KLASIFIKASI TWEET CYBERBULLYING DENGAN MENGGUNAKAN ALGORITMA SVM DAN XGBOOST

Felix Fernando [1)]

[1)] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta 11440 Indonesia
email : felix.535210052@stu.untar.ac.id

ABSTRACT

Penelitian ini bertujuan untuk melakukan klasifikasi tweet cyberbullying dengan menggunakan dua pendekatan machine learning yaitu algoritma SVM dan XGBoost. Data input yang digunakan untuk analisis merupakan hasil pengambilan data tweet secara acak yang telah dilabelkan, dilakukan ekstraksi fitur dan pelatihan model menggunakan kedua algoritma tersebut. Hasil eksperimen menunjukkan bahwa XGBoost memberikan kinerja yang sedikit lebih unggul dalam mengklasifikasikan tweet cyberbullying dibandingkan SVM. Temuan ini memberikan kontribusi dalam pengembangan metode klasifikasi untuk mendeteksi cyberbullying di media sosial sehingga dapat membantu dalam memitigasi dampak negatif dari cyberbullying. Penelitian ini juga menunjukkan potensi penggunaan algoritma XGBoost dalam konteks deteksi cyberbullying di platform media sosial seperti Twitter.

Key words

Cyberbullying, SVM, XGBoost, Klasifikasi Teks, Twitter

1. Pendahuluan

Perkembangan teknologi digital dan akses internet telah mempermudah pelaku bullying untuk mencari korbannya secara online [1]. Dalam 5 tahun terakhir, persentase terjadinya kasus cyberbullying telah meningkat, pada kisaran 6.0% hingga 46.3% [2]. Para peneliti memiliki pandangan yang berbeda-beda terhadap definisi cyberbullying. Namun, mereka sepakat bahwa perilaku yang bersifat mengolok-olok seperti pelecehan, penistaan, dan penghinaan yang terjadi di ruang digital diklasifikasikan sebagai cyberbullying [3]. Berdasarkan penelitian yang telah dilakukan, ditemukan bahwa cyberbullying memiliki dampak negatif yang lebih parah jika dibandingkan dengan bullying tradisional. Hal ini disebabkan oleh video dan gambar yang dapat digunakan sebagai media untuk melecehkan, menista, dan menghina orang lain [4].

Berdasarkan survey yang dilakukan oleh statista.com, per Januari 2023 Indonesia merupakan negara dengan jumlah pengguna internet terbesar keempat, sejumlah 212.9 juta pengguna [5]. Pada tahun 2019, survei yang sama juga pernah dilakukan oleh kata.co.id, dan ditemukan bahwa pengguna internet yang terdaftar adalah sebanyak 130 juta pengguna [6]. Salah satu kontributor terbesar terhadap peningkatan tersebut adalah remaja berusia 15-19 tahun, dengan keterampilan teknologi mereka yang mempermudah mereka untuk mendapatkan informasi dari internet. Rasa keingintahuan pada fase labil inilah yang menyebabkan cyberbullying menjadi masalah yang serius [7].

Penelitian ini bertujuan untuk memberikan kontribusi dalam memitigasi dampak negatif dari cyberbullying, karena di Indonesia kasus cyberbullying sulit untuk terungkap. Fase labil mereka menyebabkan Remaja sangat rentan terhadap berbagai kenakalan dan penyimpangan, termasuk perilaku bullying baik secara tradisional maupun digital [8].

Beberapa penelitian terkait telah dilakukan sebelum pembuatan tulisan ini, diantaranya merupakan analisis sentimen tweet e-learning yang menggunakan metode k-Nearest Neighbor dengan nilai akurasi sebesar 53.03% [9], dan klasifikasi tweet cyberbullying menggunakan SentiStrength dengan nilai akurasi sebesar 60.5% [10].

2. Metode Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan merupakan data tweet pengguna Twitter yang diperoleh dari situs kaggle [11]. Data tersebut merupakan data yang bersumber dari sebuah penelitian cyberbullying yang terkait [12]. Data yang digunakan memiliki sampel sebanyak 47692 dengan teks tweet dan label tipe cyberbullying, seperti pada Tabel 1.

