Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

Penerapan Algoritma Regresi Linier pada Prediksi Tarif *Influencer* Media Sosial

Muhammad Riza Alifi*, Hashri Hayati, Cholid Fauzi

Jurusan Teknik Komputer dan Informatika, Politeknik Negeri Bandung, Bandung
Jl. Gegerkalong Hilir, Ciwaruga, Kec. Parongpong, Kabupaten Bandung Barat, Jawa Barat, Indonesia
Email: [1,*]muhammad.riza@polban.ac.id, [2]hashri.hayati@polban.ac.id, [3]cholid.fauzi@polban.ac.id
Email Penulis Korespondensi: muhammad.riza@polban.ac.id
Submitted: **11/10/2022**; Accepted: **30/10/2022**; Published: **31/9910/2022**

Abstrak−Disrupsi media sosial melahirkan profesi *influencer* yang memiliki kemampuan untuk mempengaruhi minat audiens terhadap produk dan jasa yang dipromosikan. Memanfaatkan jasa *influencer* pada media sosial dinilai lebih unggul dibandingkan media konvensional karena performa promosi lebih terukur. Umumnya tarif *influencer* menyesuaikan jumlah pengikut (*follower*), keterlibatan (*engagement*), dan jangkauan (*reach*). Namun, penetapan tarif tersebut tidak memiliki standar referensi, sehingga berpotensi menimbulkan kerugian bagi salah satu pihak. Penelitian ini bertujuan untuk memberikan solusi berupa model prediksi tarif *influencer* berbasis *machine learning* yang dapat dijadikan referensi untuk meminimalisir dampak kerugian baik bagi *influencer* dalam menawarkan tarif, maupun klien dalam menerima tawaran tarif. Tahapan penelitian ini terdiri dari studi pustaka, pengumpulan data, pra-pemrosesan data, pembangunan model regresi linier, dan evaluasi model. Penelitian ini menghasilkan lima varian model. Salah satu model terbaik menghasilkan nilai MAE: 145401.484375, MSE: 7.222241e+10 dan RMSE: 268742.250, yang dipengaruhi oleh nilai *hyperparameter learning rate*: 0.001 dan *epoch*: 1.000. Model tersebut belum sepenuhnya mampu mewakili data mayoritas yang diujicoba. Salah satu penyebab model belum ideal adalah terbatasnya jumlah *dataset* yang digunakan untuk proses *training*, hanya 161 baris disertai dengan 4 atribut yang berkorelasi positif. Namun, dalam perspektif penggunaan *dataset* yang sangat terbatas, model yang dibangun pada penelitian ini cukup berhasil karena beberapa hasil prediksi sangat mendekati nilai aktual, salah satunya terdapat nilai prediksi yang memiliki selisih *error* −347,69.

Kata Kunci: Prediksi Tarif Influencer; Influencer Media Sosial; Machine Learning; Regresi Linier

Abstract−The influencer industry has emerged as a result of social media disruption, and its members can affect audience interest in the goods and services being advertised. Because advertising performance on social media is more quantifiable than it is with traditional media, using influencer services is thought to be preferable. Influencer rates are often dependent on reach, engagement, and follower count. However, since there is no reference standard used in determining the prices, it could harm one of the parties. In order to reduce the impact of losses for both influencers in giving rates and clients in accepting rate offers, this study intends to propose a solution in the form of a machine learning-based influencer rate prediction model that can be used as a reference. The stages of this study are literature review, data gathering, pre-processing of the data, linear regression model development, and model evaluation. Five different models were produced as a result of this investigation. One of the best models has an MAE of 145401.484375, an MSE of 7.222241e+10, and an RMSE of 268742.250. These findings are affected by the hyperparameter learning rate of 0.001 and the epoch of 1,000. Most of the test data have not been completely represented by the model. The little number of datasets utilized for training, only 161 rows with 4 positively correlated attributes, is one of the reasons why the model is not really optimal. Nevertheless, from the standpoint of using a relatively small dataset, the model developed in this study is quite successful because several of the prediction results are fairly near to the real value, one of which is the prediction value with an error difference of −347.69.

Keywords: Influencer Rate Prediction; Social Media Influencer; Machine Learning; Linear Regression

1. PENDAHULUAN

Keberadaan media sosial telah mendorong transformasi aktivitas pemasaran [1], [2] dan strategi promosi [3]–[5] yang sebelumnya berpusat pada pemanfaatan media konvensional, seperti: TV, Radio, Koran dan lainnya, perlahan menjadi lebih terdistribusi [6]. Hal ini karena setiap pengguna media sosial memiliki potensi dan segmentasi masing-masing dalam mempengaruhi dan membangkitkan minat para audiens. Selain itu, media sosial memiliki potensi jangkauan audiens yang luas, sama seperti media konvensional [6], namun bisa lebih terukur karena pada media sosial dapat merekam setiap interaksi para audiens [7], [8]. Beberapa penggiat media sosial pun ada yang menekuninya menjadi suatu profesi, umumnya dikenal oleh publik sebagai *influencer* [3], [9], [10].

Influencer menyediakan tarif (*rate card*) yang sangat beragam dalam memasarkan produk atau jasa melalui konten di media sosial. Umumnya tarif *influencer* mengikuti jumlah pengikut (*follower*) yang dimiliki [11]. Namun, jumlah pengikut tidak selalu menjamin untuk memperoleh performa yang baik, karena tidak menutup kemungkinan pertumbuhan pengikut pada akun *influencer* dibangun dengan cara yang tidak organik atau adanya rekayasa [12] dengan memanfaatkan bot dan sindikasi [12], [13]. Terdapat satuan metrik yang dapat dijadikan sebagai indikator performa selain jumlah pengikut, yaitu keterlibatan (*engagement*), dan jangkauan (*reach*) yang tercatat pada akun *influencer* [12]–[15]. Pengukuran tarif berdasarkan beberapa satuan metrik tersebut dapat dipertimbangkan untuk memperoleh referensi tarif yang representatif terhadap performa yang dihasilkan oleh *influencer*.

Tarif yang berdasarkan performa aktual *influencer* diharapkan dapat meminimalisir isu atau potensi negatif yang mungkin timbul [16], karena dengan tidak adanya standar referensi untuk menentukkan tarif ini dapat

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

merugikan *influencer*, maupun klien. Kerugian yang dapat dialami oleh *influencer*, terutama bagi yang baru menjalani profesi tersebut, yaitu menawarkan tarif terlalu rendah padahal memiliki potensi tinggi. Sebaliknya, jika *influencer* menawarkan tarif terlalu tinggi padahal memiliki potensi rendah, maka klien tidak akan berkesinambungan untuk berminat kembali menggunakan jasanya. Sementara itu, kerugian yang mungkin dialami oleh klien yaitu menerima penawaran tarif terlalu tinggi padahal memiliki potensi rendah.

