

Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data *Spam* dan *Non-Spam*

Prismahardi Aji Riyantoko¹, Tresna Maulana Fahrudin¹, Dwi Arman Prasetya¹
Trimono¹, Tahta Dari Timur²

¹Program Studi Sains Data

prismahardi.aji.ds@upnjatim.ac.id, tresna.maulana.ds@upnjatim.ac.id, dwi.arman.sada@upnjatim.ac.id, trimono.stat@upnjatim.ac.id

²Gains Associate, Vilnius, Lithuania

tahtadaritimoer@gmail.com

Corresponding author email: prismahardi.aji.ds@upnjatim.ac.id

Abstract: Sentiment analysis has become a useful tool for doing data analysis and classification based on words, phrases, or documents. Previously, researchers conducted extensive research on sentiment analysis using a variety of algorithms and models. Based on previous research, the results of the sentiment analysis have a negative impact on model performance and data type. At the moment, researchers are using the LSTM and BERT models to classify SMS data into spam and non-spam. The researcher using TD-IDF and GloVe algorithm to determine the weighting of the values represented in vectors in each word to optimize the results of value accuracy. Regardless of the results obtained, the methods BERT and LSTM have a value accuracy sensitivity of 99.35% and 98.22%, respectively. The results present that the completion of spam and non-spam dataset classification is very effective and efficient. Tests were also carried out using disaster twitter data, but the level of accuracy of the values decreased. Therefore, it can be supposed that the different types of datasets considerably affect the performance of the temptation model.

Keywords: Model Analysis, LSTM, BERT, Weighting

Abstrak: Analisis sentimen menjadi dasar untuk melakukan analisis dan klasifikasi data berdasarkan kata, frasa, atau dokumen. Penelitian terkait analisis sentimen telah banyak dilakukan oleh peneliti sebelumnya dengan menggunakan berbagai algoritma dan jenis model. Berdasarkan penelitian sebelumnya hasil analisis sentimen dipengaruhi performansi model dan jenis data. Saat ini peneliti menggunakan model LSTM dan BERT untuk menyelesaikan klasifikasi pada data sms yang terdiri dari *spam* dan *non-spam*. Untuk mengoptimalkan hasil pada akurasi nilai, peneliti menggunakan algoritma TD-IDF dan GloVe untuk menentukan pembobotan nilai yang direpresentasikan dalam vektor pada setiap kata. Adapun hasil yang sudah kami peroleh, metode BERT dan LSTM memiliki tingkat keakurasian nilai secara terurut sebesar 99,35% dan 98,22%. Hal tersebut menunjukkan bahwa dalam penyelesaian klasifikasi dataset *spam* dan *non-spam* sangat efektif dan efisien. Pengujian juga dilakukan dengan menggunakan data *disaster twitter*, tetapi tingkat keakurasian nilai mengalami penurunan. Oleh karena itu, dapat disimpulkan bahwa perbedaan jenis dataset sangat mempengaruhi performansi model uji coba.

Kata kunci: Analisis Model, LSTM, BERT, Pembobotan

I. PENDAHULUAN

Menurut laporan terbaru dari *We Are Social*, Indonesia akan memiliki 175,4 juta pengguna internet pada tahun 2020. Dibandingkan tahun sebelumnya, jumlah pengguna internet di Tanah Air meningkat 17% atau 25 juta. Berdasarkan jumlah penduduk Indonesia sebesar 272,1 juta, ini berarti 64% setengah dari penduduk Indonesia memiliki akses ke dunia maya. Salah satu kunci penting keberhasilan penetrasi internet di Indonesia adalah pembangunan infrastruktur yang menjangkau daerah-daerah terpencil di Indonesia [1]. Perkembangan teknologi informasi khususnya media online telah mengubah cara masyarakat berkomunikasi. Media online banyak digunakan oleh masyarakat untuk menyampaikan pendapat, pengalaman dan keprihatinan lainnya.

