

**SISTEM AGREGASI BERITA DENGAN KLASIFIKASI, *CLUSTERING*
DAN PERINGKASAN OTOMATIS**

SKRIPSI

Diajukan untuk Memenuhi Sebagian dari
Syarat Memperoleh Gelar Sarjana Komputer
Program Studi Ilmu Komputer



Disusun Oleh:
Ihsan Ghazi Zulfikar
2103303

**PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
BANDUNG
2025**

**SISTEM AGREGASI BERITA DENGAN KLASIFIKASI, *CLUSTERING*
DAN PERINGKASAN OTOMATIS**

Oleh:
Ihsan Ghazi Zulfikar
2103303

diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer
pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

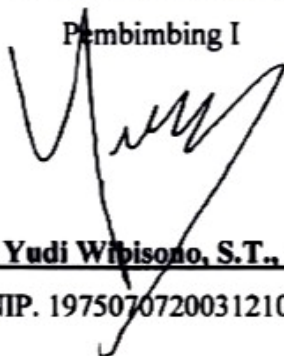
© Ihsan Ghazi Zulfikar
Universitas Pendidikan Indonesia
Agustus 2025

Hak Cipta Dilindungi Undang-Undang
Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian, dengan dicetak
ulang, di foto kopi, atau cara lainnya tanpa izin dari penulis

**SISTEM AGREGASI BERITA DENGAN KLASIFIKASI, CLUSTERING DAN
PERINGKASAN OTOMATIS**

Disetujui dan disahkan oleh:

Pembimbing I



Dr. Yudi Wibisono, S.T., M.T.

NIP. 197507072003121003

Pembimbing II



Dr. Asep Wahyudin, S.Kom., M.T.

NIP. 197112232006041001

Mengetahui,

Ketua Program Studi Ilmu Komputer



Dr. Muhamad Nursalman, M.T.

NIP. 197909292006041002

PERNYATAAN BEBAS PLAGIARISME

Saya yang bertanda tangan di bawah ini:

Nama : Ihsan Ghazi Zulfikar
NIM : 2103303
Program Studi : Ilmu Komputer
Judul Karya : Sistem Agregasi Berita dengan Klasifikasi, *Clustering*
dan Peringkasan Otomatis

Dengan ini menyatakan bahwa karya tulis ini merupakan hasil kerja saya sendiri. Saya menjamin bahwa seluruh isi karya ini, baik sebagian maupun keseluruhan, bukan merupakan plagiarisme dari karya orang lain, kecuali pada bagian yang telah dinyatakan dan disebutkan sumbernya dengan jelas.

Jika di kemudian hari ditemukan pelanggaran terhadap etika akademik atau unsur plagiarisme, saya bersedia menerima sanksi sesuai peraturan yang berlaku di Universitas Pendidikan Indonesia.

Bandung, Agustus 2025

Yang membuat pernyataan,



Ihsan Ghazi Zulfikar

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas rahmat, taufik, dan karunia-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul “Sistem Agregasi Berita dengan Klasifikasi, *Clustering* dan Peringkasan Otomatis” dengan baik. Penyusunan skripsi ini ditujukan untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana Komputer pada jenjang S1 Program Studi Ilmu Komputer, Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia. Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang bersifat membangun di masa yang akan datang.

Bandung, Agustus 2025

Penulis,



Ihsan Ghazi Zulfikar

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada berbagai pihak yang telah memberikan dukungan, bantuan, dan doa selama proses penyusunan skripsi ini, terutama kepada:

1. Allah SWT yang telah memberikan kesehatan, kesempatan, dan kelancaran dalam menyelesaikan skripsi ini.
2. Kedua orang tua, Ati Rohaeti dan Dudi Atwendi, yang selalu memberikan doa, motivasi, dan dukungan moril maupun materiil tanpa henti.
3. Bapak Dr. Yudi Wibisono, S.T., M.T. selaku pembimbing I dan Bapak Dr. Asep Wahyudin, S.Kom., M.T. selaku pembimbing II yang telah memberikan arahan, masukan, serta bimbingan selama proses penelitian dan penulisan skripsi ini.
4. Seluruh dosen dan staf Program Studi Ilmu Komputer, FPMIPA UPI yang telah membekali penulis dengan ilmu dan pengalaman selama masa studi.
5. Rekan-rekan mahasiswa kelas C2 yang selalu memberi semangat dan saling membantu.
6. Semua pihak yang tidak dapat disebutkan satu per satu, yang telah membantu secara langsung maupun tidak langsung dalam penyelesaian skripsi ini.

Bandung, Agustus 2025

Penulis,



Ihsan Khozi Zulfikar

SISTEM AGREGASI BERITA DENGAN KLASIFIKASI, *CLUSTERING* DAN PERINGKASAN OTOMATIS

Oleh

Ihsan Ghazi Zulfikar - ihsanghozizulfikar@upi.edu

2103303

ABSTRAK

Berita *online* telah menjadi sumber utama informasi bagi masyarakat Indonesia, namun banyaknya berita online sering kali membuat pembaca kesulitan menemukan berita yang sesuai minat. Penelitian ini bertujuan mengembangkan sistem agregasi berita berbasis web yang dapat melakukan klasifikasi, clustering dan peringkasan secara otomatis terhadap artikel berita *online* berbahasa Indonesia menggunakan kecerdasan buatan. Sistem ini dibangun menggunakan pendekatan pengembangan ADDIE dan implementasi perangkat lunak dengan metode *Waterfall*. Proses dimulai dari scraping artikel dari Kompas dan TribunNews pada Juni 2025, menghasilkan 37.187 artikel. Model klasifikasi yang digunakan yaitu BLSTM-2DCNN yang dilatih menggunakan dataset Indosum mencapai akurasi 86% dan F1-score 0.85. *Clustering* dilakukan menggunakan k-means berdasarkan kategori dan rentang waktu lima hari. Evaluasi menggunakan CH-Index menunjukkan nilai rata-rata 80.87. Untuk peringkasan, digunakan model BART yang di-*fine-tuning*, dengan ROUGE-1, ROUGE-2, dan ROUGE-L masing masing meningkat menjadi 0.6434, 0.5510, dan 0.6077. Sistem yang dibangun menyajikan artikel yang telah diklasifikasikan, dikelompokkan dan diringkas melalui antarmuka berbasis Flutter dan backend Flask.

Kata Kunci: Agregasi Berita, *Clustering*, Klasifikasi, Peringkasan

INTELLIGENT NEWS AGGREGATION SYSTEM WITH AUTOMATIC CLASSIFICATION, CLUSTERING, AND SUMMARIZATION

Arranged by

Ihsan Ghazi Zulfikar - ihsanghozizulfikar@upi.edu

2103303

ABSTRACT

Online news has become the primary source of information for Indonesian society; however, the overwhelming number of news articles often makes it difficult for readers to find content relevant to their interests. This research aims to develop a web-based news aggregation system capable of automatically performing classification, clustering, and summarization of Indonesian online news articles using artificial intelligence. The system was designed using the ADDIE development framework and implemented with the Waterfall software development method. Data were collected by scraping articles from Kompas and TribunNews in June 2025, resulting in 37,187 articles. The classification model employed a BLSTM-2DCNN trained on the Indosum dataset, achieving an accuracy of 86% and an F1-score of 0.85. Clustering was performed using k-means based on news categories and a five-day time window, with evaluation using the CH-Index yielding an average score of 80.87. For summarization, a fine-tuned BART model was applied, producing ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.6434, 0.5510, and 0.6077, respectively. The developed system presents classified, clustered, and summarized articles through a Flutter-based interface with a Flask backend, enabling readers to access information more efficiently and comprehensively.

Keywords: *Classification, Clustering, Summarization, News Aggregation.*

DAFTAR ISI

PERNYATAAN BEBAS PLAGIARISME	iii
KATA PENGANTAR	iv
UCAPAN TERIMA KASIH.....	v
ABSTRAK.....	vi
<i>ABSTRACT</i>	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	6
1.3 Batasan Masalah	6
1.4 Tujuan Penelitian	7
1.5 Manfaat Penulisan.....	7
BAB II TINJAUAN PUSTAKA.....	9
2.1 Peta Literatur.....	9
2.2 Berita.....	9
2.3 Platform agregasi Berita	10
2.4 <i>Web Scraping</i>	12
2.5 <i>Machine Learning</i>	13
2.6 Klasifikasi	14
2.7 Klasifikasi Teks	15
2.8 <i>Clustering</i>	20
2.9 <i>Clustering</i> Teks.....	21
2.10 Transformer.....	23
2.11 <i>Large Language Model</i>	27
2.12 Peringkasan Teks	28
2.13 Dataset.....	31
2.14 Penelitian Terkait	32
BAB III METODE PENELITIAN.....	34
3.1 Desain Penelitian	34
3.2 Alat Penelitian.....	37

3.3	Bahan Penelitian	38
BAB IV HASIL DAN PEMBAHASAN		39
4.1	<i>Exploratory Data Analysis</i>	39
4.2	Pra-proses Data	41
4.3	Skenario Eksperimen	45
4.4	Pembuatan Model	45
4.4.1	Pembuatan Model Klasifikasi	45
4.4.2	Perancangan Model Clustering	49
4.4.3	<i>Fine-Tuning</i> Model Peringkasan	51
4.5	Perancangan Sistem Agregasi Berita	54
4.5.1	<i>Scraping</i>	54
4.5.2	Klasifikasi	56
4.5.3	Clustering.....	58
4.5.4	Peringkasan.....	63
4.5.5	Antarmuka dan Implementasi Sistem	65
4.6	Evaluasi Sistem.....	67
4.7	Hubungan Antara Klasifikasi, Clustering, dan Peringkasan.....	67
4.8	Pembahasan.....	69
BAB V SIMPULAN DAN SARAN		72
5.1	Kesimpulan	72
5.2	Saran	73
DAFTAR PUSTAKA		74

DAFTAR GAMBAR

Gambar 1.1 Contoh Aplikasi Agregasi Berita Opera News (Sumber: Tangkapan layar aplikasi Opera News, diambil 11 April 2024)	2
Gambar 1.2 Contoh Implementasi Klasifikasi dan Clustering Berita pada Website Agregasi Berita Google Berita (Sumber: Google Berita, diakses dari https://news.google.com pada 7 Mei 2025)	4
Gambar 1.3 Contoh Fitur Ringkasan pada Aplikasi Agregasi Berita Dailyhunt (Sumber: Tangkapan layar aplikasi Dailyhunt, diambil 11 April 2024).....	5
Gambar 2.1 Peta Literatur	9
Gambar 2.2 Contoh Aplikasi Agregasi Berita CNN Indonesia pada Perangkat Mobile (Sumber: Tangkapan layar aplikasi CNN Indonesia, diambil 7 Mei 2025)	11
Gambar 2.3 Proses Web Scraping (Persson, 2019).....	12
Gambar 2.4 Contoh Arsitektur BLSTM-2DCNN (Peng dkk., 2016)	16
Gambar 2.5 Arsitektur Transformer (Vaswani dkk., 2017)	24
Gambar 2.6 Multi Head Attention (Jamil, 2023)	26
Gambar 2.7 Informasi kolom pada dataset IndoSum.....	32
Gambar 2.8 Contoh Sistem Agregasi Berita (Khodra dkk., 2018)	33
Gambar 3.1 Desain Penelitian.....	34
Gambar 4.1 Jumlah artikel unik berdasarkan kolom id.....	39
Gambar 4.2 Distribusi jumlah artikel per kategori.....	40
Gambar 4.3 Distribusi jumlah kata pada artikel.....	40
Gambar 4.4 Distribusi jumlah kata pada ringkasan	41
Gambar 4.5 File pada Dataset IndoSum	42
Gambar 4.6 Grafik Akurasi dan Loss Model Klasifikasi.....	48
Gambar 4.7 Grafik Loss <i>Fine-Tuning</i> BART	54
Gambar 4.8 Sampel Hasil Scraping Berita	55
Gambar 4.9 Informasi Data Hasil Scraping Berita	56
Gambar 4.10 Informasi Data Setelah Tahapan Klasifikasi.....	57
Gambar 4.11 Informasi Hasil Clustering Berita Tanggal 1-5 Juni	60

Gambar 4.12 Halaman Utama Sistem Agregasi Berita.....	66
Gambar 4.13 Halaman Daftar Berita pada Sistem Agregasi Berita.....	67

DAFTAR TABEL

Tabel 2.1 Contoh <i>Bag of Word</i>	22
Tabel 2.2 Perbandingan Model BART, ASGARD, dan MultiBART-GAT pada dataset CNN/DailyMail.....	31
Tabel 2.3 Kolom pada dataset IndoSum	31
Tabel 4.1 Salah Satu Baris IndoSum.....	42
Tabel 4.2 Perbandingan Sebelum dan Sesudah Pra-proses Kolom Paragraphs....	44
Tabel 4.3 Arsitektur Model Klasifikasi.....	47
Tabel 4.4 Hyperparameter untuk Model Klasifikasi.....	47
Tabel 4.5 Hasil Evaluasi Model Klasifikasi.....	49
Tabel 4.6 Perbandingan CH-Index Clustering per Kategori pada Dataset Indosum	50
Tabel 4.7 Hyperparameter untuk <i>Fine Tuning</i> Model BART.....	52
Tabel 4.8 Hasil Perhitungan Nilai ROUGE	52
Tabel 4.9 Distribusi Hasil Klasifikasi Berita	57
Tabel 4.10 Contoh Judul Berita per Kategori Hasil Klasifikasi.....	58
Tabel 4.11 Hasil Evaluasi Clustering pada Beberapa Rentang Waktu dan Kategori	60
Tabel 4.12 Contoh Hasil Clustering Kategori Olahraga pada 23-27 Juni 2025....	61
Tabel 4.13 Judul Berita Terdekat dengan Centroid pada Kategori Olahraga Tanggal 23-27 Juni 2025.....	62
Tabel 4.14 Contoh Hasil Peringkasan Kategori Olahraga pada 23-27 Juni 2025.	64
Tabel 4.15 Hyperparameter untuk Model Klasifikasi B	68
Tabel 4.16 Perbandingan Hubungan Klasifikasi, Clustering, dan Peringkasan....	69

DAFTAR PUSTAKA

- Alonso, O., Fetterly, D., & Manasse, M. (2013). Duplicate news story detection revisited. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8281 LNCS, 203–214. https://doi.org/10.1007/978-3-642-45068-6_18
- Annur, C. (2023). *Ada 185 Juta Pengguna Internet di Indonesia pada Januari 2024*. Databooks. <https://databoks.katadata.co.id/datapublish/2024/02/27/ada-185-juta-pengguna-internet-di-indonesia-pada-januari-2024>
- Asiri, S. (2022). *An Introduction to Classification in Machine Learning*. BuiltIn. <https://builtin.com/machine-learning/classification-machine-learning>
- Caliński, T., & Harabasz, J. (2007). *Communications in Statistics - Theory and Methods A dendrite method for cluster analysis*. 3272(1974).
- Calzada, J., & Gil, R. (2020). What do news aggregators do? Evidence from google news in Spain and Germany. *Marketing Science*, 39(1), 134–167. <https://doi.org/10.1287/mksc.2019.1150>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3). <https://doi.org/10.1145/3641289>
- Chen, T., Wang, X., Yue, T., Bai, X., Le, C. X., & Wang, W. (2023). Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers. *Applied Sciences (Switzerland)*, 13(13). <https://doi.org/10.3390/app13137753>
- Claussen, J., & Peukert, C. (2019). Obtaining Data from the Internet: A Guide to Data Crawling in Management Research. *SSRN Electronic Journal*, 1–38. <https://doi.org/10.2139/ssrn.3403799>
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019, Mlm*, 4171–4186. <https://aclanthology.org/N19-1423.pdf>
- Dewan Pers. (2024). *SURVEI LANSKAP MEDIA PERS INDONESIA*. <https://dewanpers.or.id/berita/detail/2554/survei-lanskap-media-pers-indonesia>
- El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165, 1–46. <https://doi.org/10.1016/j.eswa.2020.113679>
- Farid, A. S. (2023). Changing the Paradigm of Traditional Journalism to Digital Journalism: Impact on Professionalism and Journalism Credibility. *Journal International Dakwah and Communication*, 3(1), 22–32. <https://doi.org/10.55849/jidc.v3i1.374>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. <http://arxiv.org/abs/2203.05794>
- Gurung, P., & Wagh, R. (2017). A study on Topic Identification using K means clustering algorithm: Big vs. Small Documents. 10(2), 221–233.

- <http://www.ripublication.com>
- Jamil, U. (2023). *Transformer from Scratch*. <https://github.com/hkproj/transformer-from-scratch-notes>
- Khodra, M. L., Wibisono, Y., Fauzan, A., & Rahman, B. (2018). *Pembangunan Sistem Cerdas Agregator Berita Bahasa Indonesia*. <https://osf.io/preprints/inarxiv/xmjyq/%0Ahttps://osf.io/preprints/inarxiv/xmjyq/download>
- Kulkarni, A., & Shivananda, A. (2019). Natural language processing recipes: Unlocking text data with machine learning and deep learning using python. In *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python*. <https://doi.org/10.1007/978-1-4842-4267-4>
- Kurniawan, K., & Louvan, S. (2018). IndoSum: A New Benchmark Dataset for Indonesian Text Summarization. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 215–220. <https://doi.org/10.1109/IALP.2018.8629109>
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A Survey on Text Classification: From Shallow to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, 37(4). <http://arxiv.org/abs/2008.00364>
- Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Japanese Circulation Journal*. <https://doi.org/10.1253/jcj.34.1213>
- Liu, Y., Shi, K., He, K. S., Ye, L., Fabbri, A. R., Liu, P., Radev, D., & Cohan, A. (2023). *On Learning to Summarize with Large Language Models as References*. <http://arxiv.org/abs/2305.14239>
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). *Deep Learning Based Text Classification: A Comprehensive Review*. 1(1), 1–43. <http://arxiv.org/abs/2004.03705>
- Mohamed, A., Ibrahim, M., Yasser, M., Ayman, M., Gamil, M., & Hassan, W. (2020). News aggregator and efficient summarization system. *International Journal of Advanced Computer Science and Applications*, 11(6), 636–641. <https://doi.org/10.14569/IJACSA.2020.0110677>
- Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. In *Artificial Intelligence Review* (Vol. 56, Issue 7). Springer Netherlands. <https://doi.org/10.1007/s10462-022-10325-y>
- Peng, Z., Qi, Z., Zheng, S., Xu, J., Bao, H., & Bo, X. (2016). *Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling.pdf* (pp. 3485–3495).
- Persson, E. (2019). Evaluating tools and techniques for web scraping. *Degree Project in Computer Science and Engineering, SECOND CYC*, 1–97. <https://www.diva-portal.org/smash/get/diva2:1415998/FULLTEXT01.pdf>
- Ramdhani, K. (2019). Text Mining Pada News Aggregator Untuk Sistem Rekomendasi Berita. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699.
- Rashid, J., Shah, S. M. A., & Irtaza, A. (2020). An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-means

- Clustering. *Mehran University Research Journal of Engineering and Technology*, 39(1), 213–222. <https://doi.org/10.22581/muet1982.2001.20>
- Ratna, D., Afifah, J., & Kaka, A. (2025). *Prosiding Seminar Nasional Manajemen Peran Media Massa dalam Membentuk Opini Publik*. 4(1), 1296–1298.
- Rich, C. (2010). *WRITING & REPORTING NEWS: News Scene 3.0 A Coaching Methode*.
- Schubiger, R. (2024). *German Summarization with Large Language Models*.
- Sreenivasulu, G., Thulasi Chitra, N., Sujatha, B., & Venu Madhav, K. (2022). Text Summarization Using Natural Language Processing. *Lecture Notes in Networks and Systems*, 191(January), 653–663. https://doi.org/10.1007/978-981-16-0739-4_62
- Suyal, H., Panwar, A., & Singh Negi, A. (2014). Text Clustering Algorithms: A Review. *International Journal of Computer Applications*, 96(24), 36–40. <https://doi.org/10.5120/16946-7075>
- Thakkar, B., Shah, R., Shah, V., & Machado, S. (2023). *AI Based News Aggregator and Summarizer*. 25(09), 147–153.
- Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4), 267–276. <https://doi.org/10.1007/BF02289263>
- Tian, Q. (2022). Impact of Social Media News Overload on Social Media News Avoidance and Filtering: Moderating Effect of Media Literacy. *Frontiers in Psychology*, 13(March). <https://doi.org/10.3389/fpsyg.2022.862626>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-Decem(Nips)*, 5999–6009.
- Vyawahare, D. H. R. (2022). Machine Learning: A Solution Approach for Complex Problems. *Interantional Journal of Scientific Research in Engineering and Management*, 06(04), 8–10. <https://doi.org/10.55041/ijrsrem16123>
- Wahono, N. V., Wibowo, A., & Intan, R. (2015). Aplikasi Indonesian News Aggregator Berbasis Android yang Didukung oleh Sistem Rekomendasi. *Jurnal Infra*, 3(1), 121–127. <https://publication.petra.ac.id/index.php/teknik-informatika/article/view/2802>
- Yeasmin, S., Kuri, R., Mahamudul Hasan Rana, A. R. M., Uddin, A., Sala Uddin Pathan, A. Q. M., & Riaz, H. (2021). Multi-category Bangla News Classification using Machine Learning Classifiers and Multi-layer Dense Neural Network. *International Journal of Advanced Computer Science and Applications*, 12(5), 757–767. <https://doi.org/10.14569/IJACSA.2021.0120588>