Tabel 1 Data Tweet

Tweet_text	Cyberbullying_type
In other words #katandandre, your food was crapilicious! #mkr	not_cyberbullying

Hey dumb fuck celebs stop doing something for people for publicity on Facebook... Wtf happen to life u niggers are cowards.	ethnicity
@discerningmumin Islam has never been a resistance to oppression. It has always been source of oppression to both believers and non believer	religion
Here at home. Neighbors pick on my family and I. Mind you my son is autistic. It feels like high school. They call us names attack us for no reason and bully us all the time. Can't step on my front porch without them doing something to us	age
Masters @ManhattaKnight I mean he's gay, but he uses gendered slurs and makes rape jokes	gender
@ikralla fyi, it looks like I was caught by it. I'm not a botter, so...	other_cyberbullying

2.2 Pra-proses Data

Sebelum melakukan klasifikasi, adapun tahapan ini yang berfungsi untuk melakukan transformasi pada data mentah menjadi data yang terstruktur dan siap diproses. Pada tahapan ini, beberapa pra-proses yang akan dilakukan antara lain case folding, filtering, tokenizing, dan stemming [13]:

1. Case Folding merupakan teknik untuk melakukan konversi dari huruf pada keseluruhan dokumen menjadi huruf kecil [13]. Implementasi dari teknik ini dapat dilihat pada Tabel 2.

Tabel 2 Teks setelah dilakukan Case Folding

Tweet_text	Cyberbullying_type
in other words #katandandre, your food was craplicious! #mkr	not_cyberbullying
why is #aussietv so white? #mkr #theblock #imacelebrityau #today #sunrise #studio10 #neighbours #wonderlandten #etc	not_cyberbullying
@xochitlsuckkks a classy whore? Or more red velvet cupcakes?	not_cyberbullying

2. Filtering diperlukan untuk membersihkan elemen yang tidak dibutuhkan atau tidak memiliki korelasi dengan cyberbullying pada teks, pada kasus ini elemen-elemen tersebut merupakan URL, simbol, angka, huruf yang berulang, dan kata sambung [14]. Bentuk teks setelah dilakukan pembersihan dapat dilihat pada Tabel 3.

Tabel 3 Teks setelah dilakukan Filtering

Tweet_text	Cyberbullying_type
words katandandre food craplicious mkr	not_cyberbullying
aussietv white mkr theblock imacelebrityau today sunrise studi neighbours wonderlandten etc	not_cyberbullying
xochitlsuckkks classy whore red velvet cupcakes	not_cyberbullying

3. Tokenizing atau tokenization berguna untuk memisahkan kalimat menjadi satuan kata dengan tujuan agar setiap kata dapat dibobotkan dan mempermudah proses klasifikasi dan mengurangi cost komputatif secara signifikan [15]. Kalimat yang sudah dipisah dapat dilihat pada Tabel 4.

Tabel 4 Teks setelah dilakukan Tokenization

Tweet_text	Cyberbullying_type
[words, katandandre, food, crapilicious, mkr]	not_cyberbullying
[aussietv,white,mkr,theblock, imacelebrityau,today,sunrise,studi, neighbours,wonderlandten,etc]	not_cyberbullying
[xochitlsuckkks,classy,whore,red, velvet,cupcakes]	not_cyberbullying

4. Stemming merupakan proses penyederhanaan kata, menghilangkan imbuhannya. Tahap ini merupakan proses terpenting pada praproses teks dikarenakan hasil dari tahap ini dapat mempengaruhi langsung pada hasil akhir klasifikasi [16]. Hasil akhir tahap pra-proses dapat dilihat pada Tabel 5.

Tabel 5 Teks setelah dilakukan Stemming

Tweet_text	Cyberbullying_type
[word,katandandr,food,crapilici, mkr]	not_cyberbullying
[aussietv,white,mkr,theblock, imacelebrityau, today,sunris,studi, neighbour,wonderlandten,etc]	not_cyberbullying
[xochitlsuckkk,classi,whore,red, velvet,cupcak]	not_cyberbullying

2.3 TF-IDF Vectorizer

Sebelum proses fitting, metode TF-IDF (Term Frequency-Inverse Document Frequency) akan digunakan untuk menghitung bobot atau frekuensi munculnya setiap kata [19]. Kata diambil dari data input yang sudah berbentuk token akan ditransformasi dari string panjang menjadi vektor agar dapat diproses sebagai input [14]. Metode ini populer karena efisien, mudah untuk diimplementasikan dan menawarkan performa yang baik [13]. Vektorisasi TF-IDF melakukan perhitungan dengan rumus sederhana yang dapat dilihat pada Persamaan 1 dan Persamaan 2 [19].