Saat ini perkembangan teknologi semakin pesat, memudahkan dan mempercepat penyebaran informasi melalui media online (Facebook, Twitter), blog, atau website resmi institusi [2]. Kemudahan dan kecepatan media online dapat mengubah cara masyarakat mengkonsumsi berita. Surat kabar dan majalah fisik telah digantikan oleh media online seperti berita online dan weblog. Di dunia sekarang ini, orang ingin mengkonsumsi berita sebanyak mungkin dari berbagai sumber yang mereka anggap penting atau sesuai dengan kepentingan mereka sendiri. Media online telah

mengembangkan strategi yang efektif untuk mendapatkan perhatian. Media online, ketika meliput peristiwa, mengungkapkan pendapatnya tentang entitas yang terdiri dari orang, tempat, dan bahkan benda [3]. Untuk alasan ini, layanan peringkat sentimen interaktif ditawarkan oleh saluran yang berbeda di beberapa situs web berita, dan berita dapat dikategorikan sebagai negatif dan positif. Media online menghasilkan banyak jenis berita, termasuk bisnis, politik, kesehatan, olahraga, dan sains. Diantaranya, ekonomi menjadi salah satu topik pembahasan yang menarik. Ekonomi secara langsung mempengaruhi warga, bisnis dan bahkan pasar tradisional, tergantung pada situasi ekonomi negara [4].

Analisis sentimen (SA) sering disebut juga sebagai *opinion mining*, yang digunakan untuk menganalisis atau mengklasifikasikan pengguna berdasarkan kata, frasa, atau dokumen [5]. Beberapa penelitian telah mengeksplorasi analisis sentimen menggunakan data, model, dan tingkat akurasi yang berbeda diantaranya

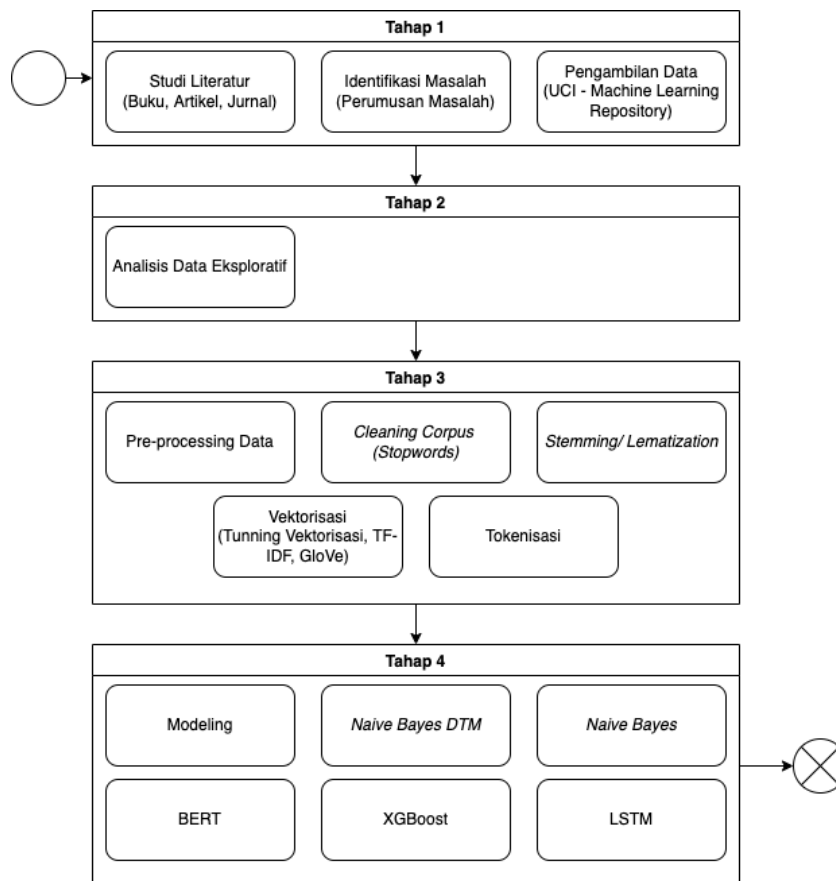
Tabel 1. Artikel, Data, dan Model untuk Sentimen Analisis

No	Data	Model	Tingkat Akurasi	Referensi
1	detik finance	LSTM	51%	[6]
		LSTM-CNN	53%	
		CNN-LSTM	62%	
2	sms	Linear Regresion	95%	[7]
		Naïve Bayes	93%	
3	sms	Naïve Bayes	90%	[10]
		SVM	81%	
4	google play store	BERT	99%	[8]
5	twitter	Bi-GRU	97.7%	[9]
6	twitter	TF-IDF - SVM	85%	[11]
7	twitter	GloVe – SVM	79,52%	[12]

Berdasarkan penelitian yang sudah dilakukan [6-12] didapatkan hasil akurasi yang bervariasi dalam menentukan analisis sentiment, hal ini disebabkan oleh perbedaan metode dan dataset yang digunakan. Oleh sebab itu, peneliti tertarik untuk melakukan pengujian data sms menggunakan pendekatan terbobot dengan algoritma **GloVe** (*Global Vectors for Word Representations*) dan **TF-IDF** (*Term Frequency-Inverse Document Frequency*) serta pendekatan model **BERT** (*Bidirectional Encoder Representations from Transformers*), Naïve Bayes, Naïve Bayes DTM, XGBoost, dan **LSTM** (*Long Short-Term Memory*).

II. METODE PENELITIAN

Metode yang kami gunakan pada penelitian ini terdiri dari beberapa tahapan yang tervisualisasi pada diagram alir pada gambar 1 berikut ini



Gambar 1. Diagram Alur Penelitian

Penelitian ini dimulai dengan melakukan survei penelitian melalui tahap 1 yaitu membaca studi literatur dari buku, artikel, atau jurnal yang sudah terpublikasi. Kemudian melakukan klasifikasi dataset, model atau metode untuk menyelesaikan masalah, serta melihat nilai akurasi setiap model. Setelah itu melakukan gap analisis untuk merumuskan masalah yang akan diteliti. Kami melakukan pengambilan data melalui website terbuka UCI – Machine Learning. Selanjutnya melakukan analisis data dengan melakukan pembersihan data dan visualisasi data awal. Setelah itu, melakukan *pre-processing data* dengan melakukan pembersihan data dengan menggunakan library *stopwords*. Pada tahap *Stemming* atau *Lematization* yakni mengembalikan kata ke awal menjadi bentuk asal yang disesuaikan dengan aturan Bahasa. Tokenisasi bertujuan untuk melakukan pemisahan setiap kata yang Menyusun sebuah dokumen, dengan memisahkan kata diantara spasi, dan karakter lainnya. Untuk melakukan pembobotan kami menggunakan tunning, TF-IDF, dan GloVe. Pada tahap akhir, kami menggunakan pengujian dan pendekatan melalui model *Naïve Bayes DTM*, *Naïve Bayes*, *XGBoost*, *BERT*, dan *LSTM*.

2.1. Pengambilan Data

Pada penelitian ini, kami mengambil data dari UCI – Machine Learning “*SMS Spam Collection Data Set*”. Pada data tersebut terdapat 5572 sms yang diekstrak dari *Grumbletext Web site*. Data terdiri dari tiga atribut yaitu *target*, *message*, dan *message_len*. Tujuan pada penelitian ini untuk mengklasifikasi sms *spam* dan *non-spam* (ham), sehingga dapat menyatukan kategori dari pesan spam.

2.2. Preprocessing

1) *Case Folding*

Pada tahap ini melakukan konversi atau mengubah *text* menjadi bentuk standart, sehingga data masukan atau primer akan diubah menjadi huruf kecil atau *lowercase*.

2) *Tokenisasi*

Pada tahap ini melakukan kemudahan *preprocess text* untuk memisahkan kata menjadi token atau pemisahan kata satu per satu. Sehingga, dilakukan penghapusan tanda baca yang tidak digunakan.

3) *Filtering*

Pada tahap ini melakukan *filtering* dengan menggunakan *stopwords* untuk menghilangkan kata-kata yang tidak memiliki arti signifikan.

4) *Normalisasi*

Pada tahap ini melakukan penyeragaman kata yang memiliki makna ganda atau sama, yang dapat terjadi reduksi kata atau pengulangan data. Karena pengulangan data akan menyebabkan proses komputasi menjadi tidak efektif.

5) *Stemming atau Lemmatization*

Pada tahap ini bertujuan untuk mengembalikan kata yang berimbuhan menjadi kata induk.

6) *Term Weighting.*

A. *Term Frequency – Inverse Document Frequency (TF-IDF)*

Langkah selanjutnya setelah preprocessing, data sms kemudian disimpan dan diproses untuk menghasilkan nilai TF-IDF. Proses pembobotan dilakukan pada setiap kalimat dokumen. Nilai yang dihasilkan dihitung dari jumlah kata dalam sebuah kalimat. Prosedur Term Frequency (TF) menghitung jumlah kata yang muncul pada setiap kalimat dokumen, sedangkan prosedur Inverse Document Frequency (IDF) menunjukkan seberapa sering kata muncul dalam dokumen. Perhitungan TF-IDF ditentukan pada Persamaan 1.

