p-ISNN : 2338-4697
e-ISSN : 2579-3322

Implementasi LDA untuk Pengelompokan Topik Twitter Bertagar #Mypertamina

Hery Oktafiandi
Teknologi Rekayasa Perangkat Lunak Politeknik Sawunggalih Aji
Email : heryokta@gmail.com

Abstrak

Media sosial twitter banyak digunakan oleh pengguna sebagai media komunikasi dan informasi. Selain sebagai alat komunikasi twitter digunakan untuk mendapatkan data penelitian yang dibutuhkan. Penggunaan tagar twitter menjadi acuan trending berita atau isu yang sedang berkembang di masyarakat. Trending yang sedang hangat dibicarakan saat ini adalah tetang aplikasi Mypertamina. Penelitian ini mengambil data dari twitter dengan tagar #Mypertamina dengan banyak data twitter sebanyak 149 tweet, dari data yang didapat maka akan diklasterkan menggunakan topic modelling metode Latent Dirichlet Allocation (LDA). Kelebihan dari metode LDA adalah dapat mengklasterkan, meringkas, dan menghubungkan data dalam jumlah yang banyak. Penelitian ini menghasilkan 3 kluster data dengan nilai coherence terbesar 0.468
Kata kunci *Mypertamina, Twitter, LDA.*

Abstract

Twitter social media is widely used by users as a medium of communication and information. Apart from being a communication tool, Twitter is used to obtain the required research data. The use of the twitter hashtag becomes a reference for trending news or issues that are developing in the community. The trend that is currently being discussed is the Mypertamina application. This study takes data from twitter with the hashtag #Mypertamina with a lot of twitter data as many as 149 tweets, from the data obtained it will be clustered using topic modeling with the Latent Dirichlet Allocation (LDA) method. The advantage of the LDA method is that it can cluster, summarize, and link large amounts of data. This study resulted in 3 data clusters with the largest coherence value of 0.4618
Keywords: *Mypertamina, Twitter, LDA*

1. Pendahuluan

Layanan twitter yang paling populer sekarang adalah layanan *microblogging* yang digunakan pada media internet(Binsaeed et al., 2020). Ketertarikan pada data twitter terus tumbuh karena banyak data yang dapat diolah(Kearney, 2019). Tagar pada twitter sebagai acuan trending berita atau isu yang berkembang pada masyarakat, disini kami melakukan analisi topik yang sedang trend(Carneiro et al., 2021) berlangsung dari pertengahan bulan juni hingga juli 2022 yaitu #Mypertamina, adanya perkembangan *e-commerce* menghadirkan suatu inovasi pembayaran elektronik(Ibrahim & Karina Moeliono, 2020)yang akan digunakan sebagai alat pembayaran pada pengisian bahan bakar, hal ini banyak pengguna Twitter yang membahas soal aplikasi ini hingga menjadi trending topik seluruh indonesia.

Penelitian sebelumnya meneliti topik kesehatan di Indonesia menggunakan metode topic modeling LDA yang menggunakan sumber data dari penelitian khusus dibidang kesehatan Indonesia dan *scraping* data dari jurnal SINTA yang diproleh dari bulan januari 2020 dan memilik scraping sebanyak 11269 penelitian(Sistem et al., 2021), pada penelitian ini menghasilkan 3 kelompok model. Fazha Rahhid d.k.k (Rashif et al., 2021) melakukan penelitian mengenai cuitan akun bot pada tagar#covid19 menggunakan metode LDA menggunakan data 128 tweet dan mendapatkan hasil 5 topik teratas

Penelitian ini menggunakan Metode LDA (Latent Dirichlet Allocation) untuk mengetahui macam-macam topik yang dibahas terhadap penyebaran informasi di social media Twitter yang membahas aplikasi buatan pertamina dengan menggunakan tagar #mypertamina.

1.1. Topic Modeling

Dasar dari topic modeling adalah kluster kata yang berisi kata-kata yang menyusun kluster

p-ISNN : 2338-4697
e-ISSN : 2579-3322

tersebut serta memiliki kemungkinan dari beberapa cluster katadengan masing – masing probabilitas. Topic modelling telah mencapai kemajuan yang mengesankan dengan perkembangan cepat model generatif saraf(Yu et al., 2022).

Sejumlah dokumen mempunyai penyebaran probabilitas kluster, tiap kata diambil dari salah satu kluster tersebut.

Dalam sesi ini kami juga menggunakan beberapa teknik pemodelan salah satunya yaitu topik pemodelan, adalah teknik paling kuat untuk text-mining, penemuan data dan menemukan hubungan antar data dan dokumen teks. Para peneliti telah menerbitkan banyak artikel dan jurnal di bidang pemodelan yang mencakup perangkat lunak rekayasa , ilmu politik, bahasam dan ilmu medic(Jelodar et al., 2019).

1.2. *Latent Dirichlet Allocation* (LDA)

Salah satu metode yang dapat digunakan dalam pemodelan topik yaitu latent dirichlet allocation (LDA). Latent dirichlet allocation (LDA) memiliki keunggulan bisa mengklasterkan data yang jumlahnya besar dibandingkan metode pemodelan topik yang lain serta dapat diimplementasikan untuk mengidentifikasi topik dalam jurnal ilmiah, klasifikasi, dan pengelompokan(Putri et al., 2021).

1.3. Nilai *Coherence*

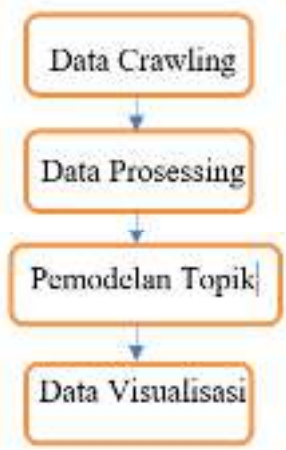
Nilai *coherence* digunakan untuk mengevaluasi model pada topic modelling, nilai koherensi yang tinggi akan menghasilkan model topik yang baik.

2. Pembahasan

Ada beberapa tahap – tahap dalam penelitian ini , yaitu :

1. *Data Crawling*
2. *Data Processing*
3. Pemodelan topic
4. Data visualisasi

Berikut penjelasan dari masing – masing tahapan penelitian.

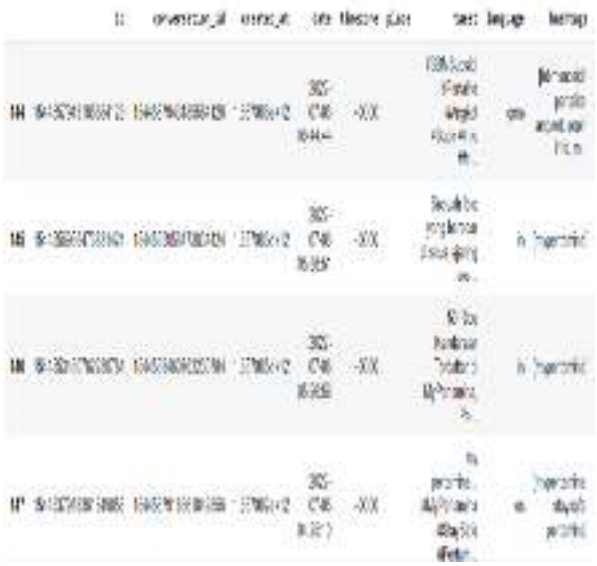


Gambar 1. Tahapan Metode Penelitian

2.1. *Data Crawling*

Crawling data yang menggunakan HTTP GET dengan kredensiap biasa di hadirkan oleh konsol pengembang Twitter(Sohail et al., 2021), data diambil dari media sosial twitter API(Putri et al., 2021). *Crawling* data dengan mengumpulkan tweet dengan tagar #mypertamina dengan bahasa Indonesia dijalankan pada bahasa pemorograman *Python* yang mendukung banyak bahasa manusia(Qi et al., 2020).

Data yang digunakan pada penelitian ini dengan mengambil data pada twitter bertagar #Mypertamina dengan proses *crawling*. Gambar 2 merupakan contoh dari hasil crawling data tweet tagar#Mypertamina.



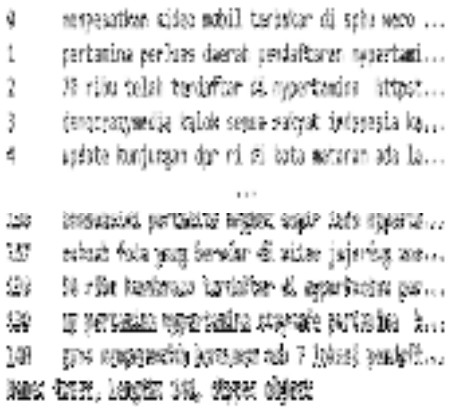
Gambar 2 Kumpulan Data Tweet

p-ISNN : 2338-4697
e-ISSN : 2579-3322

2.2. Data Prosessing

Selanjutnya dilakukan preprosessing data yang bertujuan membersihkan data sehingga bisa digunakan pada tahap pemodelan., oleh karena itu pemrosesan data adalah proses iterasi interaktif dan membutuhkan umpan balik korektif berkelanjutan(Sun et al., 2018). Data yang sama dikelompokkan dalam indek sebagai pembeda antar dokumen. Data yang telah dikelompokkan akan melalui proses tokenisas yang berperan penting dalam NLP(Ahmadi, 2020) untuk memecah menjadi token.

Data yang terkumpul selanjutnya dilakukan preprossing data, tahapan awal preprossesing data yaitu yang dengan mengambil kolom tweet saja, kemudian dilakukan *case folding*, *remove mention*, *remove hastag*, *remove number*, *tokenisasi*, *stemming*, dan *stopword* bahasa Indonesia. Setelah smua proses dilakukan maka data yang didapat sudah bersih dan siap diolah ke tahap pemodelan LDA. Gambar 3 merupakan hasil dari proses preprosessing data.

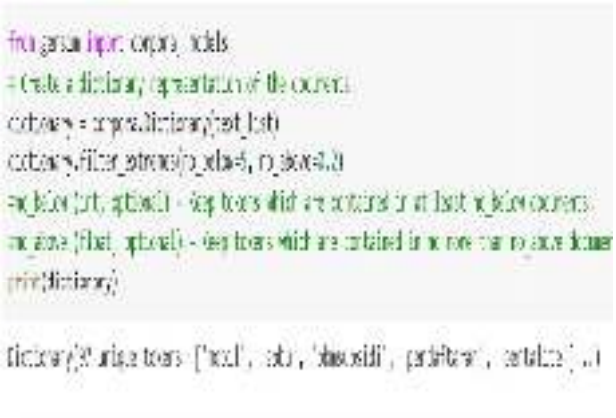


Gambar 3 Hasil Tweet yang telah melalui prosessing

2.3. Pemodelan Topik

Langkah selanjutnya adalah membangun model LDA. LDA berfungsi sebagai pengklasteran data, dan meringkas data dalam jumlah yang besar, oleh karena itu dengan pemodelan topik LDA sama dengan pendekatan *unsupervised*(Purwitasari et al., 2021). Pemodelan dilakukan dengan membentuk model bigram dan trigram menggunakan *tool gesim* model dan *gesim corpora* dengan *python*. Setelah pemodelan, dilakukan perhitungan *coherence* untuk mengevaluasi model topic yang terbentuk. (Impraimakis & Smyth, 2022). Perhitungan coherence dengan memanggil modul *coherencemodul* (Hasan et al., 2021),

Langkah pertama pemodelan *topic modelling* yaitu dengan membuat model *bigram* dan *trigram*. *Bigram* merupakan kemunculan dua kata yang sama, sedangkan *trigram* adalah kemunculan tiga kata yang sama. Kemudian membuat *dictionary* dari data *text* sebelum pemodelan LDA, membangun *dictionary* dengan memmanfaatkan modul *gesim corpora* dengan memanggil modul *Dictionary*. Modul *gesim* modul digunakan untuk memanggil modul LDAModul sebagai dasar pembentukkan model LDA. Dua input utama untuk membangun model LDA adalah *dictionary* terlihat pada gambar 4 dan *corpus* terlihat pada gambar 5. Gambar 6 merupakan hasil dari pemodelan LDA, yang menghasilkan 3 topik model teratas.



Gambar 4 Pembentukan Dictionary\

Gambar 5 Pembentukan Corpus

p-ISNN : 2338-4697
e-ISSN : 2579-3322

Gambar 6. Pemodelan Topik Modeling

Hasil dari pemodelan dapat dilihat pada tabel 1 berupa 10 kata kunci teratas pada topik – topik yang dibentuk.

Tabel 1. Hasil Topik Modeling

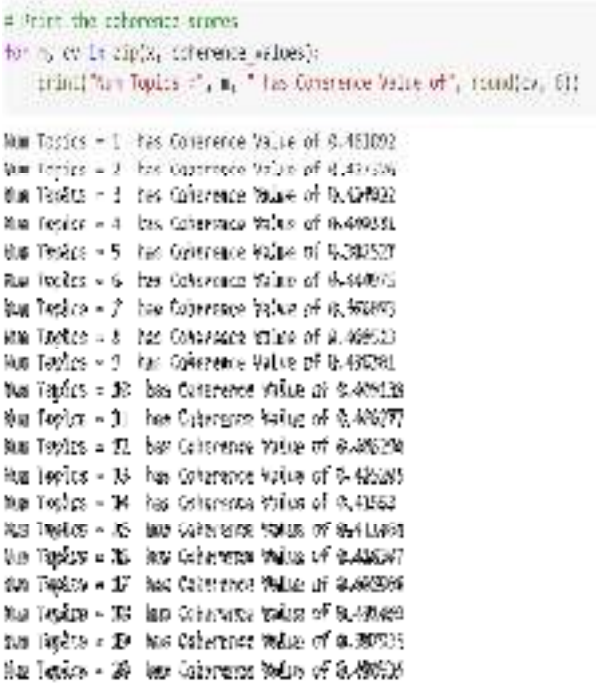
Topik	Probabilitas* Kata
0	0.053*"pertalite+0.033*"kendaraan" +0.032 *"bbmbersubsidi"+0.024*"spbu" +0.025"pakai" +0.025"mobil" 0.02*"sudah"+0.019*"dan" +0.019*"untuk"
1	+0.52mypertamina_call135 +0.023dan+0.022call135 +0.022ada +0.019mmc_milenial 0.018my_pertamina +0.017my +0.017kamu+0.017yg +0.017lebihbaikpertamax
2	0.027bisa+0.026subsidi +0.026mypertamina_call135 +0.026penggunaan +0.023pertalite+0.022jadi +0.021daftar +0.021bikin+0.020data +0.019punya

Dari hasil topik modeling yang didapat maka dilakukan analisa pada hasil topic yang terbentuk. Tabel 2 merupakan analisa hasil topic modeling yang telah dianalisa.

Tabel 2. Hasil Analisa Topik

Topik	Hasil Analisa Topik
0	Kendaraan mobil memaakai bbm bersubsidi
1	Mypertamina_call135 lebih baik pertamax
2	Mypertamina punya data daftar pengguna bbm bersubsidi pertalite

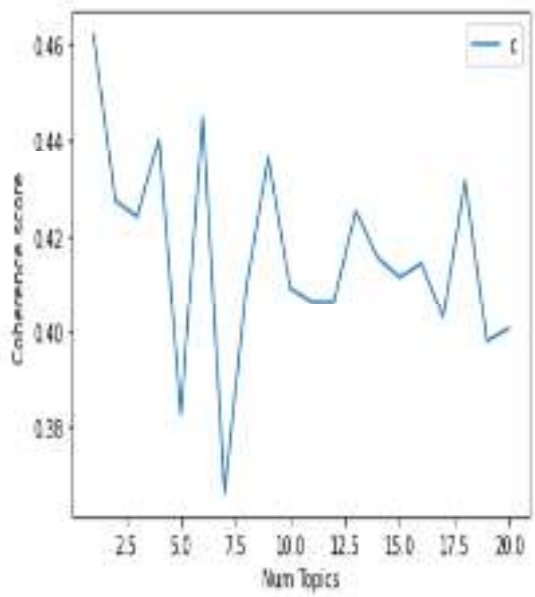
Tahapan selanjutnya adalah perhitungan nilai *coherence*, gambar 5 merupakan hasil dari nilai *coherence* yang didapat



Gambar 7 Nilai Coherence

Semakin besar nilai *coherence* maka *interpretasi topic modeling* yang dihasilkan semakin baik. Nilai *coherence* yang terbaik pada penlitian ini benilai 0.418 terdapat pada topik 1. Nilai coherence juga bisa dilihat pada gambar 6 yang merupakan gambar grafik nilai *coherence*

p-ISNN : 2338-4697
e-ISSN : 2579-3322



Gambar 8. Grafik Nilai Coherence

2.4. Data Visualisasi

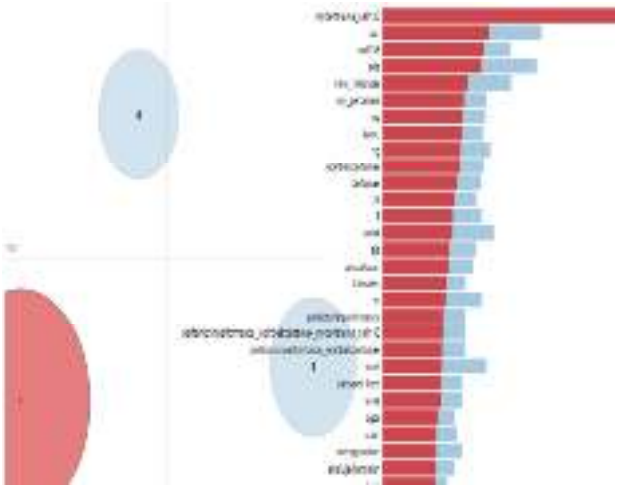
Visualisasi penerapan LDA menggunakan grafik dan *bar-chart* dengan memanfaatkan *tool PylDAvis* pada *python* (Sezer & Ozbayoglu, 2020). *PylDAvis* digunakan untuk menterjemahkan kluster yang terbentuk sesuai dengan data.

Pemodelan LDA telah dibangun selanjutnya visualisasi hasil pemodelan. Visualisasi data menggunakan modul *PyDAvis* pada *python*. Visualisasi membantu menterjemahkan topic sesuai dengan kumpulan data. Pada gambar visualisasi terlihat 2 gambar, sebelah kiri menunjukkan topic secara keseluruhan dan dapat terlihat hubungan antar topik. Bagian sebelah kanan menunjukkan distribusi frekuensi kata pada setiap topik.

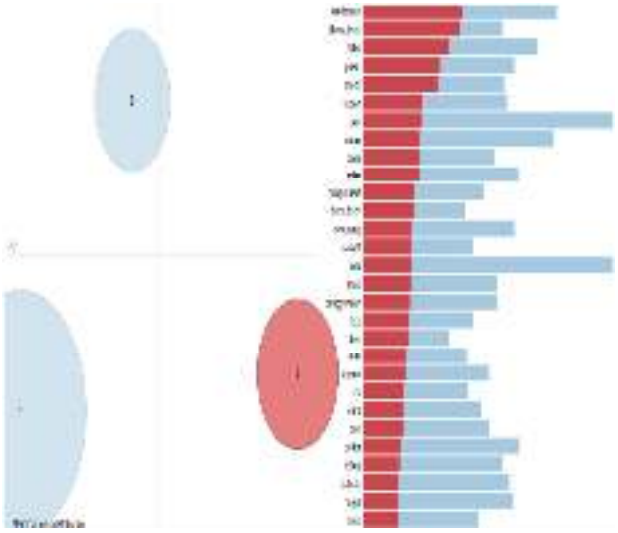
Pada penelitian ini terdapat 3 model topik yang terbangun. Topik 1 terlihat pada gambar 7 terdapat 30 *terminologi* kata yang terdapat pada topic. Pada topic 1 membahas tentang mypertamina_call135, mypertamina, call135, lebih baik pertamax, pertamax, pendaftaran, transaksi, ini, ribet, dan transaksi bbm.

Topik 2 dapat dilihat pada gambar 8 membahas tentang kendaraan, bbm subsidi ,spbu, pakai, mobil, sudah, masyarakat, bersubdisidi, rakyat.

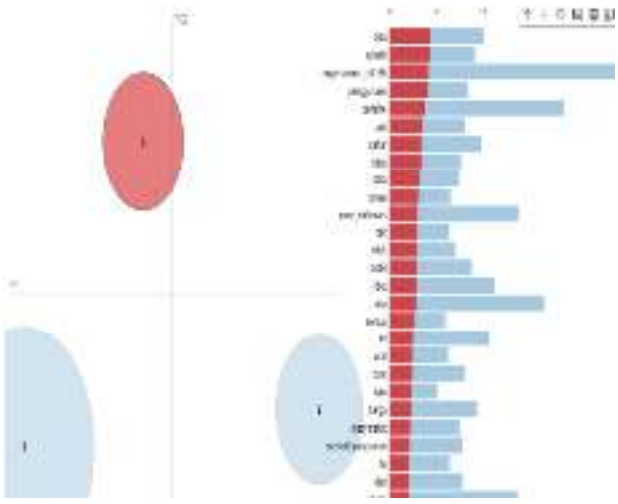
Topik 3 terlihat pada gambar 9, diantaranya membahas tentang bisa, subsidi, mypertamina_call135, penggunaan, pertalite, cara, bikin, daftar.



Gambar 9. Topik 1



Gambar 10 Topik 2



Gambar 11 Topik 3

p-ISNN : 2338-4697
e-ISSN : 2579-3322

3. Kesimpulan

Metode pengklasteran LDA pada twitter bertagar #mypertamina mengelompokkan data twitter menjadi 3 buah topic. Dari sejumlah topic yang dipilih diperoleh nilai *coherence* tertinggi sebesar 0.4618. Melihat hasil topic modeling yang terbentuk dapat disimpulkan tweet dengan tagar#Mypertamina membahas tentang Mypertamina_call135, pertamax, bbm bersubsidi dan pertalite. Kekurangan pada penelitian ini belum adanya perbandingan perhitungan manual, penelitian ini hanya menggunakan modul yang ada pada python.

.

Daftar Pustaka

Ahmadi, S. (2020). A Tokenization System for the Kurdish Language. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 114–127.

Binsaeed, K., Stringhini, G., & Youssef, A. E. (2020). Detecting Spam in Twitter Microblogging Services: A Novel Machine Learning Approach based on Domain Popularity. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 11).

Carneiro, A., Matos, M. J., Uriarte, E., & Santana, L. (2021). Trending topics on coumarin and its derivatives in 2020. In *Molecules* (Vol. 26, Issue 2). MDPI AG. https://doi.org/10.3390/molecules26020501

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in latent dirichlet allocation (lda). *Advances in Intelligent Systems and Computing*, *1309*, 341–354. https://doi.org/10.1007/978-981-33-4673-4_27

Ibrahim, R. M., & Karina Moeliono, N. N. (2020). Persepsi manfaat, kepercayaan, efikasi diri, kemudahan penggunaan, keamanan terhadap persepsi konsumen pada my pertamina (Studi pada penggunaan my pertamina kota Bandung. *Jurnal Ilmiah Mahasiswa Ekonomi Manajemen Accredited SINTA*, *4*(2), 396–413.

Impraimakis, M., & Smyth, A. W. (2022). Input–parameter–state estimation of limited information wind-excited systems using a sequential Kalman filter. *Structural Control and Health Monitoring*, *29*(4). https://doi.org/10.1002/stc.2919

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. In *Multimedia Tools and Applications* (Vol. 78, Issue 11). https://doi.org/10.1007/s11042-018-6894-4

Kearney, M. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, *4*(42), 1829. https://doi.org/10.21105/joss.01829

Purwitasari, D., Aida Muflichah, Novrindah Alvi Hasanah, & Agus Zainal Arifin. (2021). Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 421–428. https://doi.org/10.29207/resti.v5i3.3049

Putri, S. A., Kusuma, P. D., & Setianingsih, C. (2021). Clustering Topik Pada Data Sentimen BPJS Kesehatan Menggunakan Metode Laten Dirichlet Allocation. *E-Proceeding of Engineering*, *8*(5), 6097–6105.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Rashif, F., Ihza Perwira Nirvana, G., Alif Noor, M., & Aini Rakhmawati, N. (2021). Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19 LDA Implementation for Topic of Bot's Tweets with #Covid-19 Hashtag. *Cogito Smart Journal* |, *7*(1), 170–181.

Sezer, O. B., & Ozbayoglu, A. M. (2020). Financial trading model with stock bar chart image time series with deep convolutional neural networks. *Intelligent Automation and Soft Computing*, *26*(2), 323–334. https://doi.org/10.31209/2018.100000065

Sistem, R., Sahria, Y., & Fudholi, D. H. (2021). JURNAL RESTI Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode. *Jurnal Rekayasa Sistem Dan Teknologi Informasi (RESTI)*, *1*(10), 336–344.

Sohail, S. S., Khan, M. M., Arsalan, M., Khan, A., Siddiqui, J., Hasan, S. H., & Alam, M. A. (2021). *Crawling Twitter data through API: A technical/legal perspective*.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: A

p-ISNN : 2338-4697
e-ISSN : 2579-3322

review. In *Journal of Healthcare Engineering* (Vol. 2018). Hindawi Limited. https://doi.org/10.1155/2018/4302425

Yu, P., Xie, S., Ma, X., Jia, B., Pang, B., Gao, R., Zhu, Y., Zhu, S.-C., & Wu, Y. N. (2022). *Latent Diffusion Energy-Based Model for Interpretable Text Modeling*. *2020*.