$$W_{ij} = tf_{ij} \times idf \quad \text{.....(1)}$$

$$idf = log\ (\frac{N}{df_j})$$ .....(2)

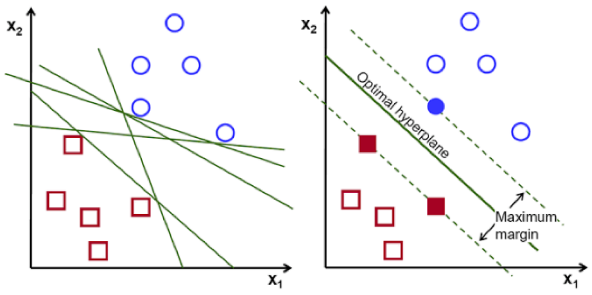
Dimana W adalah bobot dokumen TF-IDF dan N adalah jumlah dokumen. Jumlah variabel direpresentasikan sebagai i, dan j adalah jumlah data, serta t sebagai frekuensi variabel dalam data.

2.4 Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu algoritma klasifikasi, yang karakteristiknya adalah mencari hyperplane, yaitu garis yang memisahkan data antar kelas atau kategori seperti pada Gambar 1. SVM mencari hyperplane yang memiliki margin terbesar, yaitu jarak antar data tiap kelas terhadap hyperplane. Data yang paling dekat akan disebut sebagai support vector. SVM melakukan perhitungan hyperplane dengan Persamaan 3 [17].

$$w.x + b = 0$$ .....(3)

Dimana W adalah nilai parameter hyperplane yang dicari, x adalah data input dan b adalah nilai bias.



Gambar 1. Hyperplane dan Maximum Margin SVM

Margin yang dicari oleh hyperplane dapat dimaksimalkan dengan menggunakan sebuah persamaan garis seperti pada Persamaan 4 [18]:

$$ax + by + c = 0$$ .....(4)

Persamaan garis awal diubah dengan mengganti variabel x menjadi $x1$ dan variabel y menjadi $x2$. Konstanta a diubah menjadi $w1$ dan konstanta b diubah menjadi $w2$. Persamaan tersebut kemudian diubah untuk dimensi d menjadi persamaan yang lebih umum, yaitu pada Persamaan 5 dan Persamaan 6 [18].

$$\sum_{j=1}^{d} = w_j x_i + c = 0$$ .....(5)

$$g(x) = < w, x > + c = 0$$ .....(6)

2.5 XGBoost

XGBoost (eXtreme Gradient Boosting) merupakan salah satu algoritma machine learning, yang dapat digunakan untuk klasifikasi dan prediksi. XGBoost menggunakan beberapa decision tree sebagai metode boostingnya [19]. XGBoost merupakan algoritma yang bagus digunakan pada data berukuran kecil ke sedang [20]. Bobot pada setiap pohon dihitung menggunakan Persamaan 7.

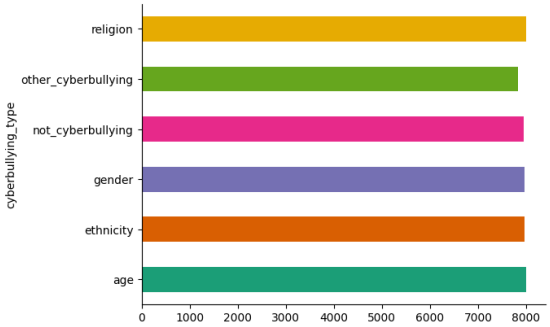
$$\hat{y}_i = \sum_k^k f_k(x_i), f_k \in F$$ .....(7)

Dimana $\hat{y}_i$ adalah prediksi untuk observasi ke-i, f_k adalah pohon ke-k, x_i adalah vektor fitur untuk observasi ke-i, dan F adalah himpunan pohon yang membentuk model.

