MENORMALISASIKAN TEKS PADA BOT SISTEM INFORMASI AKADEMIK MENGGUNAKAN ALGORITMA *DAMERAU–LEVENSHTEIN DISTANCE DAN PREFIX TREE* (STUDI KASUS: UNIVERSITAS TEKNOKRAT INDONESIA)

Muhammad Thomas Fadhila Yahya[1], Sigit Doni Ramdan[2],
[1]Informatika
[2]Teknik Elektro
*) sigitpapazola@gmail.com

Abstrak

Penelitian ini dilakukan atas dasar permasalahan pada bot yang sulit memahami dan merespons pesan dengan tepat karena terdapat kesalahan pengetikan, tata bahasa dan penggunaan bahasa yang buruk dalam pesan pengguna. Sistem ini terdiri tujuh tahapan normalisasi yaitu normalisasi garis baru, normalisasi huruf kecil, normalisasi karakter berulang, normalisasi spasi, tokenisasi, normalisasi kata dasar, dan pengecekan ejaan. Berdasarkan hasil penelitian dan implementasi, diketahui bahwa menggunakan algoritma *Damerau–Levenshtein Distance* dan fungsi Perhitungan Kedekatan Huruf menghasilkan nilai mean *average precision* sebesar 0,86. Dan menggunakan *Prefix Tree* menghasilkan waktu proses sebesar 0.004 detik untuk kata dengan panjang mulai dari 3 karakter, bertambah 0.002 detik untuk setiap karakter.

Kata Kunci: *Chatbot*, Menormalisasikan, *Damerau–Levenshtein Distance*, *Prefix Tree*

PENDAHULUAN

(Damayanti & Sumiati, 2018), (Pasha, 2017), (Suryono et al., 2018)Dunia saat ini tengah memasuki era revolusi industri 4.0 atau revolusi industri dunia keempat di mana teknologi menjadi basis dalam setiap aspek kehidupan manusia. Aspek pendidikan pun tak luput dari pengaruh revolusi industri 4.0. Dunia pendidikan dituntut mengikuti perkembangan teknologi yang sedang berkembang saat ini dengan cara memanfaatkan teknologi informasi sebagai fasilitas dalam membantu proses belajar dan mengajar.

Dari tiga algoritma di atas, algoritma jarak string yang dipilih untuk normalisasi teks yang dapat memperbaiki kesalahan pengetikan adalah *Damerau-Levenshtein distance*. Dalam penelitian Raffael Vogler pada tahun 2013 yang melakukan pembandingan dengan sembilan algoritma jarak string dan menghasilkan kesimpulan bahwa algoritma yang cocok untuk menangani masalah kesalahan pengetikan adalah variasi Levenshtein adalah yang paling baik. Algoritma ini dinilai yang paling cocok karena dalam algoritma ini, penghitungan operasi yang diperlukan untuk mengubah string a menjadi string b lebih lengkap yaitu empat operasi (penyisipan, penghapusan, penggantian, dan transposisi) (Vogler, 2013) berbeda dengan *Hamming Distance* yang hanya menghitung operasi substitusi dan LCS hanya operasi penyisipan dan penghapusan (Borman, 2016), (Ahmad et al., 2021), (Napianto et al., 2017).

Oleh karena itu, untuk mengatasi permasalahan di atas, yaitu kesalahan pengetikan dan bentuk kata tidak baku dalam pesan percakapan pengguna bot, penulis melakukan penelitian "Menormalisasikan Teks Pada bot Sistem Informasi Akademik Menggunakan Algoritma *Damerau–Levenshtein Distance dan Prefix Tree* (Studi Kasus: Universitas Teknokrat Indonesia)" di mana nantinya diharapkan dapat menyelesaikan permasalahan dan meningkatkan chatbot dalam memahami pesan dan merespon dengan tepat pesan pengguna (Hana et al., 2019), (Suaidah & Sidni, 2018), (Rulyana & Borman, 2014).

KAJIAN PUSTAKA

Sub-bagian I

Oleh Tri Sony Saragih (2017) dari Departemen Ilmu Komputer Matematika Dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor dengan judul Normalisasi Teks pada *Twitter* Berbahasa Indonesia Menggunakan Algoritme Jarak String pada R. Latar belakang masalah dalam penelitian ini adalah data *tweet* sering menggunakan kata yang tidak baku sesuai bahasa Indonesia sehingga sulit digunakan untuk *text mining* (Bakri, 2017), (Nabila et al., 2021). (Darwis et al., 2017).

