarmuka dari pengujian model.

VI. References

[1] brs, "Kominfo : Pengguna Internet di Indonesia 63 Juta Orang," Kementrian Komunikasi dan Informatika Republik Indonesia, 7 November 2013. [Online]. Available: https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker. [Accessed 13 November 2019].

[2] B. Clinten, "Pengguna Aktif Harian Twitter Indonesia Diklaim Terbanyak," Kompas, 30 Oktober 2019. [Online]. Available: https://tekno.kompas.com/read/2019/10/30/16062477/pengguna-aktif-harian-twitter-indonesia-diklaim-terbanyak. [Accessed 2019 November 2019].

[3] S. Hinduja and J. W. Patchin, "Bullying, Cyberbullying, and Suicide," *International Academy for Suicide Research,* vol. 14, p. 209, 2010.

[4] B. A. Laksana, "Mensos: 84% Anak Usia 12-17 Tahun Mengalami Bullying," Detik News, 21 Juli 2017. [Online]. Available: https://news.detik.com/berita/3568407/mensos-84-anak-usia-12-17-tahun-mengalami-bullying. [Accessed 13 November 2019].

[5] G. Xiang, B. Fan, L. Wang, J. I. Hong and C. P. Rose, "Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus," 2012.

[6] A. F. Hidayatullah, A. A. F. Yusuf, K. P. Juwairi and A. N. R. Nayoan, "Identifikasi Konten Kasar pada Tweet Bahasa Indonesia," *Jurnal Linguistik Komputasional,* vol. 2, no. 1, p. 3, 2019.

[7] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," *Association for the Advancement of Artificial,* pp. 5-6, 2015.

[8] V. Balakrishnan, S. Khan, T. Fernandez and H. R. Arabnia, "Cyberbullying detection on twitter using Big Five and Dark Triad features," *Personality and Individual Differences,* no. 141, pp. 252-257, 2019.

[9] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *Journal of Physics,* vol. 801, 2017.

[10] D. Satria and MUSHTHOFA, "Perbandingan Metode Ekstraksi Ciri Histogram dan PCA untuk," *Jurnal Ilmu Komputer dan Agri-Informatika,* vol. 2, pp. 20-28, 2013.