3. Hasil Percobaan

Pada penelitian ini, klasifikasi apakah sebuah tweet termasuk cyberbullying dilakukan dengan algoritma SVM dan XGBoost. Setelah klasifikasi dengan kedua algoritma selesai dilakukan, hasil evaluasi kedua algoritma dari classification report yang dihitung menggunakan confusion matrix akan dibandingkan. Klasifikasi akan dilakukan dengan menggunakan data tweet yang diambil dari website kaggle, data ini memiliki sampel sebanyak 47692 dengan teks tweet dan label tipe cyberbullying yang dipecah menjadi beberapa kelas:

1. not_cyberbullying: Kelas ini mewakili tweet yang tidak bersifat mencela.
2. ethnicity: Kelas ini mewakili tweet yang bersifat mencela etnis atau suku.
3. religion: Kelas ini mewakili tweet yang bersifat mencela hal-hal yang berkaitan dengan agama.
4. age: Kelas ini mewakili tweet yang mencela hal-hal yang berkaitan dengan usia.
5. gender: Kelas ini mewakili tweet yang menghina atau mencela orientasi seksual orang lain.
6. other_cyberbullying: Kelas ini mewakili tweet cyberbullying yang tidak masuk kedalam keempat kelas diatas.



Gambar 2 Distribusi Data

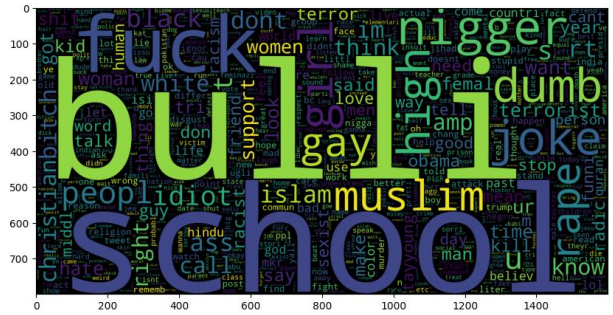
Berdasarkan Gambar 2, kita dapat melihat bahwa jumlah data telah terdistribusi secara merata untuk setiap kelas. Jumlah data setiap kelas dapat dilihat pada Tabel 6. Skenario eksperimen pada penelitian ini membagi

data menjadi dua bagian yaitu data testing dan training dengan rasio 80 : 20 agar tidak terjadi overfitting ataupun underfitting.

Tabel 6 Jumlah Data setiap Kelas

Kelas	Jumlah
not_cyberbullying	7945
gender	7973
religion	7998
other_cyberbullying	7823
age	7992
ethnicity	7961

Hasil dari tahapan pra-proses (stemming) telah menghasilkan statistik jumlah kemunculan kembali setiap kata pada dataset. Total kata yang menjadi fitur adalah sebanyak 329778 kata, dan kata-kata tersebut dapat direpresentasikan menggunakan grafik WordCloud, yang dapat dilihat pada Gambar 3. Pada grafik WordCloud, ukuran kata pada grafik menunjukkan frekuensi kemunculannya pada dataset yang digunakan.

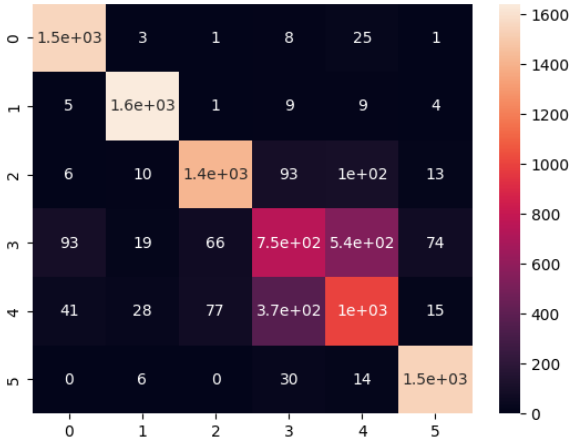


Gambar 3 WordCloud Data Tweet

Pada algoritma Support Vector Machine, dengan menggunakan kernel linear akurasi yang dihasilkan adalah 82.65%. Hasil evaluasi kinerja algoritma SVM yang dihitung menggunakan confusion matrix dapat dilihat melalui classification report dan heatmap pada Gambar 4 dan Tabel 7.