Pada penelitian Khogasteh et al. (2021) mengangkat isu transparansi dan keaslian (*authenticity*) profil akun *influencer*, dengan mencoba melakukan prediksi jangkauan aktual audiens berdasarkan satuan metrik jumlah pengikut dan keterlibatan menggunakan algoritma *machine learning*--regresi linier [12]. Penelitian Arora et al (2019). mengajukan mekanisme pengukuran indeks *influencer* untuk menyelidiki dampak keterlibatan, jangkauan, sentimen dan pertumbuhan keterlibatan pada *influencer* dengan pendekatan *machine learning*--regresi [17]. Kedua studi [12], [17] memiliki kesamaan dalam menggunakan metode regresi untuk mengukur nilai kuantitatif, serta menyertakan satuan metrik keterlibatan sebagai bagian yang penting.

Studi mengenai prediksi dalam menentukkan tarif *influencer* saat ini masih belum ditemukan. Namun, terdapat beberapa studi [12], [17] yang beririsan karena memiliki kesamaan objektif dalam hal melakukan pengukuran secara kuantitatif. Kedua studi tersebut melakukan pengukuran performa (variabel dependen) *influencer* berdasarkan satuan metrik yang dinilai relevan sebagai indikator performa (variabel independen).

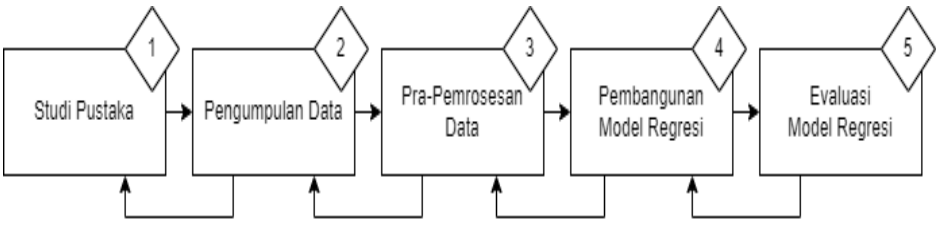
Studi pertama mencoba menjawab pertanyaan, (1) "bagaimana jangkauan konten *influencer* dapat diprediksi menggunakan *machine learning* berdasarkan satuan metrik keterlibatan sebagai masukan?" (2) "bagaimana kualitas hasil prediksi yang dihasilkan?". Studi ini memanfaatkan metode regresi linier untuk menjawab pertanyaan tersebut, bahwa satuan metrik keterlibatan dan jumlah pengikut memiliki korelasi linieritas yang signifikan untuk memprediksi jangkauan. Serta, kualitas hasil prediksi berdasarkan keterlibatan dan jumlah pengikut dapat dilihat berdasarkan nilai R-Squared (0,7011 dan 0,3809) dan Mean Absolute Error (14453,12 dan 10156,02). Semakin tinggi nilai R-Squared (dalam skala satu) mengindikasikan bahwa semakin baik variabel independen menjelaskan variabilitas terhadap variabel dependen. Sementara itu, semakin rendah nilai Mean Absolute Error mengindikasikan bahwa semakin tinggi akurasi model prediksi yang dihasilkan [12].

Studi kedua mengajukan suatu mekanisme untuk mengukur indeks *influencer* pada beberapa platform media sosial dengan pendekatan *machine learning*--regresi linier, antara lain: Ordinary Least Squares (OLS), K-NN Regression, Support Vector Regression (SVR) dan Lasso Regression. Hasil dari penelitian ini menunjukkan bahwa pendekatan *machine learning ensemble* yang terdiri dari perpaduan keempat algoritma *machine learning* menghasilkan akurasi paling tinggi (93,7%), diikuti algoritma *machine learning* K-NN Regression (93,6%). Selain itu, studi ini memberikan wawasan bahwa satuan metrik keterlibatan, jangkauan, sentimen dan pertumbuhan keterlibatan memiliki pengaruh terhadap nilai indeks *influencer* [17].

Studi terkait yang telah dibahas [12], [17], belum spesifik membahas pengukuran tarif *influencer*. Namun, terdapat potensi untuk mengadopsi metode pengukuran jangkauan konten *influencer* menjadi pengukuran tarif *influencer*, termasuk metode evaluasi yang terkait, dan penggunaan satuan metrik yang dapat berperan sebagai variabel independen terhadap tarif sebagai variabel dependen. Berdasarkan potensi yang ditemukan dari beberapa studi tersebut, penelitian yang akan diajukan yaitu melakukan prediksi tarif *influencer* menggunakan algoritma *machine learning*--regresi linier, dengan menyertakan satuan metrik keterlibatan, jangkauan dan jumlah pengikut. Pemilihan satuan metrik pada penelitian ini akan menyesuaikan ketersediaan data yang memungkinkan dikoleksi melalui akses publik saja, sehingga tidak perlu meminta data hasil analisis ke setiap *influencer* karena akan membutuhkan waktu yang sangat lama dan tentunya tidak semua *influencer* terbuka memberikan data yang diperlukan. Selanjutnya penelitian yang akan dilakukan mempertimbangkan penggunaan alat bantu berupa *framework* Tensorflow dalam membentuk model regresi berbasis *neural network* [18]. Alasan memilih *framework* Tensorflow untuk mendukung penelitian ini, yaitu didasarkan pada adanya dukungan terhadap penerapan *deep learning* yang merupakan *state-of-the-art* pada bidang *machine learning* saat ini, serta menawarkan fleksibilitas dan skalabilitas dalam melakukan analisis data dan pemodelan untuk penyelesaian berbagai masalah kompleks pendidikan dan perilaku sains dengan lebih mudah dan cepat [19].

2. METODOLOGI PENELITIAN

Tujuan penelitian ini yaitu membangun suatu model regresi untuk memprediksi tarif *influencer* berdasarkan satuan metrik keterlibatan, jangkauan dan jumlah pengikut. Penelitian ini dilakukan melalui beberapa tahap.