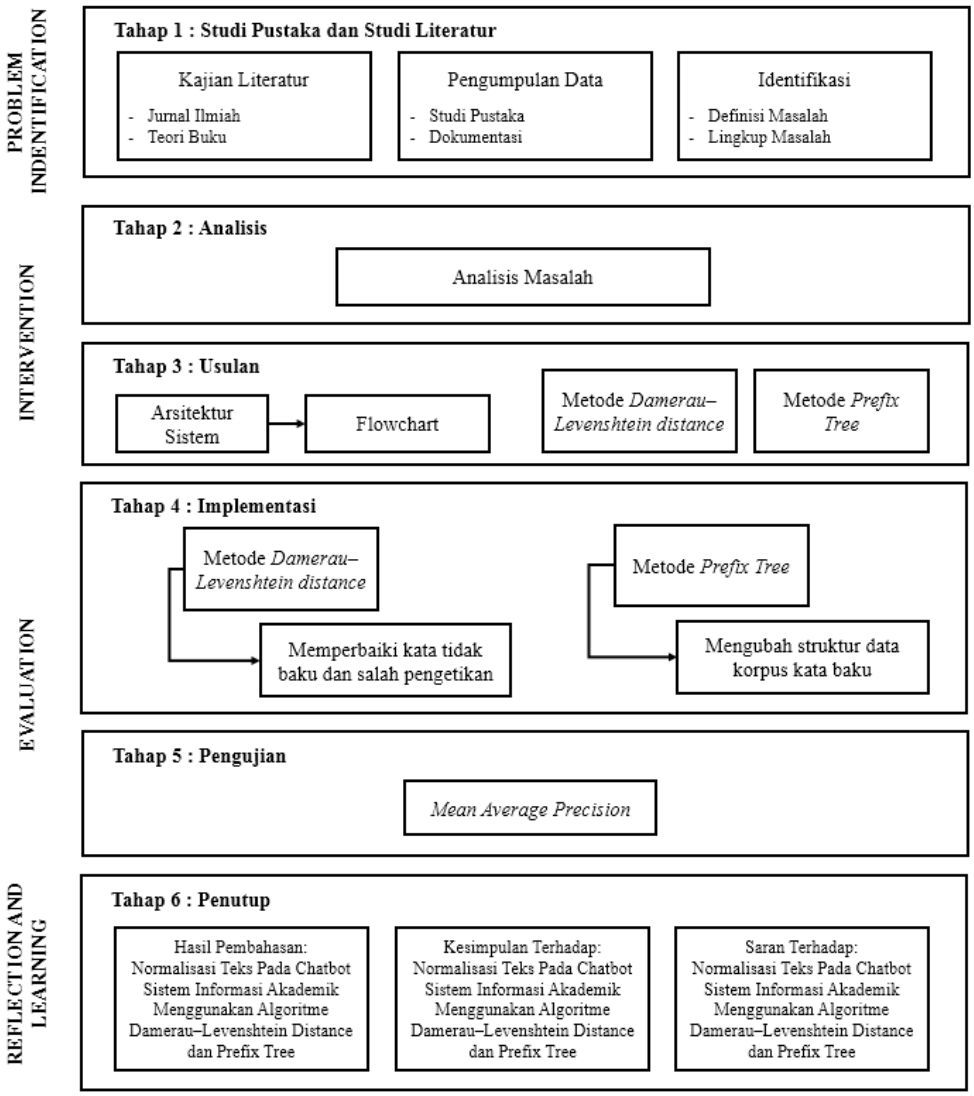
Oleh Imam Thoiba, Arief Setyantob, dan Suwanto Raharjoc (2018) dari Magister Teknik Informatika, Universitas Amikom Yogyakarta dengan judul Pengaruh Normalisasi Teks Dengan *Text Expansion* Dalam Deteksi Komentar Spam Pada Youtube. Di mana dalam penelitian yang dilakukan oleh penulis mengangkat masalah bagaimana komentar spam pada video-video yang ada di Youtube menjadi hal yang sangat meresahkan pemilik channel dan menjadi celah untuk mendapatkan keuntungan dari Youtube dengan cara yang ilegal (Wibisono et al., 2020), (Tenriawali, 2018), (Darwis, 2017).