Tabel 7 Classification Report Algoritma SVM

	Precision	Recall	f1-score
0	0.91	0.98	0.94
1	0.96	0.98	0.97
2	0.91	0.86	0.88
3	0.60	0.49	0.54
4	0.59	0.65	0.62
5	0.93	0.97	0.95
accuracy			0.83
macro avg	0.82	0.82	0.82
weighted avg	0.82	0.83	0.82

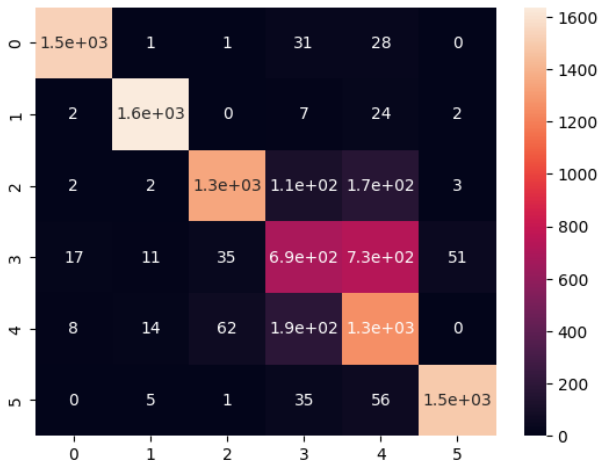


Gambar 4 Heatmap SVM

Heatmap dapat menunjukkan korelasi antar kelas pada dataset. Angka 0 hingga 5 merepresentasikan setiap kelas pada dataset. Grafik ini dapat merepresentasikan kemampuan algoritma dalam menghasilkan prediksi yang tepat. Berdasarkan Gambar 4, diagonal utama pada heatmap adalah titik dua kelas yang sama bertemu, sehingga model klasifikasi yang dapat memberikan prediksi akurat akan memiliki nilai yang tinggi pada diagonal utama heatmap. Titik 1,0 dengan nilai tertinggi 1.6e+03 (1600) menunjukkan bahwa model SVM dapat memprediksi tweet cyberbullying gender dengan sangat baik. Pada algoritma XGBoost, akurasi yang dihasilkan adalah 83.24%, sedikit lebih baik dibandingkan dengan SVM. Hasil evaluasi kinerja algoritma XGBoost yang dihitung menggunakan confusion matrix dapat dilihat melalui classification report dan heatmap pada Gambar 5 dan Tabel 8.

Tabel 8 Classification Report Algoritma XGBoost

	Precision	Recall	f1-score
0	0.98	0.96	0.97
1	0.98	0.98	0.98
2	0.93	0.82	0.87
3	0.65	0.45	0.53
4	0.55	0.82	0.66
5	0.96	0.94	0.95
accuracy			0.83
macro avg	0.84	0.83	0.83
weighted avg	0.85	0.83	0.83



Gambar 5 Heatmap XGBoost

Pada Gambar 5, Heatmap algoritma XGBoost menunjukkan hasil yang serupa seperti pada heatmap algoritma SVM. Berdasarkan heatmap dari kedua algoritma, terdapat sebuah pola dimana kedua model memiliki performa yang lebih rendah dalam memprediksi kelas 3, yang dapat diindikasikan oleh nilai heatmap yang rendah seperti 1.9e+02 pada 3,4 (190). Hal ini dapat disebabkan oleh kelas 3 yaitu other_cyberbullying yang tidak memiliki ciri khusus seperti jenis cyberbullying lainnya yang telah terkategori seperti religion, age, gender, dan ethnicity.

4. Kesimpulan

Dalam penelitian ini, klasifikasi tweet cyberbullying telah dilakukan menggunakan algoritma SVM dan XGBoost. Hasil yang diperoleh menunjukkan bahwa SVM memberikan akurasi sebesar 82.65%, sementara XGBoost memberikan akurasi sebesar 83.24%. Dengan demikian, dapat disimpulkan bahwa XGBoost sedikit lebih unggul dalam hal akurasi dibandingkan dengan SVM pada penelitian ini. Selain itu, kelebihan XGBoost meliputi kecepatan dalam menangani data yang besar, fleksibilitas dalam penggunaan untuk berbagai tujuan, dan kemampuannya dalam mengurangi overfitting. Untuk pengembangan selanjutnya, disarankan untuk melakukan optimasi parameter model agar dapat meningkatkan akurasi dan performa, serta mempertimbangkan penggunaan teknik ensemble learning untuk menggabungkan keunggulan dari kedua model..