Gambar 1. Metode Penelitian Prediksi Tarif Influencer menggunakan Machine Learning

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

2.1 Studi Pustaka

Mengidentifikasi masalah berdasarkan kajian teoritis dan studi literatur dari penelitian-penelitian dengan cakupan topik: perolehan data *influencer*, penerapan algoritma *machine learning*--regresi linier yang terkait dengan *influencer*, dan evaluasi model regresi. Serta, referensi untuk mempertimbangkan alat bantu *framework* yang akan digunakan untuk membentuk model *machine learning*.

2.2 Pengumpulan Data

Pengumpulan data dilakukan dengan cara mengunjungi platform kolaborasi *influencer* dan melakukan *web scraping* akun *influencer* di media sosial. Pemilihan akun *influencer* memiliki kriteria minimal dengan jumlah 5.000 pengikut dan bergerak di bidang *fashion and beauty*. Penentuan jumlah minimal pengikut ini ditentukan berdasarkan fokus klasifikasi *influencer* pada tingkat *micro-influencer* yang kurang dari 15.000 pengikut [1] dan pemantauan langsung beberapa akun *influencer* yang dianggap memiliki pengaruh pada kisaran 5.000 pengikut. Data yang dikumpulkan mencakup *followers*, *following*, *likes*, *posts*, *views* dan *rates*. Terkadang *influencer* menawarkan *rates* atau tarif yang berbeda untuk UMKM dan korporasi. Tarif *influencer* yang digunakan pada data penelitian ini adalah tarif untuk UMKM.

Media sosial yang saat ini sangat diminati oleh para *influencer* dan lebih berpotensi menjaring atensi audiens terkait *endorsement*, yaitu Instagram dan TikTok. *Dataset* yang telah diperoleh bersumber dari Instagram dan TikTok. Namun, pada penelitian ini hanya menggunakan *dataset* yang bersumber dari media sosial TikTok, karena terdapat keterbatasan waktu untuk memproses data mentah yang tidak terstruktur.

2.3 Pra-pemrosesan Data

Dataset yang telah dihasilkan pada tahap sebelumnya masih perlu dipersiapkan agar dapat digunakan untuk proses pembentukan model regresi. Pra-pemrosesan pada kasus regresi umumnya menangani data yang kosong atau tidak memiliki nilai. Data yang kosong tersebut, jika memungkinkan dapat digantikan dengan nilai rata-rata diantara baris data yang berdekatan. Selain itu, karena peroleh *dataset* melalui proses *web scraping*, perlu dipastikan pada *dataset* tidak mengandung *script* atau karakter-karakter spesial yang tidak diperlukan.

2.4 Pembangunan Model Regresi

Model regresi dibangun dengan algoritma *machine learning*--regresi linier menggunakan alat bantu *framework* Tensorflow untuk memperoleh nilai prediksi tarif *influencer* (variabel dependen), dan satuan metrik keterlibatan (seperti *likes* dan *views*), jangkauan dan jumlah pengikut (variabel independen). Selain itu, akan dilakukan eksperimen untuk memperoleh varian kombinasi beberapa satuan metrik yang paling berpengaruh terhadap performa model. Penelitian ini menggunakan multi-variabel input sebagai variabel independen. Formula (1) menunjukan formula regresi yang digunakan dengan $\hat{Y}$ merupakan prediksi tarif *influencer*, X_1, X_2, X_3, dan X_4 masing-masing mewakili variabel independen yang digunakan yaitu *likes*, *views*, *posts*, dan *followers*, b_1, b_2, b_3, dan b_4 masing-masing mewakili bobot atau estimasi koefisien setiap variabel independen, dan b_0 mewakili konstanta nilai $\hat{Y}$ saat seluruh variabel independen bernilai 0. Algoritma *machine learning* digunakan untuk menemukan nilai b_0, b_1, b_2, b_3, dan b_4 yang memberikan hasil evaluasi terbaik sesuai data X_1, X_2, X_3, dan X_4 yang disiapkan.

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 \quad (1)$$

2.4 Evaluasi Model

Metrik evaluasi yang digunakan untuk mengevaluasi model regresi yang dihasilkan, yaitu: Mean Absolute Error (MAE), Mean Square Error (MSE) dan Root Mean Square Error (RMSE) untuk mengukur performa model dengan melakukan kuantifikasi seberapa baik model menyesuaikan dengan *dataset*. Tentunya, ketiga metrik tersebut telah didukung oleh framework Tensorflow sehingga dapat langsung digunakan saat proses *training* berlangsung. Formula (2), (3), dan (4) menunjukan formula yang digunakan dalam evaluasi model pada penelitian ini.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{real} - y_i^{pred}\right| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{real} - y_i^{pred}\right)^2 \quad (3)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{real} - y_i^{pred}\right)^2} \quad (4)$$

3. HASIL DAN PEMBAHASAN

Hasil penelitian mencakup penentuan *dataset* yang menghasilkan *dataset* sesuai dengan format yang didukung untuk proses pembentukan model, pra-pemrosesan data menghasilkan data yang memiliki konsistensi format dan nilai yang sudah dinormalisasi, pembangunan model regresi menghasilkan lima buah model prediksi yang sesuai

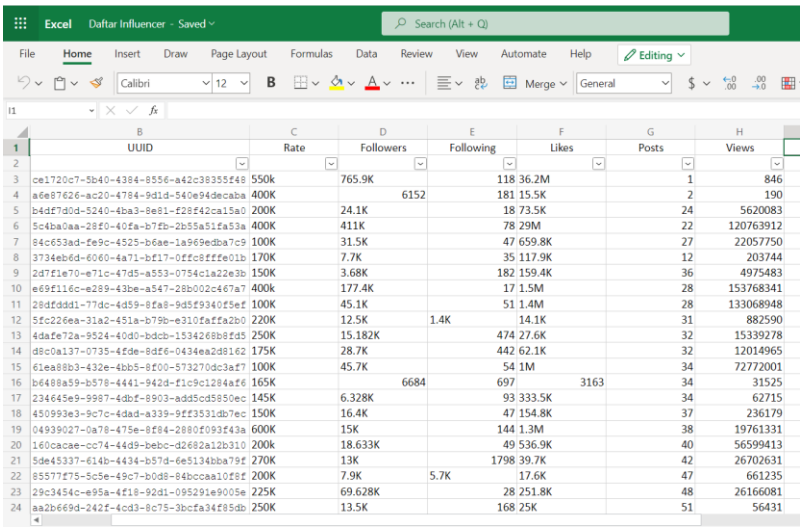
Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

dengan beberapa skenario pembentukan model, dan evaluasi model menghasilkan nilai pengukuran semua model untuk selanjutnya dilakukan ujicoba terhadap model terpilih berdasarkan performa terbaik.