$$W_{ij} = tf_{ij} \times Idf_{ij}$$

$$Idf_{ij} = (\log(N/df)) \quad (1)$$

B. *Word Embedding GloVe*

Glove merupakan teknik penyisipan kata atau representasi kata yang digunakan untuk merepresentasikan vektor kata dari setiap dokumen. Tujuan dari proses representasi kata menggunakan metode GloVe adalah untuk mendapatkan hubungan semantik antar kata dari matriks *co-occurrence*. Menanamkan kata GloVe membentuk korpus dengan membuat matriks kemunculan bersama yang menunjukkan hubungan antara kata dan frekuensinya dalam dokumen. GloVe memudahkan untuk menangani sejumlah besar data pelatihan. Dalam penelitian ini, data yang digunakan adalah data sms spam dan non-spam.

2.3. Modelling

1) *Naïve Bayes*

Naïve Bayes merupakan salah satu metode klasifikasi yang mendasari ingeren dalam algoritma dengan melihat korelasi dari atribut data independent satu dengan lainnya. Secara semantik Naïve Bayes identik memiliki hubungan yang kuat antara kata yang satu dengan lainnya. Berikut persamaan Naïve Bayes

$$P(a_i = x_i | C_k) = (P = (x_i | C_k) = N_{ki}/N_k \quad (2)$$

2) *XGBoost*

XGBoost merupakan algoritma yang meminimalkan fungsi objektif yang diregularisasi dengan melakukan penggabungan fungsi untuk mendapatkan kompleksitas model yang efektif.

3) *BERT*

BERT merupakan sebuah metode deep learning yang telah memberikan hasil yang efektif pada berbagai penyelesaian dalam *Natural Language Processing (NLP)*. BERT memiliki

enam layer transformer yang dibentuk pada proses *encoding* dan *decoding* yang menyumbangkan file untuk proses *training* dengan kompleks, konfigurasi tinggi, dan waktu komputasi yang lama. BERT melakukan proses dengan sebuah kata yang direpresentasikan melalui *embedding* dengan perhitungan *multi-headed attention* pada setiap layer sebelumnya.

4) LSTM

LSTM merupakan sebuah metode deep learning yang menggunakan konsep tiga gerbang (*gate*) yakni gerbang masukan, gerbang *forget*, dan gerbang luaran. Salah satu komponen gerbang pada LSTM digunakan untuk melakukan control informasi yang masuk kedalam memori untuk menyelesaikan masalah gradien. Pengulangan koneksi dapat menyebabkan penambahan keadaan atau memori kedalam jaringan dan memungkinkan untuk memanfaatkan pengamatan yang terurut. Fungsi memori pada metode LSTM untuk menyajikan antrian luaran.

2.4. Evaluasi

Untuk melakukan evaluasi klasifikasi data diperlukan beberapa parameter yang sering digunakan yakni yakni nilai *accuracy*, *precision*, *recall*, dan *F1-Score* sebagai evaluator untuk mengetahui seberapa baik dan efektif sistem yang telah dijalankan. Berikut formulasi metrik untuk evaluasi klasifikasi data

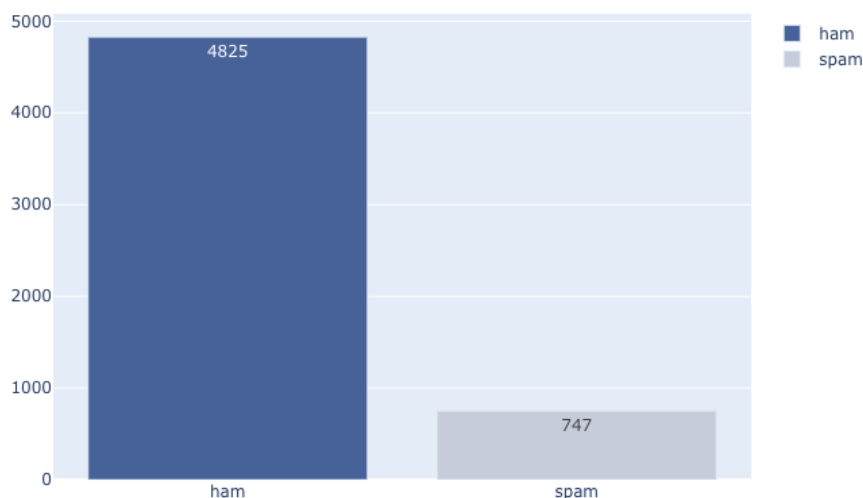
Tabel 2. Parameter untuk Evaluasi Data

No	Data	Model
1	<i>accuracy</i>	$(TP+TN)/(TP+FP+FN+TN)$
2	<i>precision</i>	$TP/(TP+FP)$
3	<i>recall</i>	$TP/(TP+FN)$
4	<i>F1-Score</i>	$2*(R*P)/(R+P)$

Performa sistem dievaluasi berdasarkan hasil eksperimen sistem analisis sentimen. Berdasarkan parameter evaluator tersebut nilai yang dihasilkan antara 0-1 atau 0-100%.

III. HASIL DAN PEMBAHASAN

Pada tahap hasil dan pembahasan penelitian ini menggunakan data 747 *spam* dan 482 *non-spam* yang divisualisasikan pada gambar 2. Seperti yang dapat dilihat pada Gambar 2, jumlah data tidak seimbang, jadi untuk selanjutnya dapat mengelola data dengan cara resampling.



Gambar 2. Dataset *spam* dan *non-spam* (ham)

Sekarang akan telah dilakukan rekayasa data untuk memudahkan perhitungan model pada proses klasifikasi data dengan melakukan preprocessing data. Hal pertama yang dilakukan adalah dengan membersihkan data yang dapat dilihat pada tabel 3 berikut ini

Tabel 3. Data yang sudah dibersihkan

no	target	message	message_len	Message_clean
0	ham	Go until jurong point, crazy.. Available only ...	20	go until jurong point crazy available only in ...
1	ham	Ok lar... Joking wif u oni...	6	ok lar joking wif u oni
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	28	free entry in a wkly comp to win fa cup final...
3	ham	U dun say so early hor... U c already then say...	11	u dun say so early hor u c already then say
4	ham	Nah I don't think he goes to usf, he lives aro...	13	nah i dont think he goes to usf he lives aroun...
...

Pada Tabel 3 menampilkan kategori target, pesan sms, jumlah setiap kata per sms, dan pembersihan data yang dibuat menjadi huruf kecil semua atau *lowercase*. Pada Tabel 3 didapatkan dari algoritma *stopwords* dengan menggunakan *stemming* atau *lemmatization* seperti berikut ini

Tabel 4. Stemming dan Lemmatization

No	Stemming	Lemmatization
0	Adjustable → adjust	Was → to be
1	Formality → formaliti	Better → good
2	Formality → formal	Meeting → meeting
...

Pada Tabel 4 telah dilakukan proses stemming dan lemmatization untuk pengubahan kata-kata dan menyamakan kata (sinonim) sehingga ketika sudah dilakukan cleaning, kata yang didapatkan menjadi tersusun sesuai dengan algoritma vektorisasi. Pada algoritma *steeming* terdapat tiga tahapan untuk mempercepat proses simplifikasi yaitu dengan *porterstemmer*, *snowballstemmer*, dan *lancastersetmmer*.

Selanjutnya disajikan proses pembobotan dengan TD-IDF yang divisualisasikan dalam bentuk array seperti Tabel 5 berikut ini

Tabel 5. Pembobotan dalam bentuk Array pada TD-IDF dan GloVe

TD-IDF	GloVe
array([[2, 3179, 274, ..., 0, 0, 0],	array([[0. , 0. , 0. , ... , 0.34658 , ... , 0.070204 , 0.44509 , 0.24147999]
[8, 236, 527, ..., 0, 0, 0],	[-0.57832998, -0.0036551 , ... , ...]
[9, 356, 588, ..., 0, 0, 0],	[-0.078894 , 0.46160001, ... , ...]
..., [6724, 1002, 6725, ..., 0, 0, 0],	[0.57779002, ..., 0.26352 , ... , ...]
[138, 1251, 1603, ..., 0, 0, 0],	[0.59397 , 0.26741001] , ... , ...]
[1986, 378, 170, ..., 0, 0, 0]], dtype=int32)	[-0.12002 , -1.23870003, -0.23303001, ..., 0.13658001, -0.61848003, 0.049843]

Tabel 5. Nilai Percobaan dengan Menggunakan Dataset Lain

No	Model	Accuracy	Precision	Recall	F1-Score
1	XGBoost	77,25%	75,67%	74,45%	76,12%
2	Glove-LSTM	80,56%	67,09%	86,38%	75,52%

Berdasarkan Tabel 4 dan Tabel 5 dengan menggunakan beberapa metode yang sama tetapi dataset yang berbeda tingkat akurasi nilai yang didapatkan sangat berbeda. Hal ini menunjukkan bahwa dataset sangat mempengaruhi bahwa hasil yang diperoleh dengan data yang berbeda tidak selalu sama dan memuaskan.

IV. KESIMPULAN

Pada penelitian klasifikasi data *spam* dan *non-spam* dari data sms dengan menggunakan bervariasi metode didapatkan hasil yang baik dan tingkat akurasi yang tinggi pada metode BERT, pada proses pengujian klasifikasi. Dengan menggunakan data training dan data test yang sudah ditentukan didapatkan tingkat akurasi nilai metode BERT sebesar 99,35%. Tetapi dilain hal, hasil penelitian dengan menggunakan dataset yang berbeda tetapi masih dalam kategori *text mining* atau NLP, kami mendapatkan hasil yang kurang memuaskan. Hal ini menunjukkan bahwa dataset sangat mempengaruhi hasil uji coba. Untuk kedepannya, kami akan melakukan ujicoba dengan menggunakan bervariasi data untuk menguji metode yang sudah kami jalankan saat ini. Ketertarikan kami akan NLP sangat besar, sehingga kedepannya kami membuka Kerjasama penelitian untuk mengembangkan pemilihan metode yang optimal dalam melakukan klasifikasi data.

UCAPAN TERIMA KASIH

Terima kasih kepada rekan-rekan yang telah membantu dalam menyelesaikan artikel ini, dalam bentuk akses dataset pada web site terbuka dan referensi algoritma dalam menyelesaikan dan menjawab permasalahan klasifikasi data.

DAFTAR PUSTAKA

1. Detiknet.com, 2020, Riset: Ada 175,2 Juta Pengguna Internet di Indonesia, <https://inet.detik.com/cyberlife/d-4907674/riset-ada-1752-juta-pengguna-internet-di-indonesia>.
2. Meng, X., Liu, M., Wu, Q., 2020, Prediction of Rice Yield via Stacked LSTM, International Journal of Agricultural and Environmental Information Systems (IJAEIS), No. 1, Vol. 11, Hal. 86–95
3. Reis, J., Benevenuto, F., Olmo, P., Prates, R., Kwak, H., An, J., 2015, Breaking the news: First impressions matter on online news, The 9th International Conference on Web and Social Media, oxford, 26-29 Mei
4. Zamahsyari., Nurwidyantoro, A., 2017, Sentiment analysis of economic news in Bahasa Indonesia using majority vote classifier, 2016 International Conference on Data and Software Engineering (ICoDSE 2016), Denpasar, 26 - 27 Oktober
5. Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., 2016, Structure of Social Media Conversations, arXiv preprint, arXiv:1609.09028
6. Hermanto, D.T., Setyanto, A., Lutfhi, E.T. Algoritma LSTM-CNN untuk Sentimen Klasifikasi dengan Word2vec pada Media Online. Citec Journal, Vol. 8, No. 1, Januari 2021. ISSN: 2354-5771.
7. Revi yanti, F. Azhar, Y., Marthasari, G.I. Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression. Repositor, Vol. 3, No. 4., Agustus 2021. Pp. 387-392.
8. Atmaja, R.M.R.W.P.K., Yustanti., W. Analisis Sentimen Customer Review Aplikasi Ruang Guru dengan Metode BERT (Bidirectional Encoder Representations from Transformers). JEISBI (Journal of Emerging Information System and Business Intelligence): Volume 2, Number 3, 2021.
9. Faturrohman, F., Rosmala, D. Analisis Sentimen Sosial Media dengan Metode Bidirectional Gated Recurrent Unit. Diseminasi-FTI.
10. Utami, L.D., Yusuf, L., dan Nurlela, D. Komparasi Algoritma Naïve Bayes dan Support Vectors Machine pada Analisis Sentimen SMS HAM dan SPAM. nforek : Jurnal Informatika dan Teknologi. Vol. 4 No. 2, Juli 2021 Hal. 249-258. DOI : 10.29408/jit.v4i2.3665.

Seminar Nasional Sains Data 2022 (SENADA 2022)

UPN "Veteran" Jawa Timur

11. Gifari, O.I., Adha, M., Hendrawan, I.R., Durrand, F.F.S. Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY). Vol. 2, No. 1, Maret 2022.
12. Putro, M.F.A., Setiawan., E.B. Analisis Sentimen Terhadap Kebijakan Pemerintah dengan Feature Expansion Metode GloVe pada Media sosial Twitter. e-Proceeding of Engineering : Vol.9, No.1 Februari 2022