Analisis Perbandingan Tingkat Performa Algoritma SVM, Random Forest, dan Naïve Bayes untuk Klasifikasi Cyberbullying pada Media Sosial

Theofilus Arifin[1], Hans Wirjawan[2]

[12]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[1]s160420046@student.ubaya.ac.id, [2]s160420108@student.ubaya.ac.id

Abstrak

Pada Januari 2022, terdapat 4,95 miliar pengguna internet di seluruh dunia dengan waktu akses rata-rata sebanyak 135 hingga 193 menit per hari. Kemajuan teknologi di bidang informasi dan komunikasi tidak sejalan dengan perilaku masyarakat di sosial media. Pada tahun 2017, tercatat sebagian besar kasus cyberbullying berasal dari sosial media. Sosial media adalah sebuah platform digital yang digunakan untuk bersosialisasi dengan orang lain secara online. Sosial media yang paling sering digunakan di dunia pada tahun 2017 adalah Facebook, Youtube, Whatsapp, Instagram, dan Twitter. Menurut data statistik yang pernah diperoleh, 54% dari 10000 peserta survei The Annual Bullying telah mengalami tindak kekerasan *cyberbullying*. Pada penelitian ini dilakukan sebuah proses analisis sentimen *cyberbullying* yang disampaikan dari berbagai sosial media yang ada di dunia. Analisis sentimen ini digunakan untuk menentukan apakah teks tersebut memiliki emosional *cyberbullying* atau tidak. Jumlah data yang digunakan sebanyak 46000 teks yang berbeda dengan rincian kurang lebih 8000 teks untuk setiap kategori yang ada yaitu *cyberbullying* usia, *cyberbullying* etnis, *cyberbullying* jenis kelamin, *cyberbullying* agama, *cyberbullying* lainnya dan bukan *cyberbullying* dan paling tidak ditemukan 1000 teks lebih yang mengandung "*fuck"*. Metode penelitian ini menggunakan fitur TF-IDF(Term Frequency-Inverse Document Frequency) dan 3 model untuk mengklasifikasikannya yaitu SVM (Support Vector Machine), RF (Random Forest), dan Naive Bayes. Berdasarkan hasil penelitian yang dilakukan Algoritma SVM dan Random Forest memiliki performa yang terbaik dengan *evaluation matrix* mencapai *precision* 82%, *recall* 83%, *accuracy* 83% dan *precision* 83%, *recall* 82%, *accuracy* 82%.

Kata Kunci: *cyberbullying, svm, random forest, naïve bayes, media sosial*.

Comparative Algorithm Performance Analysis of SVM, Random Forest, and Naïve Bayes for Cyberbullying Classification on Social Media

Abstract

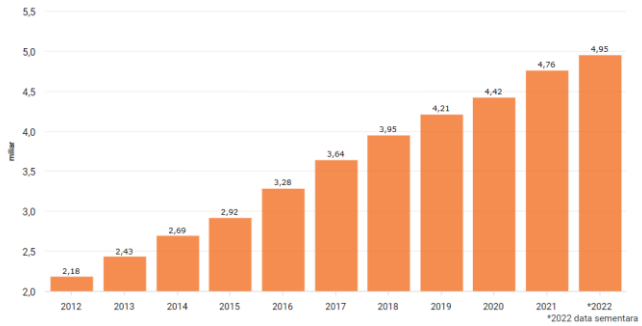
In January 2022, the number of Internet users in the world has reached 4,95 billion with an average of activity of 135 to 193 minutes per day. Technological advances in information gathering and communication are not in line with the improvements in people's behavior on social media. It is recorded that most of cyberbullying incidents in 2017 originate from social media. Social media are media technologies that facilitate interaction between people on the Internet. The most used social media in the world are Youtube, Instagram, Snapchat, Whatsapp, dan Twitter. There is a static data indicating that 54% of participants in The Annual Bullying Survey have experienced cyberbullying. For this research, a sentiment analysis was performed on a collection of texts from several social media platforms around the world. There are about 46000 different texts with an approximately 8000 text for each category, namely age cyberbullying, ethnicity cyberbullying, gender cyberbullying, religion cyberbullying, other type of cyberbullying and not cyberbullying and approximately 1000 text consist word "fuck". Sentiment analysis is the process of classifying sentiments in text, whether or not the text contains cyberbullying emotions. This research classifies the type of cyberbullying using the TF-IDF (Term Inversion Frequency Document) function and 3 models namely SVM (Support Vector Machine), RF (Random Forest) and Naive Bayes. Result highlight that SVM and Random Forest performed the best and achieved a precision 82%, recall 83%, accuracy 83% and precision 83%, recall 82%, accuracy 82% using evaluation matrix.