3.1 Penentuan Dataset

Penentuan kriteria jumlah pengikut dan pemilahan atribut dapat berdampak terhadap jumlah *dataset* yang akan digunakan. Pada penelitian ini *dataset* yang digunakan bersumber dari media sosial sebanyak 161 baris. *Dataset* tersebut kemudian disimpan dalam format *file* CSV (*Comma-Separated Values*) untuk memudahkan proses memuat data saat mulai membentuk model *machine learning*.

Setelah *dataset* terkumpul selanjutnya akan dilakukan pemilahan atribut yang sesuai pada *dataset* untuk digunakan pada penelitian. Hal ini perlu dilakukan sebelum memasuki ke tahap pra-pemrosesan data agar proses pra-pemrosesan data lebih efektif. Pemilahan atribut data yang akan digunakan menyesuaikan dengan ketersediaan yang diperoleh pada seluruh atribut. Hasil pemilahan atribut untuk *dataset* yang digunakan, dapat dilihat pada Gambar 2, yaitu: *rate*, *followers*, *following*, *likes*, *posts*, dan *views*.



	UUID	Rate	Followers	Following	Likes	Posts	Views
3	ce1720c7-5b40-4384-8556-a42c38355f48	550k	765.9K	118	36.2M	1	846
4	a6e87626-ac20-4784-9d1d-540e94decaba	400K	6152	181	15.5K	2	190
5	b4df7d0d-5240-4ba3-8e81-f28f42ca15a0	200K	24.1K	18	73.5K	24	5620083
6	5c4ba0aa-28f0-40fa-b7fb-2b55a51fa53a	400K	411K	78	29M	22	120763912
7	84c653ad-fe9c-4525-b6ae-1a969edba7c9	100K	31.5K	47	659.8K	27	22057750
8	3734eb6d-6060-4a71-bf17-0ffc8fffe01b	170K	7.7K	35	117.9K	12	203744
9	2d7f1e70-e71c-47d5-a553-0754c1a22e3b	150K	3.68K	182	159.4K	36	4975483
10	e69f116c-e289-43be-a547-28b002c467a7	400k	177.4K	17	1.5M	28	153768341
11	28dfddd1-77dc-4d59-8fa8-9d5f9340f5ef	100K	45.1K	51	1.4M	28	133068948
12	5fc226ea-31a2-451a-b79b-e310faffa2b0	220K	12.5K	1.4K	14.1K	31	882590
13	4dafe72a-9524-40d0-bdcb-1534268b8fd5	250K	15.182K	474	27.6K	32	15339278
14	d8c0a137-0735-4fde-8df6-0434ea2d8162	175K	28.7K	442	62.1K	32	12014965
15	61ea88b3-432e-4bb5-8f00-573270dc3af7	100K	45.7K	54	1M	34	72772001
16	b6488a59-b578-4441-942d-f1c9c1284af6	165K	6684	697	3163	34	31525
17	234645e9-9987-4dbf-8903-add5cd5850ec	145K	6.328K	93	333.5K	34	62715
18	450993e3-9c7c-4dad-a339-9ff3531db7ec	150K	16.4K	47	154.8K	37	236179
19	04939027-0a78-475e-8f84-2880f093f43a	600K	15K	144	1.3M	38	19761331
20	160cacae-cc74-44d9-bebc-d2682a12b310	200k	18.633K	49	536.9K	40	56599413
21	5de45337-614b-4434-b57d-6e5134bba79f	270K	13K	1798	39.7K	42	26702631
22	85577f75-5c5e-49c7-b0d8-84bccaa10f8f	200K	7.9K	5.7K	17.6K	47	661235
23	29c3454c-e95a-4f18-92d1-095291e9005e	225K	69.628K	28	251.8K	48	26166081
24	aa2b669d-242f-4cd3-8c75-3bcfa34f85db	250K	13.5K	168	25K	51	56431

Gambar 2. Dataset Influencer TikTok sebelum Pra-pemrosesan

3.2 Pra-pemrosesan Data

Salah satu isu yang paling menonjol pada *dataset* yang ditunjukkan pada Gambar 2 yaitu terdapat inkonsistensi satuan pada seluruh atribut. Sebagai contoh ada yang menggunakan satuan "K", yang bermakna ribuan atau terdapat 3 buah nol setelah angka, selanjutnya satuan "M", yang bermakna jutaan atau terdapat 6 buah nol setelah angka. Selain penggunaan satuan, terdapat juga penempatan tanda titik dan koma yang tidak konsisten ditengah-tengah angka.

Setelah pra-pemrosesan data dilakukan, selanjutnya *dataset* yang telah dipersiapkan dapat dirangkum dalam bentuk statistik deskriptif seperti yang ditunjukkan pada Gambar 3. Pada Gambar 3 cakupan statistik yang ditampilkan dapat merangkum karakteristik data, mulai dari tendensi sentral, dispersi, dan bentuk distribusi kumpulan data. Terdapat hal yang juga perlu diperhatikan untuk memastikan bahwa *dataset* tidak mengandung nilai NaN atau *null*. Jika data yang tidak mengandung nilai dibiarkan, tentunya akan berdampak terhadap kualitas model yang dibentuk, karena pada dasarnya kualitas *dataset* yang digunakan untuk *training* memiliki pengaruh kuat terhadap kualitas model.

Pada Gambar 4 terdapat pratinjau *sample dataset* yang akan digunakan. Terdapat atribut atau kolom UUID yang merupakan *identifier* untuk masing-masing baris data. Tentunya, *identifier* tersebut tidak diperlukan pada proses pembentukan atau pembangunan model *machine learning*—regresi, sehingga atribut tersebut dapat dihilangkan. Total atribut yang akan digunakan sebanyak 6 atribut. yang diklasifikasikan menjadi variabel independen/bebas dan dependen/terikat (target), dapat ditunjukkan pada Tabel 1.

	count	mean	std	min	25%	50%	75%	max
Rate	129.0	2.528682e+05	1.692744e+05	50000.0	150000.0	200000.0	300000.0	1.000000e+06
Followers	129.0	2.179221e+05	6.780998e+05	4500.0	13700.0	54200.0	142900.0	6.962800e+06
Following	129.0	1.206698e+03	4.730458e+03	0.0	68.0	168.0	410.0	5.090000e+04
Likes	129.0	3.410518e+06	6.694341e+06	5700.0	111600.0	954300.0	3200000.0	3.620000e+07
Posts	129.0	3.300853e+02	3.327631e+02	1.0	96.0	192.0	468.0	1.919000e+03
Views	129.0	2.465905e+08	4.920380e+08	190.0	7645178.0	80731677.0	235985262.0	2.666094e+09

Gambar 3. Dataset Influencer TikTok setelah Pra-pemrosesan disertai Statistik Deskriptif

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

	UUID	Rate	Followers	Following	Likes	Posts	Views
156	89e8f39c-90e3-411c-ae95-2ad60fd4e8de	100000	45700	54	1000000	34	72772001
157	645584c0-9784-47a3-8561-f2087c61f6e3	140000	87900	128	2500000	643	355975243
158	168a7037-6187-4d61-9dd0-3edcdb8f4a3e	150000	11600	1400	12700	121	457666
159	32b3e104-2c55-4561-a9d5-276d2c41bc20	150000	94200	258	3200000	937	274957936
160	d96e6dbb-5152-4b9f-8fc5-90506ed0b2b3	270000	209200	38	1500000	564	169286390

Gambar 4. Deskriptif Sample Dataset Influencer TikTok setelah Pra-pemrosesan

Tabel 1. Klasifikasi Atribut

No.	Atribut	Klasifikasi
1	Rate	Dependen
2	Followers	Independen
3	Following	Independen
4	Likes	Independen
5	Posts	Independen
6	Views	Independen

Pada Tabel 1 menunjukkan bahwa terdapat hanya satu atribut yang merupakan "target" atau label yang nilainya akan diprediksi, yaitu *rate*. Nilai atribut *rate* tersebut selanjutnya akan dipengaruhi oleh atribut-atribut lainnya, yaitu *followers*, *following*, *likes*, *posts* dan *views*.

Pada tahap pra-pemrosesan, pustaka yang digunakan yaitu Pandas dan Normalizer yang merupakan salah satu fitur penting pada Tensorflow untuk pra-pemrosesan. Hasil penggunaan pandas dapat ditunjukkan pada Gambar 3 dan Gambar 4 untuk mempelajari karakteristik dan memanipulasi *dataset*. Sedangkan Normalizer Tensorflow diperlukan untuk melakukan normalisasi atribut atau fitur yang kontinyu [20]. Sebagai contoh pada Figure 3 pada atribut *likes*, terdapat nilai minimum 5.700 dan maksimum sekitar 3.600.000. Sementara itu, pada atribut *views* terdapat nilai minimum 190 dan maksimum sekitar 2.666.000. Tentu antara kedua atribut tersebut memiliki kesenjangan rentang nilai yang sangat jauh, termasuk juga atribut yang merupakan variabel independen lainnya sehingga perlu dilakukan normalisasi.

Tabel 2. Perbandingan Hasil Pra-pemrosesan dengan Normalizer

No.	Dengan Normalizer				Tanpa Normalizer			
	Likes	Views	Posts	Followers	Likes	Views	Posts	Followers
1	6.800.000	610.072.223	688	283.300	0.51	0.74	1.08	0.10
2	27.400	7.002.125	377	10.000	-0.51	-0.49	0.14	-0.31
3	885.600	175.662.906	789	39.100	-0.38	-0.14	1.38	-0.26

3.3 Pembangunan Model Regresi

Pada tahap ini *framework* Tensorflow mulai digunakan menghasilkan model prediksi *influencer*. Penggunaan Tensorflow dilakukan melalui layanan komputasi awan yaitu Google Colab dengan bahasa pemograman Python. Karena berbasis Python, pustaka yang digunakan yaitu: tensorflow, matplotlib.pyplot, numpy, pandas, seaborn dan drive. Fungsi dari seluruh pustaka yang digunakan, dapat dilihat pada Tabel 3.

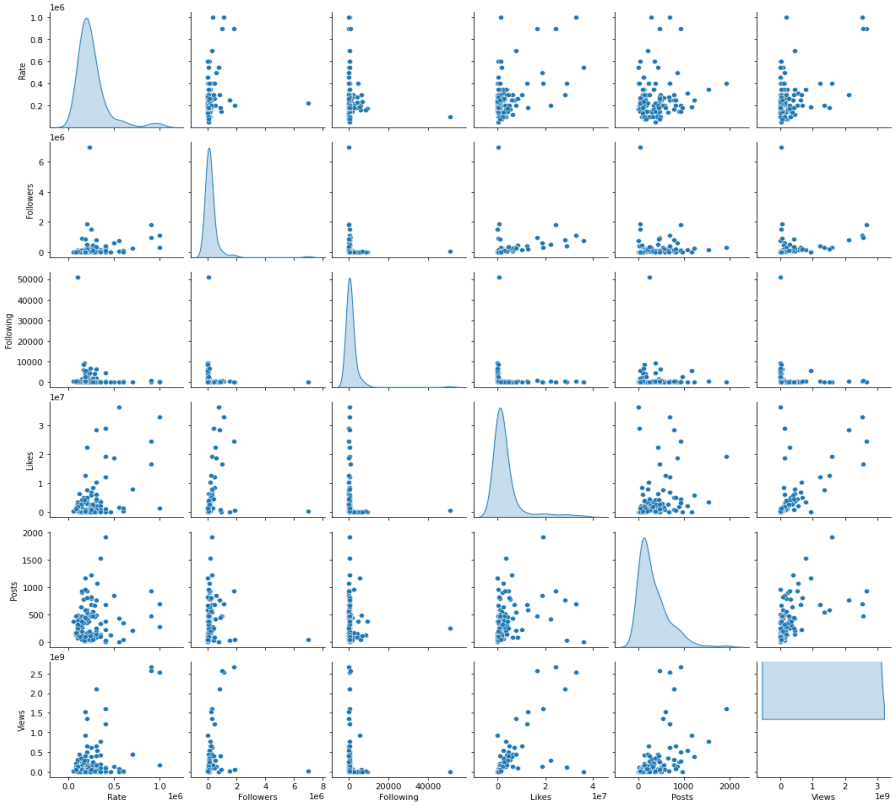
Tabel 3. Pustaka Pendukung Pembentukan Model Machine Learning

No.	Pustaka	Penjelasan
1	tensorflow	Pustaka untuk menggunakan *framework* machine learning. Dalam penelitian ini menggunakan *backend* Keras, disertai juga dengan Normalizer untuk *layer* pra-pemrosesan nilai atribut sehingga dalam rentang -1.00 sampai dengan 1.00.
2	matplotlib.pyplot	Digunakan untuk membuat *plot* atau visualisasi data.
3	numpy	Pustaka yang digunakan untuk operasi yang terkait dengan larik multi-dimensi dan matriks.
4	pandas	Pustaka untuk eksplorasi dan manipulasi dataset, yang dapat digunakan pada tahap pra-pemrosesan.
5	seaborn	Memiliki fungsi utama seperti pustaka matplotlib.pyplot, dapat digunakan untuk eksplorasi data secara visual, seperti memuat visualisasi *pairplot*, *heatmat correlation* antar atribut dalam suatu *dataset.*
6	drive	Diperlukan untuk dapat mengakses *dataset* yang tersimpan pada *cloud storage*, dalam hal ini Google Drive.