REFERENSI

[1] Vismara, M., Girone, N., Conti, D., Nicolini, G. and Dell'Osso, B., 2022. The current status of Cyberbullying research: A short review of the literature. Current Opinion in Behavioral Sciences, 46, p.101152.

[2] Zhu, C., Huang, S., Evans, R. and Zhang, W., 2021. Cyberbullying among adolescents and children: a comprehensive review of the global situation, risk factors, and preventive measures. Frontiers in public health, 9, p.634909.

[3] Barlett, C.P., Simmers, M.M., Roth, B. and Gentile, D., 2021. Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic. *The journal of social psychology*, *161*(4), pp.408-418.

[4] Syah, R. and Hermawati, I., 2018. Upaya pencegahan kasus cyberbullying bagi remaja pengguna media sosial di Indonesia. *Jurnal Penelitian Kesejahteraan Sosial*, *17*(2), pp.131-146.

[5] Ani P., 2023. Countries with the largest digital populations in the world as of January 2023, Statista. https://www.statista.com/statistics/262966/number-of-internet-users-in-selected-countries.

[6] Febriana, T. and Budiarto, A., 2019, August. Twitter dataset for hate speech and cyberbullying detection in Indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)* (Vol. 1, pp. 379-382). IEEE.

[7] Jubaidi, M. and Fadilla, N., 2020. Pengaruh Fenomena Cyberbullying Sebagai Cyber-Crime di Instagram dan Dampak Negatifnya. *Shaut Al-Maktabah: Jurnal Perpustakaan, Arsip dan Dokumentasi*, *12*(2), pp.117-134.

[8] Navisa, A., Amalia, A., Setiya, D., Afra, S. and Pinilih, S.S., 2024. Kemampuan Mitigasi Cyberbullying terhadap Resiliensi Remaja. *Journal of Educational Innovation and Public Health*, *2*(1), pp.201-215.

[9] Geofany, N. and Liza, R., 2021, October. Klasifikasi Sentimen Tweet Pada Twitter Terhadap Pembelajaran E-Learning Menggunakan Metode k-Nearest Neighbor. In *SEMINAR NASIONAL TEKNOLOGI INFORMASI & KOMUNIKASI* (Vol. 1, No. 1, pp. 380-385).

[10] Khaira, U., Johanda, R., Utomo, P.E.P. and Suratno, T., 2020. Sentiment analysis of cyberbullying on twitter using SentiStrength. *Indones. J. Artif. Intell. Data Min*, *3*(1), p.21.

[11] Larxel, Cyberbullying Classification. Retrieved December 2020. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data.

[12] 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.

[13] Rahman, O.H., Abdillah, G. and Komarudin, A., 2021. Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(1), pp.17-23.

[14] Suharmanto, B., Kurnia, S., Prabowo, H.G., Pribadi, M.N.N. and Chamidah, N., 2022, August. Klasifikasi Tweet Cyberbullying dengan Menggunakan Algoritma Random Forest. In *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya* (Vol. 3, No. 2, pp. 753-764).

[15] Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R. and Roberts, A., 2020. Comparative analysis of text classification approaches in electronic health records. *arXiv preprint arXiv:2005.06624*.

[16] Rosid, M.A., Fitrani, A.S., Astutik, I.R.I., Mulloh, N.I. and Gozali, H.A., 2020, June. Improving text preprocessing for student complaint document classification using sastrawi. In *IOP Conference Series: Materials Science and Engineering* (Vol. 874, No. 1, p. 012017). IOP Publishing.

[17] Irmanda, H.N. and Astriratma, R., 2020. Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *4*(5), pp.915-922.
[18] Fikriani, A., Asror, I. and Murti, Y.R., 2019. Klasifikasi Kepribadian Berdasarkan Data Twitter dengan Menggunakan Metode Support Vector Machine. *eProceedings of Engineering*, *6*(3).
[19] Sinaga, H.H. and Agustian, S., 2022. Perbandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter. *Perbandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter*, *8*(03), pp.107-114.
[20] Edaño, M.G.E., Gonzales, R.J., Laguda, R.C.B. and De Goma, J.C., Stressor Classification of Filipino Political Tweets Using LDA, SVM, XGBoost, Logistic Regression. In *International Conference on Industrial Engineering and Operations Management. Istanbul, Turkey*.

Felix Fernando, mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara