ISSN: 2477-5126
e-ISSN: 2548-9356

Klasifikasi Judul Berita *Clickbait* menggunakan RNN-LSTM

Widi Afandi[1], Satria Nur Saputro[2], Andini Mulia Kusumaningrum[3], Hikari Ardiansyah[4], Muhammad Hilmi Kafabi[5], Sudianto Sudianto[6*)]
[1,2,3,4,5,6]Jurusan Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto, Purwokerto
[1,2,3,4,5,6]Jl. D.I Panjaitan No. 128, Kota Purwokerto, 53147, Indonesia
email: [1]19102127@ittelkom-pwt.ac.id, [2]19102296@ittelkom-pwt.ac.id, [3]19102027@ittelkom-pwt.ac.id, [4]19102105@ittelkom-pwt.ac.id, [5]19102115@ittelkom-pwt.ac.id, [6]sudianto@ittelkom-pwt.ac.id

Abstract *− **Online news has become a lifestyle in today's era. Online news has changed the old way of people searching for and obtaining news information, which used to be from print media, and now switch to electronic media. As news is quickly and easily shared, many media producers take advantage of this opportunity by uploading news online on multiple platforms to increase traffic and news ratings. However, the information conveyed is only limited to attracting readers' attention by exaggerating the headlines in the news. Thus, news titles often do not match the content of the news. This phenomenon is commonly known as "clickbait" among the public. Therefore, the purpose of this study is to classify news with clickbait news titles and non-clickbait news titles. In addition, the classification method used in this study is the Deep Learning method using the RNN-LSTM architecture. The classification results show that the best accuracy calculation is 79% on the training data with an epoch value of 27, and the accuracy value on the validation data is 77%.***

Keywords **– NLP, RNN-LSTM, Clickbait, Classification**

Abstrak *− **Berita online telah menjadi gaya hidup di era saat ini. Berita online telah merubah cara lama masyarakat dalam mencari dan memperoleh informasi berita yang dulunya dari media cetak sekarang beralih ke media elektronik. Seiring dengan pesat dan mudahnya berita dibagikan, banyak produsen media memanfaatkan kesempatan ini dengan mengunggah berita online di beberapa platform untuk meningkatkan lalu lintas dan peringkat berita. Namun, informasi yang disampaikan hanya sebatas menarik perhatian pembaca dengan melebih-lebihkan headline pada berita. Dengan begitu, judul berita seringkali tidak sesuai dengan isi berita. Fenomena ini biasa dikenal dengan istilah "clickbait" di kalangan masyarakat. Oleh karena itu, tujuan dari penelitian ini mengklasifikasikan berita dengan judul berita clickbait dan judul berita non-clickbait. Selain itu, metode klasifikasi yang digunakan pada penelitian ini yaitu metode Deep Learning menggunakan arsitektur RNN-LSTM. Hasil klasifikasi menunjukkan perhitungan akurasi terbaik sebesar 79% pada data latih dengan nilai epoch 27 dan nilai akurasi pada data validasi sebesar 77%.***

Kata Kunci **– NLP, RNN-LSTM, Clickbait, Classification**

I. PENDAHULUAN

Seiring dengan meningkatnya laju penggunaan internet di masyarakat, membuat berita *online* kian menjadi pilihan utama dalam mengakses informasi. Berita *online* sangat mudah diakses oleh pengguna. Berita *online* mudah diperoleh dan disebarkan luaskan diberbagai perangkat. Namun, dengan mudahnya berita *online* diakses dan disebar luaskan, hal ini justru dimanfaatkan oleh beberapa produsen berita yang menyesatkan pembaca. Fenomena menyesatkan pembaca ini, terlihat antara judul berita dan konten isi yang disampaikan tidak relevan. Sehingga fenomena inilah yang dimanfaatkan oleh produsen untuk meraup keuntungan agar rating maupun jumlah pembaca meningkat saat melakukan klik pada berita. Fenomena ini disebut *clickbait*.

Konsep *clickbait* mulai diformalkan sebagai sesuatu untuk menarik pembaca mengeklik situs mereka berdasarkan potongan informasi yang menyertainya. Terutama ketika *link* tersebut mengarah ke konten yang bernilai atau menarik. *Clickbait* adalah tindakan yang disengaja untuk memberikan janji yang berlebihan atau dengan sengaja memberikan gambaran yang salah dalam judul, gambar, atau beberapa kombinasi di media sosial yang dapat memberikan harapan kepada pembaca pada saat membaca sebuah cerita di halaman website. Hal itu sengaja dibuat dan akibatnya memanfaatkan kesenjangan informasi [1].

Dari permasalahan yang ada, sebagai langkah pencegahan berita *clickbait*, maka pembaca diharapkan dapat melakukan klasifikasi berita secara dini, terutama berita *clickbait* atau bukan. Sehingga pembaca lebih cepat mengetahui konten dari berita *online* yang beredar. Pada penelitian sebelumnya, klasifikasi berita *clickbait* menggunakan LSTM dan CNN. Algoritme LSTM mendapatkan akurasi 84% dan algoritme CNN mendapatkan akurasi 88% [2], [3]. Selain itu, klasifikasi tek dengan menggunakan SAT-LSTM, penelitian ini mendapatkan akurasi 89,2% [4]. Namun penelitian sebelumnya belum menggunakan berita berkonten Bahasa Indonesia. Sehingga penlitian ini, melakukan klasifikasi berita *online* menggunakan RNN-LSTM pada berita di Indonesia.

*) **penulis korespondensi**: sudianto
Email: sudianto@ittelkom-pwt.ac.id

Oleh karena itu, tujuan penelitian ini melakukan klasifikasi yang dapat mendeteksi berita *clickbait*. Arsitektur yang akan digunakan untuk penelitian kali ini menggunakan arsitektur RNN-LSTM. LSTM atau Long Short-Term Memory merupakan model pembaruan dari model RNN yang digunakan untuk mengelola data yang bersifat sekuensial dengan menyimpan hasil informasi sebelumnya [5]. RNN adalah arsitektur jaringan saraf tiruan yang dipemrosesan secara berulang untuk memproses data masukan bersifat sekuensial [6].

II. PENELITIAN YANG TERKAIT

Pada penelitian terdahulu, terkait analisis terkait klasifikasi *clickbait* yang berasal dari *headline* berita lokal maupun dari pangkalan data. Berdasarkan penelitian terdahulu, peneliti menggunakan berbagai macam model Deep Learning untuk melakukan analisis sentimen terhadap *clickbait*. Seperti data yang berasal dari pangkalan data penelitian ilmiah ScienceDirect dan IEEE Xplore. pada penelitian ini menggunakan metode LSTM dan CNN. Metode LSTM mendapatkan akurasi 84% dan metode CNN mendapatkan akurasi 88% [2].

Pada referensi lain yang mendeteksi ujaran kebencian dari data Twitter menggunakan metode LSTM, mendapatkan acc yang sangat baik yaitu 86% [7]. Referensi selanjutnya yaitu melakukan klasifikasi tek dengan menggunakan SAT-LSTM, penelitian ini mendapatkan akurasi 89,2% [4]. Selanjutnya pada referensi yang melakukan *classification text* dengan metode LSTM, di dapatkan rata-tata akurasi 93.5% [8].

RNN atau Recurrent Neural Network bekerja seperti jaringan syaraf yang melakukan pemrosesan berulang-ulang di panggil untuk memproses input. RNN sendiri telah mengalami kemajuan cukup pesat dan telah merevolusi bidang-bidang seperti NLP [9], [10]. Sedangkan LSTM merupakan pengembangan dari model RNN, karena pada model RNN mempunyai permasalahan *vanishing gradient* [11]. Berdasarkan penelitian-penelitian sebelumnya, peneliti akan mengajukan model analisis sentimen dengan Deep Learning menggunakan metode LSTM dan RNN. Dengan melihat penelitian terdahulu, metode LSTM dipilih untuk melakukan klasiifkasi terhadap *clickbait*.

III. METODE PENELITIAN

Diagram alir dari penelitian ini dilakukan sesuai dengan alur yang ditunjukkan pada Gbr 1.

A. Persiapan Perencanaan

Menyiapkan alur dari project yang dibuat, mulai dari penentuan permasalahan, studi literatur , solusi algoritma penyelesaian masalah, penyederhanaan *clean code algorithm*.

B. Pengumpulan Dataset

Dataset berasal dari jurnal yang berisi 15000 yang telah dianotasi antara *clickbait* dan non *clickbait*, data non *clickbait* sberjumlah 8700 dan *clickbait* berjumlah 6300 [12].

C. Preprocessing Dataset

Menerapkan pembersihan dataset agar bersih dan pengoptimalan saat men-training data, tahapan-tahapan preprocessing data: (1) Mengubah huruf menjadi kecil; (2) Pembersihan simbol dan angka.

Gbr. 1 diagram alir.

D. Pembagian Data Train dan Test

Pembagian dilakukan dengan membagi 80% data *train* dan 20% data test, pembagian dilakukan secara *random* agar mendapatkan akurasi optimal [13].

E. Tokenizer

Melakukan *tokenizer* pada dataset baik pada *train* dan *test*, hal ini merupakan cara mengubah setiap kata menjadi angka yang unik sehingga model dapat memahami inputan data [14].

F. Padding dan Sequences

Dari hasil tokenizer dilakukan pada *sequences* agar setiap kalimat memiliki panjang data yang sama.

G. Pemodelan

Tahap ini melakukan arsitektur pemodelan untuk men-*training* dataset agar model dapat mengklasifikasi *headline clickbait* atau *non clickbait*.

H. Evaluasi Model

Melakukan evaluasi apakah model sudah optimal untuk digunakan sebagai peng-klasifikasian *headline* baru.

I. Visualisasi Model

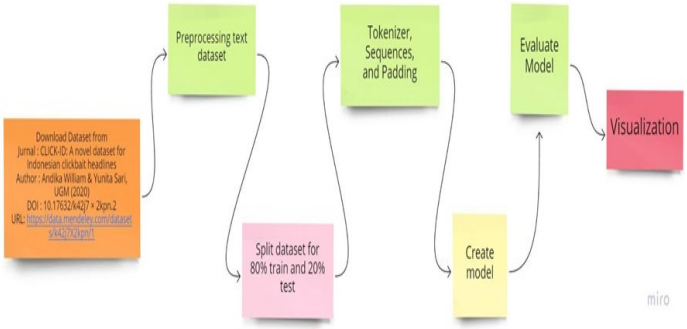
Melakukan visualisasi hasil akurasi , validasi akurasi, loss, dan validasi loss, dalam bentuk gambar sehingga mudah dipahami oleh peneliti.

ISSN: 2477-5126
e-ISSN: 2548-9356

J. Visualisasi Word Cloud

Melakukan visualisasi kata-kata yang banyak muncul pada *headline* di dalam dataset.

Langkah-langkah dalam penelitian terhadap pembuatan sistem pada gambar di bawah.



Gbr. 2 storyboard.

K. Arsitektur Model

Pada penelitian ini menggunakan arsitektur LSTM seperti gambar dibawah.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, None, 4)           60000

 dropout_2 (Dropout)         (None, None, 4)           0

 lstm_1 (LSTM)               (None, 16)                1344

 dropout_3 (Dropout)         (None, 16)                0

 dense_2 (Dense)             (None, 8)                 136

 dense_3 (Dense)             (None, 1)                 9

=================================================================
Total params: 61,489
Trainable params: 61,489
Non-trainable params: 0
_________________________________________________________________
```

Gbr. 3 Contoh arsitektur model.

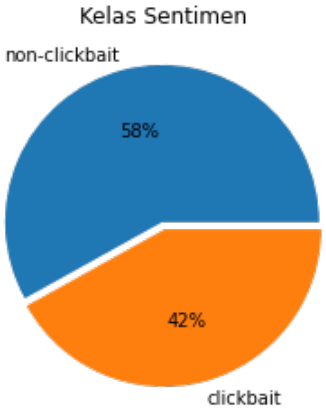
IV. HASIL DAN PEMBAHASAN

Pada penelitian ini, digunakan dataset berupa kumpulan *headline* dari beberapa media *online* yang terdapat *clickbait* atau tidak yang didapat dari penelitian sebelumnya.

	title	label	label_score
0	Masuk Radar Pilwalkot Medan, Menantu Jokowi Be...	non-clickbait	0
1	Malaysia Sudutkan RI: Isu Kabut Asap hingga In...	non-clickbait	0
2	Viral! Driver Ojol di Bekasi Antar Pesanan Mak...	clickbait	1
3	Kemensos Salurkan Rp 7,3 M bagi Korban Kerusuh...	non-clickbait	0
4	Terkait Mayat Bayi Mengenaskan di Tangerang, S...	non-clickbait	0

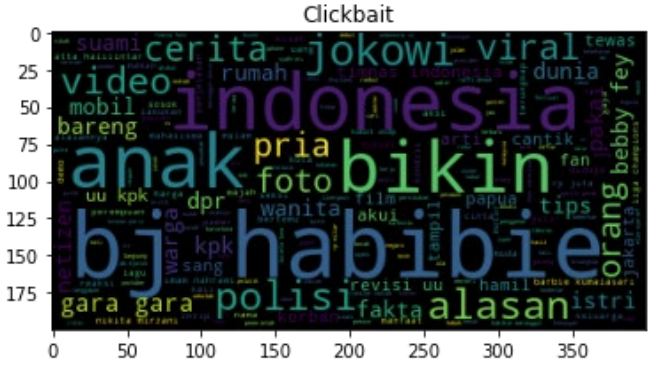
Gbr. 4 Contoh data yang digunakan.

Proporsi data dengan label *non-clickbait* dan *clickbait* ditampilkan dalam bentuk grafik seperti Gbr 4. Dapat dilihat bahwa proporsi jumlah data untuk label *non-clickbait* lebih banyak dibandingkan label *clickbait*.

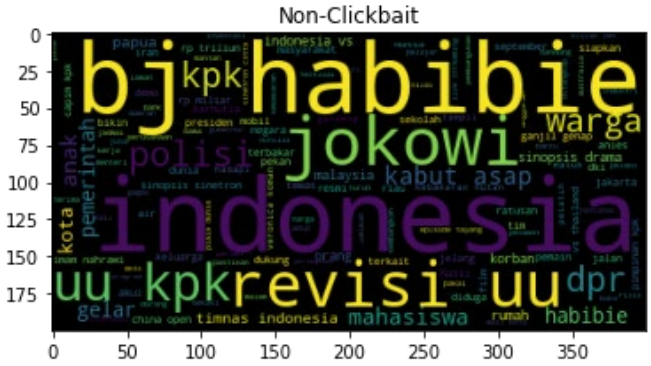


Gbr. 5 Proporsi label data non-clickbait dan clickbait.

Ditampilkan juga Wordcloud dari masing-masing data berlabel non-clickbait dan clickbait pada dataset kumpulan headline dari beberapa media online pada gambar 5 dan gambar 6.



Gbr. 6 Wordcloud data non-clickbait.



Gbr. 7 Wordcloud data clickbait.

Dilakukan proses *text-processing* untuk mengubah data yang tidak terstruktur menjadi lebih terstruktur agar dapat diolah oleh model yang juga akan mempengaruhi tingginya akurasi. Tahapan dari *text-processing* yaitu *data cleansing* dan *tokenization* yang dilakukan secara berurutan.

Pada proses *data cleansing* akan dilakukan perubahan data pada *headline* dimana setiap huruf besar diubah ke huruf kecil, menghapus tanda baca dan angka pada *dataset*. Tahap selanjutnya dilakukan *tokenization* di mana proses ini akan memisahkan kata pada kalimat kemudian diubah dalam bentuk numerik. Hasil dari *tokenization* kemudian akan

ISSN: 2477-5126
e-ISSN: 2548-9356

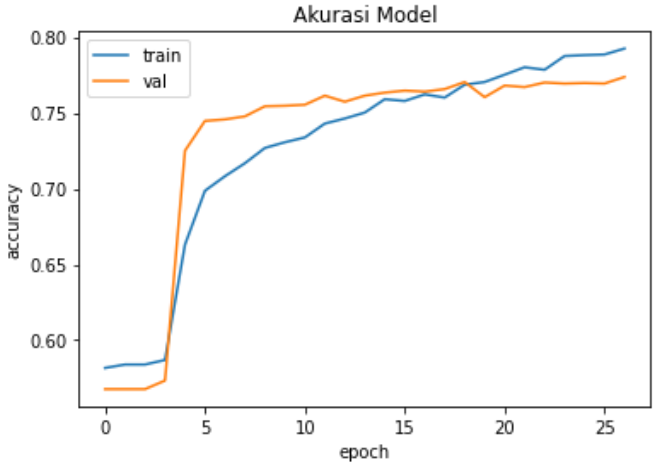
dimasukkan ke dalam larik yang disebut *sequence*. Tahap terakhir adalah *padding* di mana *padding* digunakan untuk menyamakan panjang dari setiap *sequence*.

	title	title_cleaned
0	Masuk Radar Pilwalkot Medan, Menantu Jokowi Be...	masuk radar pilwalkot medan menantu jokowi ber...
1	Malaysia Sudutkan RI: Isu Kabut Asap hingga In...	malaysia sudutkan ri isu kabut asap hingga inv...
2	Viral! Driver Ojol di Bekasi Antar Pesanan Mak...	viral driver ojol di bekasi antar pesanan maka...
3	Kemensos Salurkan Rp 7,3 M bagi Korban Kerusuh...	kemensos salurkan rp m bagi korban kerusuhan s...
4	Terkait Mayat Bayi Mengenaskan di Tangerang, S...	terkait mayat bayi mengenaskan di tangerang se...

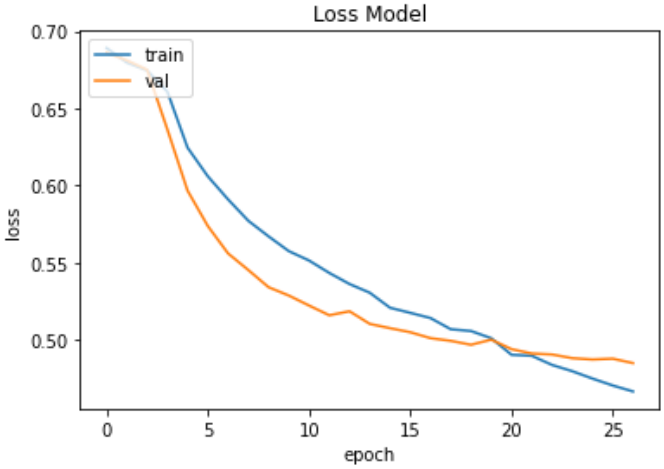
Gbr. 8 Contoh data setelah proses data cleansing.

Tahapan selanjutnya adalah membagi data menjadi data *training* dan data *testing* dengan proporsi 80:20. Data *training* digunakan untuk melatih model sedangkan data *testing* digunakan untuk mengevaluasi model apakah memiliki akurasi yang baik *pada* proses prediksi.

Tahap pemodelan digunakan model RNN-LSTM di mana model ini termasuk dalam model Deep Learning. Deep Learning dapat digunakan untuk klasifikasi data text maupun citra [15], [16]. Digunakan juga *layer dropout* yang berguna untuk menonaktifkan *perceptron* yang tidak aktif dan berfungsi mencegah *overfitting*.



Gbr. 9 Contoh kurasi dari epoch 0-100 pada data training dan data testing.



Gbr. 10 Contoh loss dari epoch 0-100 pada data training dan data testing

Dari Gbr 9 dan Gbr 10, menunjukkan bahwa model memiliki akurasi yang seimbang pada data *training* dan data *testing* di *epoch* 27 dengan sebesar 79% pada data *training* dan 77% pada data *testing*. Pada *loss* mengalami penurunan baik dari data *training* maupun *testing*. Penelitian ini digunakan *callbacks* dimana ketika hasil akurasi data *training* dan *testing* sudah melebihi target yaitu 77% maka *epoch* akan berhenti. Hasil akurasi penelitian ini sedikit lebih tinggi dari pada penelitian sebelumnya yang menggunakan arsitektur CNN dengan melakukan pemrosesan teks yang sama pada dataset mendapatkan dengan hasil akurasi rata-rata 75%. Dilakukan tes pada data dunia nyata dengan judul *headline* "Terbongkar, Artis Lutfi Aminuddin Ternyata Suka Makan Tempe" dan model mengidentifikasi bahwa 85% *headline* tersebut adalah *clickbait.*

V. KESIMPULAN

Hasil dari akurasi klasifikasi tertinggi menunjukan bahwa model memiliki akurasi yang seimbang pada data *training* dan data *testing* di *epoch* 27 dengan sebesar 79% pada data *training* dan 77% pada data *testing*. Pada *loss* mengalami penurunan baik dari data *training* maupun *testing*. Hal ini dipengaruhi data *Preprocessing*, arsitektur model, dan keseimbangan data *non clickbait* dan *clickbait*.

UCAPAN TERIMA KASIH

Penyusun mengucapkan terima kasih kepada Prodi Teknik Informatika yang telah mensupport penelitian ini. Serta semua pihak yang telah berpartisipasi dalam proses penyelesaian penelitian ini.

DAFTAR PUSTAKA

[1] V. Kumar, D. Khattar, S. Gairola, Y. Kumar Lal, and V. Varma, "Identifying clickbait: A multi-strategy approach using neural networks," *41st Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2018*, pp. 1225–1228, 2018, doi: 10.1145/3209978.3210144.

[2] R. Yunanto, A. P. Purfini, and A. Prabuwisesa, "Survei Literatur : Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning," *J. Manaj. Inform.*, vol. xx, no. 2, pp. 118–130, 2021, doi: 10.34010/jamika.v11i2.5362.

[3] S. Kaur, P. Kumar, and P. Kumaraguru, "Detecting clickbaits using two-phase hybrid CNN-LSTM biterm model," *Expert Syst. Appl.*, vol. 151, p. 113350, 2020, doi: 10.1016/j.eswa.2020.113350.

[4] R. Jing, "A Self-attention Based LSTM Network for Text Classification," *J. Phys. Conf. Ser.*, vol. 1207, no. 1, 2019, doi: 10.1088/1742-6596/1207/1/012008.

[5] F. Fauzi, F. F. Abdulloh, and I. R. Pambudi, "Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19," *Comput. Sci. Res. Its Dev. J.*, vol. 13, no. 3, pp. 141–148, 2021.

[6] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance," *Expert Syst. Appl.*, vol. 161, p. 113725, 2020, doi: 10.1016/j.eswa.2020.113725.

[7] A. Bisht, A. Singh, H. S. Bhadauria, J. Virmani, and Kriti, "Detection of hate speech and offensive language in twitter data using LSTM model," *Adv. Intell. Syst. Comput.*, vol. 1124, pp. 243–264, 2020, doi: 10.1007/978-981-15-2740-1_17.

[8] Z. H. Kilimci and S. Akyokus, "The Evaluation of Word Embedding Models and Deep Learning Algorithms for Turkish Text Classification," *UBMK 2019 - Proceedings, 4th Int. Conf. Comput. Sci. Eng.*, pp. 548–553, 2019, doi: 10.1109/UBMK.2019.8907027.

[9] P. Liu, X. Qiu, and H. Xuanjing, "Recurrent neural network for text

classification with multi-task learning," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2016-Janua, pp. 2873–2879, 2016.

[10] S. Sudianto, A. D. Sripamuji, I. Ramadhanti, R. R. Amalia, J. Saputra, and B. Prihatnowo, "Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasisifikasi Topik Berita," vol. 11, no. 2, pp. 84–91, 2022.

[11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[12] A. William and Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," *Data Br.*, vol. 32, p. 106231, 2020, doi: 10.1016/j.dib.2020.106231.

[13] S. Sudianto, P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "Comparison Of Random Forest And Support Vector Machine Methods On Twitter Sentiment Analysis (Case Study : Internet Selebgram Rachel Vennya Escape From Quarantine) Perbandingan Metode Random Forest Dan Support Vector Machine Pada Analisis Sentimen Twitt," *Jutif*, vol. 3, no. 1, pp. 141–145, 2022.

[14] Y. Puspitarani and Y. Syukriyah, "Pemanfaatan Optical Character Recognition Dan Text Feature Extraction Untuk Membangun Basisdata Pengaduan Tenaga Kerja," vol. 1, no. 3, pp. 704–710, 2020.

[15] Sudianto, Y. Herdiyeni, A. Haristu, and M. Hardhienata, "Chilli quality classification using deep learning," 2020. doi: 10.1109/ICOSICA49951.2020.9243176.

[16] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 9, pp. 103–114, 2018, doi: 10.14257/ijgdc.2018.11.9.09.