Keywords: *cyberbullying, svm, random forest, naïve bayes, social media.*

I. PENDAHULUAN

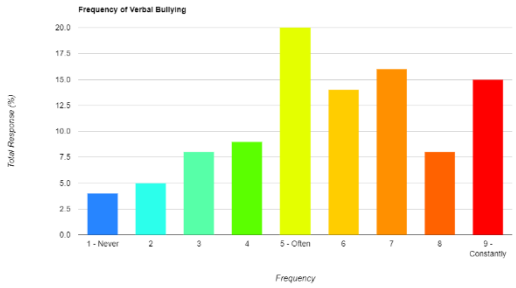
Internet adalah jaringan besar yang saling menghubungkan mulai dari jaringan-jaringan komputer yang satu ke jaringan-jaringan komputer diseluruh dunia melalui satelit. Salah satu manfaat dari perkembangan teknologi internet ini adalah sarana untuk berkomunikasi. Sarana komunikasi yang sangat populer saat ini adalah media sosial. Rulli Nasrullah [1] Media sosial adalah medium di internet yang memungkinkan pengguna merepresentasikan dirinya maupun berinteraksi, bekerja sama, berbagi, berkomunikasi dengan pengguna lain membentuk ikatan sosial secara virtual. Selain digunakan untuk melakukan komunikasi dan interaksi dengan orang lain terkadang medial sosial digunakan untuk tindakan yang kurang baik.

Ada banyak perubahan gaya hidup sejak pandemi COVID-19. Pada tahun 2021 terjadi peningkatan sebesar 7% dari tahun 2020 yaitu sebanyak 4,76 miliar orang yang telah menggunakan internet [2].



Gambar 1. Jumlah Pengguna Internet di Dunia

Ditch the Label [3] yang merupakan salah satu lembaga terbesar gerakan anti *bullying* di dunia, pernah mencatat bahwa terdapat 20% dari mereka yang mengikuti survei dan menjawab sering menerima tindakan *bully* secara verbal.



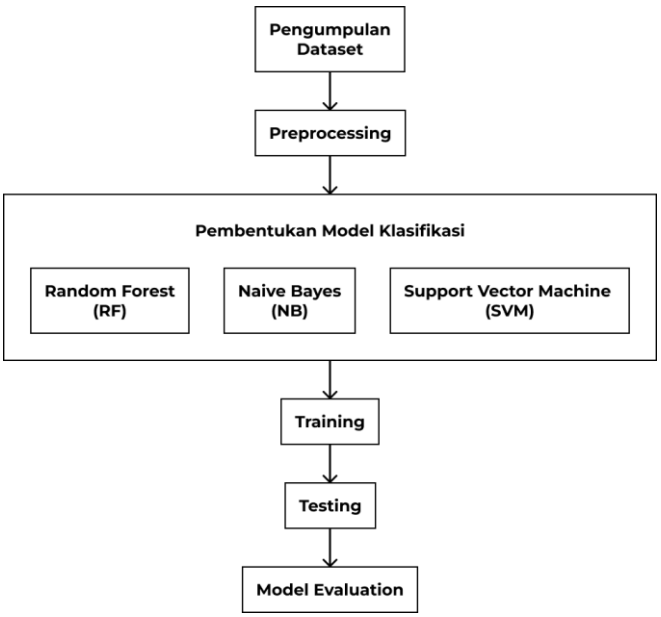
Gambar 2. Responden Ditch the Label mengenai verbal bullying

Menurut Rufa Mitsu, et al [4] *Cyberbullying* memiliki potensi untuk sering terjadi daripada perundungan fisik karena pada umumnya perundungan fisik itu hanya terjadi di kalangan tertentu seperti sekolah dan universitas. Potensi tersebut disebabkan karena tidak ada yang dapat mengawasinya terus-menerus. Tindakan *cyberbullying* ini dapat menimbulkan dampak yang negatif seperti depresi, keinginan untuk bunuh diri, narkoba, dan lain-lain.

Melihat data-data yang ada, diperlukannya sebuah metode yang dapat mengelompokkan apakah sebuah teks komentar yang ada pada media sosial termasuk dalam *cyberbullying* atau tidak. Metode tersebut adalah dengan membuat model klasifikasi *cyberbullying* dengan menggunakan algoritma Support Vector Machine (SVM), Random Forest(RF) dan Naive Bayes. Algortima ini merupakan salah satu model dari machine learning yang dapat melakukan klasifikasi teks dengan bantuan suatu algoritma yang dapat menghitung bobot setiap kata yang ada pada teks yaitu TF-IDF. Dengan demikian, proses untuk melakukan filter pada komentar media sosial dapat menjadi lebih efisien.

II. METODOLOGI PENELITIAN

Penelitian ini terdiri dari 5 tahap seperti yang ditunjukkan pada Gambar 3. Metode penelitian yang dilakukan dimulai dari tahap pengumpulan dataset, tahap pengolahan data, tahap pembentukan model klasifikasi, tahap pelatihan model, tahap uji coba model, dan tahap evaluasi model. Bagian ini akan menjelaskan proses-proses tersebut dengan lebih detail.



Gambar 3. Tahapan Proses Penelitian

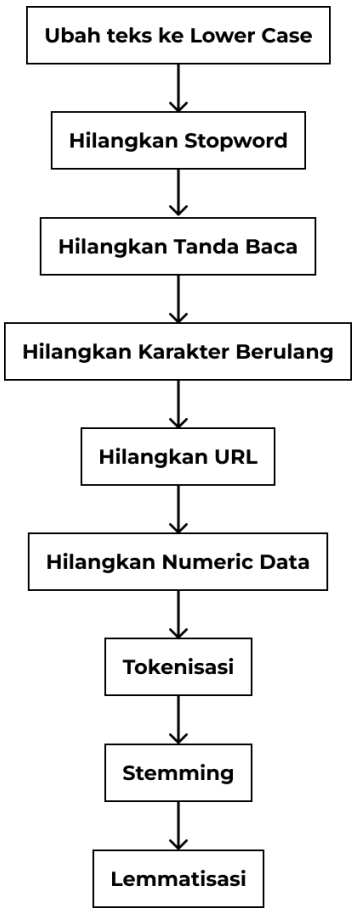
A. Pengumpulan Dataset

Proses pertama adalah pengumpulan *dataset*. *Dataset* yang digunakan pada penelitian ini adalah *Cyberbullying Classification* yang diambil dari *website* kaggle.com [5]. *Dataset* ini mengandung 46.017 data teks yang merupakan teks komen pada aplikasi twitter. Dataset ini telah dilabeli berdasarkan class *cyberbulllying* yaitu usia, etnis, jenis kelamin, agama, bukan *cyberbulllying*, *cyberbulllying* jenis lain. Masing-masing *class* terdiri atas 7998 jumlah tweet *cyberbulllying* agama, 7992 jumlah tweet *cyberbulllying*

umur, 7973 jumlah tweet *cyberbulllying* jenis kelamin, 7961 jumlah tweet *cyberbulllying* etnis, 7945 jumlah tweet bukan *cyberbulllying*, 7823 jumlah tweet *cyberbulllying* lain. Pada dataset ini tiap *class* telah memiliki jumlah data yang kurang lebih sama sehingga dapat diproses tanpa melakukan penyamaan jumlah data pada masing-masing *class*.

B. Pengolahan Data

Proses pengolahan data atau *data preprocessing* terbagi atas beberapa tahap. Tahapan-tahapam pengolahan data seperti yang ditunjukkan pada Gambar 4.

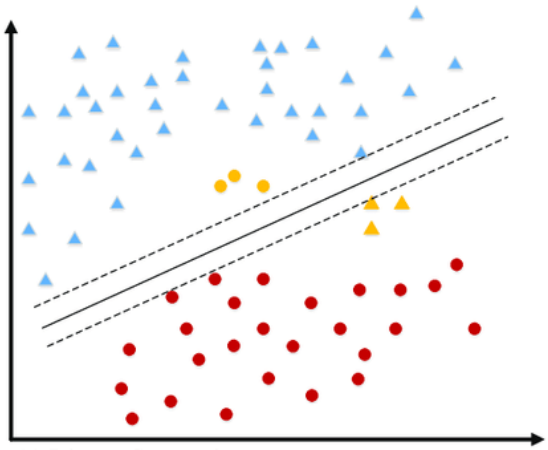


Gambar 4. Tahapan Proses Pengolahan Data

Proses pengolahan data dimulai dengan mengubah keseluruhan teks pada *dataset* ke *lower case*. Selanjutnya, daftar *stopword* dibuat dan *stopword* dihilangkan dari keseluruhan teks pada *dataset*. Proses selanjutnya adalah menghilangkan tanda baca pada keseluruhan teks di *dataset.* Lalu, menghilangkan karakter berulang pada keseluruhan teks di *dataset*. Selanjutnya adalah proses penghilangan URL pada teks. Setelah itu data numerik akan dihilangkan dari keseluruhan teks pada dataset. Lalu tokenisasi dilakukan. Tokenisasi mebuat kalimat teks menjadi array yang berisi tiap kata di dalamnya. Selanjutnya stemming akan dilakukan terhadap data yang telah ditokenisasi. Terakhir, proses lemmatisasi dilakukan dengan mengabungkan kata-kata yang bermakna sama.

C. Pembentukan Model Klasifikasi

Pada penelitian ini, terdapat 3 model yang digunakan yaitu Naïve bayes, Random Forest, dan Support Vector Machine. Ketiga model ini telah disediakan oleh library sklean dan pada penelitian ini ketiga model tersebut dibuat menggunakan library sklearn. Naïve Bayes dan Random Forest menggunakan parameter default sedangkan Support Vector Machine menggunakan SVC yaitu kondisi dimana *hyperlane* yang digunakan adalah linear. Illustrasi dari model SVC dapat dilihat pada Gambar 5.



Gambar 5. Illustrasi Model SVC

SVC yang digunakan menggunakan beberapa parameter yang telah dimodifikasi. Parameter yang telah dimodifikasi ada pada Tabel 1.

Tabel 1. Parameter SVC

Parameter	Deskripsi
Kernel	linear
c	1

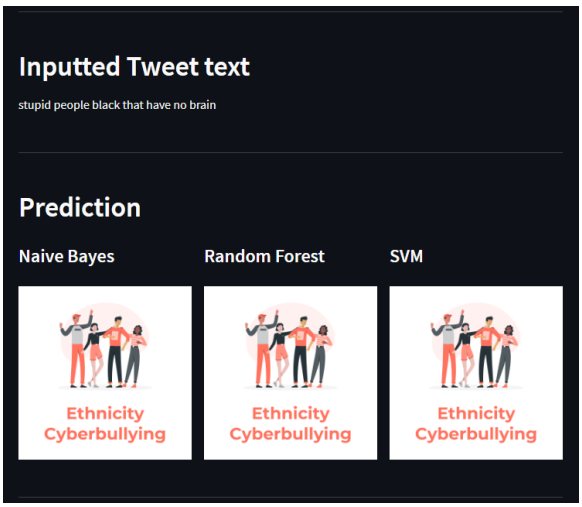
D. Training

Tahapan *training* merupakan tahapan pelatihan model menggunakan data *training* yang disebut X train dan y train. X train adalah data teks tweet sedangkan y train adalah *class* dari hasil klasifikasi teks tweet yang sudah ada. Data *training* diambil secara acak dari seluruh *text* yang ada. Seluruh *text* dibagi menjadi 70% data *training* dan 30% data *testing*. Sebelum melakukan *training* proses TF-IDF akan dilakukan terhadap X train untuk mengetahui tiap bobot setiap kata yang digunakan pada tiap data teks.

E. Testing

Testing merupakan proses percobaan klasifikasi dari model yang telah dilakukan *training*. Proses testing model dapat dilakukan pada *website* berbasis python menggunakan library steamlit yang telah disediakan [6]. Model yang telah dilakukan *training* telah tersimpan di dalam *website*. Ketika *user* melakukan input text, model akan melakukan klasifikasi menggunakan tiap model-

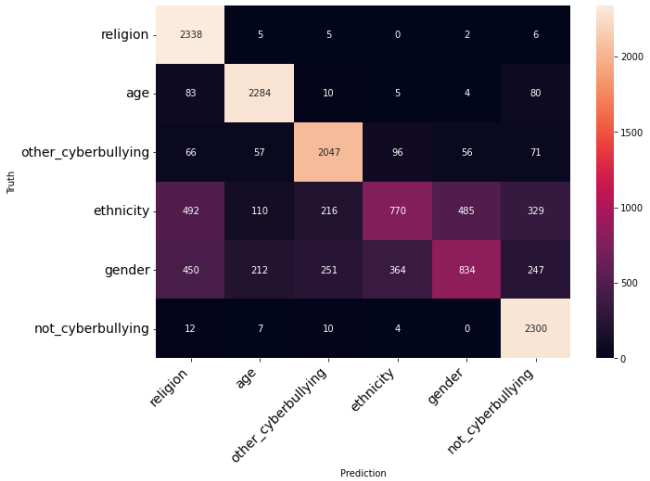
model yang ada yaitu Naïve Bayes, Random Forest, dan juga Support Vector Machine. Setelah itu hasil yang didapatkan akan ditampilkan. Hasil testing pada website dapat dilihat pada Gambar 6.



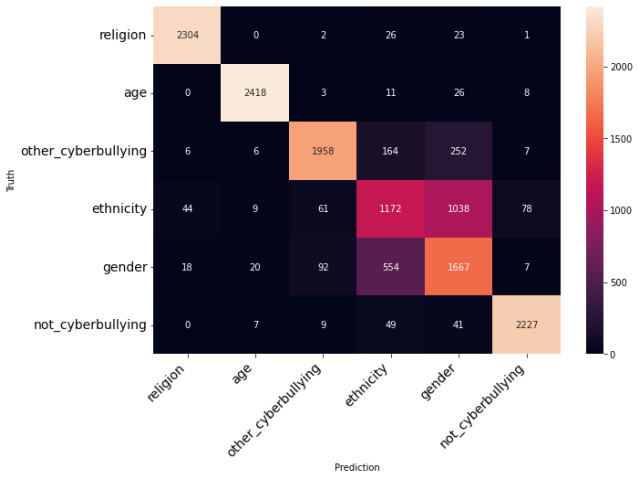
Gambar 6. Hasil testing pada *website*.

F. Validasi Model

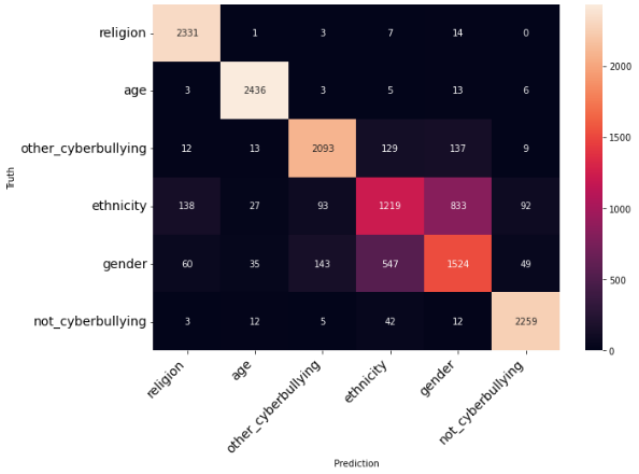
Berdasarkan hasil yang telah didapatkan dari model sebelumnya, akan dilakukan proses validasi model menggunakan penghitungan dari *confusion matrix* yang mencakup nilai *precision*, *recall*, dan *F1-score* dan juga Cross Validation Score. Pertama, pembuatan *confussion matrix* menggunakan *heatmap* untuk mempermudah pembacaan *confussion matrix* tersebut. Masing-masing heatmap untuk tiap model dapat dilihat pada Gambar 7 hingga Gambar 9.



Gambar 7. *Confussion matrix* naïve bayes.



Gambar 8. *Confussion matrix* random forest.



Gambar 9. *Confussion matrix* SVM.

Selain itu, pada tahapan ini juga dilakukan perbandingan dari akurasi pada tiap model. Rumus untuk menghitung *accuracy* model dapat dilihat pada persamaan (1), *precision* pada persamaan (2), *recall* pada persamaan (3), dan *F1-score* pada persamaan (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ score = \frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

Accuracy menunjukkan akurasi dari *classifier*. *Precission* menunjukkan rasio prediksi positif yang benar terhadap total prediksi positif. *Recall* menunjukkan rasio prediksi positif terhadap total prediksi positif. F1 menunjukkan tingkat akurasi model *dataset*. Hasil perhitungan performa tiap *classifier* menggunakan persamaan 1 hingga 4 dapat dilihat pada tabel 2.

Tabel 2. Tabel performa tiap *classifier*

Classifier	Accuracy	Precision	Recall	F1
NB	0.74	0.72	0.74	0.71
RF	0.82	0.83	0.82	0.82
SVM	0.83	0.82	0.83	0.82

III. HASIL DAN PEMBAHASAN

Dari hasil uji coba, kami memperoleh akurasi prediksi sebesar 74% dengan algoritma Naïve Bayes, 82% dengan algoritma Random Forest, dan 83% menggunakan algoritma Support Vector Machine. Setelah dilakukan penjalanan terhadap program, terlihat bahwa keakuratan prediksi metode-metode sebelumnya paling rendah pada 74% dan paling tinggi pada 83%. Hasil perbandingan akurasi tiap *classifier* dapat dilihat pada Tabel 3.

Tabel 3. Tabel akurasi tiap *classifier*

Classifier	Akurasi
NB	74%
RF	82%
SVM	83%

Hasil akurasi ini menunjukkan bahwa SVM jelas merupakan algoritma *classifier* yang paling baik untuk mengklasifikasi *dataset* ini dalam kelas-kelas *cyberbullying* yang telah ditentukan. Selain itu, *runtime* (waktu berjalan) dari algoritma-algoritma tersebut juga telah dicatat dan hasilnya ada pada Tabel 4.

Tabel 4. Tabel *runtime training* tiap *classifier*

Classifier	Runtime
NB	0.13s
RF	44m 29s
SVM	4m 52s

Hasil ini berasal dari *library* Python time. Disini, terlihat bahwa Random Forest memakan waktu terlama dalam proses *training* yaitu 44 menit 29 detik, dan Naïve Bayes memakan waktu tercepat pada *training* yaitu 0.13 detik. Dengan menghubungkan hasil lama *runtime* dengan hasil akurasi, dapat diinferensikan bahwa Naïve Bayes merupakan algoritma yang cepat namun kurang akurat pada *dataset* ini, Random Forest merupakan algoritma yang sangat lambat namun akurasi yang dihasilkan cukup baik pada *dataset* ini, dan Support Vector Machine merupakan algoritma cukup cepat walaupun tidak secepat Random Forest yaitu dengan lama waktu *training* 4 menit 52 detik dan memiliki akurasi yang paling tinggi.

IV. KESIMPULAN

Dari hasil penelitian dan pembahasan yang sudah dilakukan dapat ditarik kesimpulan bahwa algoritma SVM memiliki tingkat performa yang paling tinggi dibandingkan dengan algoritma lain yaitu Naïve Bayes dan Random Forest dalam melakukan klasifikasi *cyberbullying* dari dataset yang diperoleh dari *Cyberbullying Classification* [5] karena memiliki tingkat akurasi hingga 83% dan waktu *runtime* yang cukup cepat yaitu 4 menit 52 detik.

REFERENSI

[1] Nasrullah, Rulli. 2015. Media Sosial; Persfektif Komunikasi, Budaya, dan Sosioteknologi. Bandung : Simbiosa Rekatama Media.

[2] Databooks (2022, Januari 26), "Pengguna Internet di Dunia Capai 4,95 Miliar Orang Per Januari 2022", https://databoks.katadata.co.id/datapublish/2022/02/07/pengguna-internet-di-dunia-capai-495-miliar-orang-per-januari-2022

[3] Ditch the Label (2017, July). "The Annual Bullying Survey 2017". https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf

[4] Mitsu, R., & Dawood, E. (2022). Cyberbullying: An Overview. *Indonesian Journal of Global Health Research*, *4*(1),195-202. https://doi.org/10.37287/ijghr.v4i1.927

[5] J. Wang, K. Fu, C.T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.

[6] "UAS ML · Streamlit." Accessed December 14, 2022. https://theofilusarifin-project-ml-webapp-f4zrxg.streamlit.app/#random-forest.