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

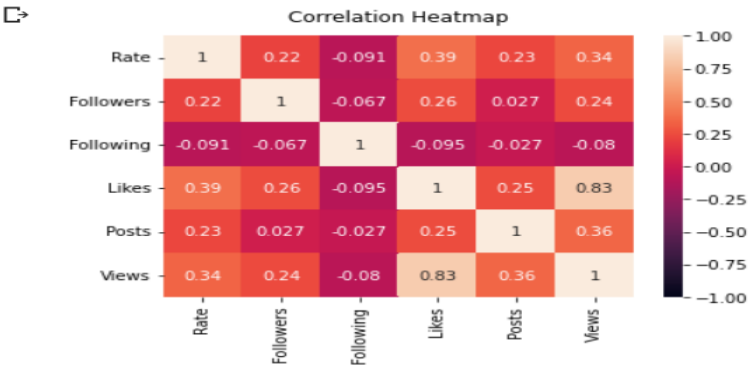
Sebelum memulai pembangunan model, akan lebih efektif jika memeriksa terlebih dahulu karakteristik *dataset* yang dimiliki, salah satunya yaitu memeriksa keterhubungan setiap atribut dalam suatu *dataset* untuk menunjukkan seberapa besar tingkat keterhubungan atau korelasi (*relationship*) antar atribut.

Pada Gambar 5 terlihat sebaran data yang dapat menunjukkan korelasi antar atribut. Pada baris pertama, yaitu atribut *rate* merupakan atribut yang paling disoroti karena merupakan atribut target atau variabel dependen, yang disandingkan terhadap atribut-atribut lainnya. Sekilas terlihat bahwa sebaran data korelasi atribut *rate* hanya terbentuk cukup baik dengan 3 atribut saja, yaitu: *likes*, *posts*, dan *views*. Sementara untuk atribut *followers* dan *following* cenderung tegak lurus secara vertikal. Tentunya, hal tersebut akan membentuk model yang kurang baik, karena kecenderungan model regresi linear membentuk gradien yang memiliki pergerakan kemiringan kearah horizontal-menaik. Selain itu, sebaran data yang padat (*density*) akan lebih membantu menghasilkan model dengan rata-rata kesalahan (*error rate*) lebih rendah.



Gambar 5. Visualisasi Korelasi Antar Atribut pada Dataset menggunakan Pairplot

Lebih lanjut mengenai keterhubungan atau korelasi atribut pada dataset, dapat dikuantifikasi dalam bentuk *correlation heatmap* yang ditunjukkan pada Gambar 6. Kuantifikasi yang ditunjukkan berdasarkan skala -1.00 sampai dengan 1.00. Semakin rendah nilai atau semakin gelap warna yang dihasilkan, semakin rendah tingkat korelasinya.

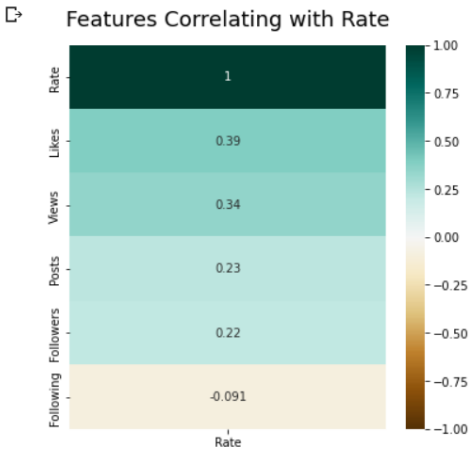


Gambar 6. Visualisasi Korelasi Antar Atribut pada Dataset menggunakan Heatmap Correlation

Dengan melakukan kuantifikasi korelasi antar atribut, menghasilkan wawasan yang lebih terukur. Pada Gambar 6 dan Gambar 7 dapat dilihat bahwa atribut *likes* memiliki korelasi yang paling kuat (0.39) dibandingkan atribut lainnya, diikuti dengan atribut *views* (0.34), atribut *posts* (0.23), dan atribut *followers* (0.22). Untuk atribut

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

following terlihat bahwa atribut tersebut tidak memiliki korelasi terhadap semua atribut karena menghasilkan nilai *minus*, sehingga atribut *following* tidak digunakan untuk membentuk model regresi linier.



Gambar 7. Visualisasi Pemeringkatan Korelasi Antar Atribut pada Dataset

Eksplorasi korelasi atribut yang paling menarik yaitu atribut *followers* merupakan atribut yang paling rendah nilai korelasinya diantara semua atribut yang bernilai positif. Hal tersebut dapat membentuk asumsi bahwa jumlah *followers* pada *influencer* tidak dapat dijadikan secara independen sebagai standar acuan untuk menentukan tarif *influencer*.

Pada tahap pembentukan model regresi linear, berdasarkan Tabel 7, 4 atribut yang akan digunakan karena memiliki nilai korelasi positif terhadap atribut *rate*, yaitu *likes*, *views*, *posts* dan *followers*. Pada Tabel 4 terlihat bahwa *layer* pertama pada setiap skenario model diawali dengan *normalization* dengan parameter (None, 4), yang mana angka 4 sesuai jumlah total atribut yang digunakan. Selanjutnya, proses pembentukkan model memiliki 5 varian skenario yang masing-masing memiliki arsitektur dan konfigurasi atau hyperparameter yang beragam.

Tabel 4. Skenario Pembentukan Model Regresi Linier

No.	Model (Kode)	Arsitektur Layer	Konfigurasi/ Hyperparameter
1	Multi Layer 1 (ML1)	Normalization (None, 4) Dense (None, 64) Dense (None, 1)	Epoch: 1.000 Validation Split: 0.2 Optimizer: Adam Learning Rate: 0.1
2	Multi Layer 2 (ML2)	Normalization (None, 4) Dense (None, 64) Dense (None, 64) Dense (None, 1)	Epoch: 1.000 Validation Split: 0.2 Optimizer: Adam Learning Rate: 0.1
3	Multi Layer 2 Learning Rate 1 (ML2-LR1)	Normalization (None, 4) Dense (None, 64) Dense (None, 64) Dense (None, 1)	Epoch: 1.000 Validation Split: 0.2 Optimizer: Adam Learning Rate: 0.01
4	Multi Layer 2 Learning Rate 2 (ML2-LR2)	Normalization (None, 4) Dense (None, 64) Dense (None, 64) Dense (None, 1)	Epoch: 1.000 Validation Split: 0.2 Optimizer: Adam Learning Rate: 0.001
5	Multi Layer 2 Learning Rate 2 Epoch 2K (ML2-LR2-E2K)	Normalization (None, 4) Dense (None, 64) Dense (None, 64) Dense (None, 1)	Epoch: 2.000 Validation Split: 0.2 Optimizer: Adam Learning Rate: 0.001

3.4 Evaluasi Model

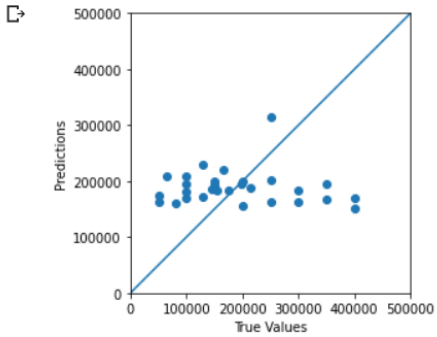
Hasil ujicoba dengan tiga metrik MAE, MSE dan RMSE terhadap 5 varian skenario yang telah ditentukan, dapat dilihat pada Tabel 5. Berdasarkan hasil yang diperoleh, ML2-LR2 memiliki performa yang jauh lebih baik ditandai oleh nilai MAE 145401.484375, MSE 145401.484375, dan RMSE 268742.250. Hal yang menarik dari model ini adalah proses pembentukannya pun diperoleh dengan waktu yang paling rendah karena dipengaruhi oleh nilai parameter epoch sebanyak 1.000 dan learning rate 0.001. Kedua parameter ini sangat berpengaruh terhadap performa model yang dibangun. Terdapat temuan bahwa dengan meningkatkan nilai epoch dua kali lipat pada model ML2-LR-E2K, belum tentu menghasilkan model yang lebih baik.

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

Tabel 5. Hasil Pengujian Varian Skenario Model Regresi Linier

No.	Kode Model	CPU Time (detik)	Wall Time (detik)	MAE	MSE	RMSE
1	ML1	47.5	56.1	170666.328125	1.426823e+11	377733.125
2	ML2	48	56.6	189690.906250	2.424685e+11	492410.875
3	ML2-LR1	49	57.1	175875.625000	1.657227e+11	407090.53125
4	ML2-LR2	47.4	56	145401.484375	7.222241e+10	268742.250
5	ML2-LR2-E2K	94	111	168133.562500	1.351930e+11	367686.0625

Pada Gambar 8, terlihat bahwa hasil prediksi untuk data pengujian menunjukkan sebaran data yang belum banyak mendekati nilai prediksi yang diharapkan. Gambar 8 dapat dimaknai dengan semakin dekat titik data uji dengan representasi garis miring berwarna biru adalah semakin baik. Hanya titik berjumlah minoritas yang dekat dengan garis miring tersebut.



Gambar 8. Visualisasi Perbandingan Nilai Prediksi dan Aktual

Untuk lebih memaknai nilai rata-rata kesalahan yang dihasilkan, Tabel 6 merupakan skenario pengujian lanjutan dengan sejumlah data terhadap model yang memiliki performa terbaik, yaitu: ML2-LR2. Pada kolom Nilai Aktual, Nilai Prediksi dan Selisih Error menggunakan satuan *rate* dalam mata uang Rupiah.

Tabel 1. Hasil Pengujian Model Lanjutan dengan Skenario Operasional

No.	Likes	Views	Posts	Followers	Nilai Aktual	Nilai Prediksi	Selisih Error
1	2.200.000	479.753.251	601	400.000	300.000	162.598,48	137.401,52
2	413.900	15.380.268	53	47.200	65.000	208.124,30	-143.124,3
3	9.700.000	1.254.898.318	688	177.200	250.000	313.866,94	-63.866,94
4	26.900	40.967.690	270	29.700	300.000	183.335,77	116.664,23
5	2.500.000	251.663.648	147	89.600	50.000	162.677,75	-112.677,75
6	55.600	1.328.854	668	7.500	200.000	200.347,69	-347,69
7	63.600	5.578.701	196	10.400	350.000	194.851,84	155.148,16
8	23.800	728.382	254	10.400	215.000	188.610,60	26.389,4
9	5.500.000	553.326.109	680	98.800	150.000	192.508,19	-42.508,19
10	1.000.000	72.772.001	34	45.700	100.000	194.798,86	-94.798,86

Pada Tabel 1 baris ke-6 terlihat ada data yang memiliki selisih *error* yang sangat rendah, minus pada kisaran Rp. -347,69. Tentu hasil prediksi data pada baris ini sangat baik. Selanjutnya, pada baris ke-7 merupakan contoh selisih *error* yang sangat jauh dari yang diharapkan, karena selisih error sampai dengan Rp. 155.148,16. Temuan selisih error yang sangat tinggi ini perlu ditelaah lebih lanjut dan tentunya sesuai dengan nilai MAE yang dihasilkan, berkisar Rp. 145.401,5.

4. KESIMPULAN

Perolehan hasil dari lima varian model yang dibangun belum sepenuhnya mampu mewakili data mayoritas yang diujicoba. Terdapat beberapa faktor penyebab model yang dibentuk masih belum sesuai dengan ketentuan ideal, yaitu: ketersediaan dataset yang belum memadai, dan rentang nilai yang terlalu lebar. Jika diperhatikan lebih lanjut dengan membandingkan dari nilai-nilai atribut *rate* yang ada pada *dataset*, terdapat nilai atribut *rate* terendah yaitu Rp. 50.000. Merujuk pada *error rate* berdasarkan metrik MAE sebesar 145401.484375 (Rp. 145.401.5), model yang dibentuk masih belum bisa digunakan untuk operasional, karena ukuran MAE hampir tiga kali lipat dari nilai *rate* terendah. Setidaknya, selisih *error rate* diharapkan lebih mendekati nilai terendah atribut *rate*. Model yang dibangun belum bisa dijadikan referensi karena terbatasnya jumlah *dataset* yang digunakan untuk proses *training*, hanya 161 baris disertai dengan 4 atribut yang berkorelasi positif, yaitu: likes (0,39), views (0,34), posts (0,23)

Journal of Information System Research (JOSH)
Volume 4, No. 1, Oktober 2022, pp 210−218
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i1.2361

dan followers (0,22). Selain itu, masih ada ruang untuk mengeksplorasi penentuan fitur (*posts* dan *followers* dapat diabaikan), penggunaan *layer* dan konfigurasi *hyperparameter* untuk menghasilkan model dengan performa yang lebih baik. Namun, dalam perspektif penggunaan *dataset* yang sangat terbatas tersebut, model yang dibangun pada penelitian ini cukup berhasil karena ada beberapa contoh data yang dapat diprediksi sangat dekat nilai aktual, salah satunya terdapat nilai prediksi yang memiliki selisih *error* −347,69.

REFERENCES

[1] A. Narassiguin and S. Sargent, "Data Science for Influencer Marketing : feature processing and quantitative analysis," *CoRR*, vol. abs/1906.05911, Oct. 2019, [Online]. Available: http://arxiv.org/abs/1906.05911
[2] D. Vrontis, A. Makrides, M. Christofi, and A. Thrassou, "Social media influencer marketing: A systematic review, integrative framework and future research agenda," *Int J Consum Stud*, vol. 45, no. 4, pp. 617–644, 2021.
[3] J. Tabellion and F.-R. Esch, "Influencer Marketing and its Impact on the Advertised Brand," *Advances in Advertising Research X: Multiple Touchpoints in Brand Communication*. Springer Fachmedien Wiesbaden, pp. 29–41, 2019. doi: 10.1007/978-3-658-24878-9_3.
[4] A. Ampountolas, G. Shaw, and S. James, "The role of social media as a distribution channel for promoting pricing strategies," *Journal of Hospitality and Tourism Insights*, 2019.
[5] K. M. Klassen, E. S. Borleis, L. Brennan, M. Reid, T. A. McCaffrey, and M. S. C. Lim, "What people 'like': Analysis of social media strategies used by food industry brands, lifestyle brands, and health promotion organizations on Facebook and Instagram," *J Med Internet Res*, vol. 20, no. 6, p. e10227, 2018.
[6] M. Chui *et al*., "The social economy: Unlocking value and productivity through social technologies." Oct. 2012. [Online]. Available: https://www.mckinsey.com/~/media/mckinsey/industries/technology%20media%20and%20telecommunications/high%20tech/our%20insights/the%20social%20economy/mgi_the_social_economy_full_report.pdf
[7] TikTok, "Understanding your analytics," *TikTok Creator Portal*. Oct. 2022. [Online]. Available: https://www.tiktok.com/creators/creator-portal/en-us/tiktok-content-strategy/understanding-your-analytics/
[8] K. Peters, Y. Chen, A. M. Kaplan, B. Ognibeni, and K. Pauwels, "Social media metrics—A framework and guidelines for managing social media," *Journal of interactive marketing*, vol. 27, no. 4, pp. 281–298, 2013.
[9] D. Bakker, "Conceptualising influencer marketing," *Journal of emerging trends in marketing and management*, vol. 1, no. 1, pp. 79–87, Oct. 2018, [Online]. Available: https://research.hanze.nl/en/publications/conceptualising-influencer-marketing
[10] C. Egger, "Identifying Key Opinion Leaders in Social Networks - An Approach to use Instagram Data to Rate and Identify Key Opinion Leader for a Specific Business Field." 2016. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:hbz:832-epub4-8450
[11] A. Handayani, "Influencer Rate Card: How to Get and Discover," Jan. 19, 2022. https://lemon.co.id/en/articles/influencer-rate-card-how-to-get-and-discover/ (accessed Oct. 20, 2022).
[12] S. Khogasteh and E. Wiorek, "Predicting Influencer Actual Reach Using Linear Regression." Oct. 2021. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1583319/FULLTEXT01.pdf
[13] C. Primasiwi, M. I. Irawan, and R. Ambarwati, "Key Performance Indicators for Influencer Marketing on Instagram," in *Proceedings of the 2nd International Conference on Business and Management of Technology (ICONBMT 2020)*, 2021, pp. 154–163. doi: https://doi.org/10.2991/aebmr.k.210510.027.
[14] R. S. Nugroho, "Bagaimana Cara Mengetahui Rate Card Influencer? Simak Penjelasannya!," *IDXChannel.com*. Oct. 2022. [Online]. Available: https://www.idxchannel.com/milenomic/bagaimana-cara-mengetahui-rate-card-influencer-simak-penjelasannya
[15] S. Wies, A. Bleier, and A. Edeling, "EXPRESS: Finding Goldilocks Influencers: How Follower Count Drives Social Media Engagement," *J Mark*, p. 00222429221125131, 2022.
[16] BBC News Indonesia, "Pemerintah Indonesia bayar influencer Rp90 miliar untuk sosialiasi kebijakan, 'buang duit yang efektif?,'" Aug. 21, 2020. https://www.bbc.com/indonesia/indonesia-53846128 (accessed Oct. 20, 2022).
[17] A. Arora, S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi, "Measuring social media influencer index- insights from facebook, Twitter and Instagram," *Journal of Retailing and Consumer Services*, vol. 49, pp. 86–101, Oct. 2019, doi: https://doi.org/10.1016/j.jretconser.2019.03.012.
[18] TensorFlow, "Basic regression: Predict fuel efficiency," *TensorFlow Tutorials*, 2022. https://www.tensorflow.org/tutorials/keras/regression (accessed Sep. 11, 2022).
[19] B. Pang, E. Nijkamp, and Y. N. Wu, "Deep Learning With TensorFlow: A Review," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 2, pp. 227–248, 2020, doi: 10.3102/1076998619872761.
[20] TensorFlow, "tf.keras.layers.Normalization," *TensorFlow API Docs - Python*, 2022. https://www.tensorflow.org/api_docs/python/tf/keras/layers/Normalization (accessed Sep. 11, 2022).