Oleh Redmond Fileno A. Alleva, Issaquah Michael J. Rozak, Bellevue Larry J. Israel (1999) dari *Microsoft Corporation* dengan judul *Text Normalization Using Context-Free Grammar*. Latar belakang masalah dalam penelitian ini adalah pengenalan ucapan sering menghasilkan output teks yang tidak *familier* bagi pengguna. Dalam penelitian ini, penulis menggunakan metode *Context-Free Grammar* (Kuswoyo, 2016), (Ayu et al., 2021), (Kuswoyo & Susardi, 2016).

METODE

Tahapan penelitian merupakan kegiatan penelitian yang dilakukan secara terencana, teratur, dan sistematis untuk mencapai tujuan tertentu .

(Suryono et al., 2019), (Riskiono et al., 2018), (Romdhoni et al., 2012) Setelah melakukan pengumpulan data yang dibutuhkan dalam penelitian, tahap selanjutnya yaitu tahapan analisis masalah. Analisis masalah adalah proses identifikasi sebab dan akibat dari suatu masalah yang terjadi pada objek penelitian agar sistem yang akan diusulkan sesuai dengan tujuan dari penelitian dan dapat mengatasi permasalahan yang sudah diidentifikasi.



Gambar 1

	id ▲ 1	username	time	text
Delete	1301	ajeng	2019-07-03 16:34:36	jadwl mata kuliah matematika diskrit
Delete	1302	ajeng	2019-07-03 16:34:50	bagaimna?
Delete	1303	oci	2019-07-03 16:35:40	kmu siapa
Delete	1304	oci	2019-07-03 16:35:45	apa kabar
Delete	1305	oci	2019-07-03 16:35:49	slamat siang
Delete	1306	oci	2019-07-03 16:35:56	hai
Delete	1307	oci	2019-07-03 16:36:10	makan apa hri ini
Delete	1308	nadya	2019-07-03 17:24:54	kmu siapa
Delete	1309	nadya	2019-07-03 17:25:02	cuaca hari ini
Delete	1310	nadya	2019-07-03 17:25:12	hri ini hari apa
Delete	1311	nadya	2019-07-03 17:25:28	mulai

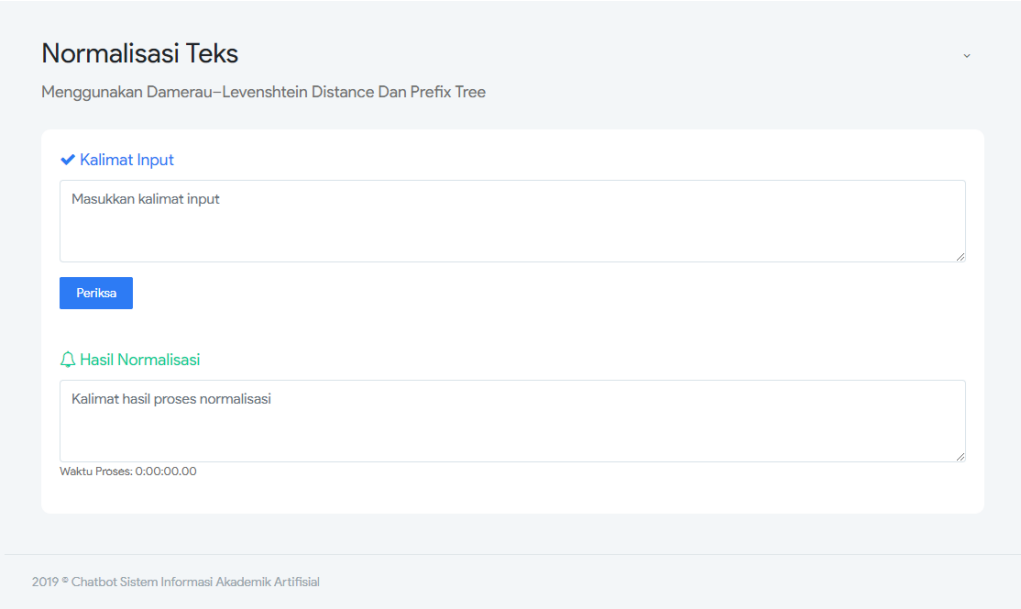
Gambar 2

Pengujian dilakukan untuk mengetahui apakah algoritma memiliki hasil perbaikan kata yang baik dan waktu proses yang cepat. Dalam penelitian ini, pengujian akan dilakukan dengan menghitung nilai precision dari setiap hasil yang nantinya akan menghasilkan nilai *average precision* dan *mean average precision*. Pengujian akan dilakukan sebanyak dua kali (Ashari, 2019), (Amanda, 2017), (Manalu & Setyadi, 2010).

HASIL DAN PEMBAHASAN

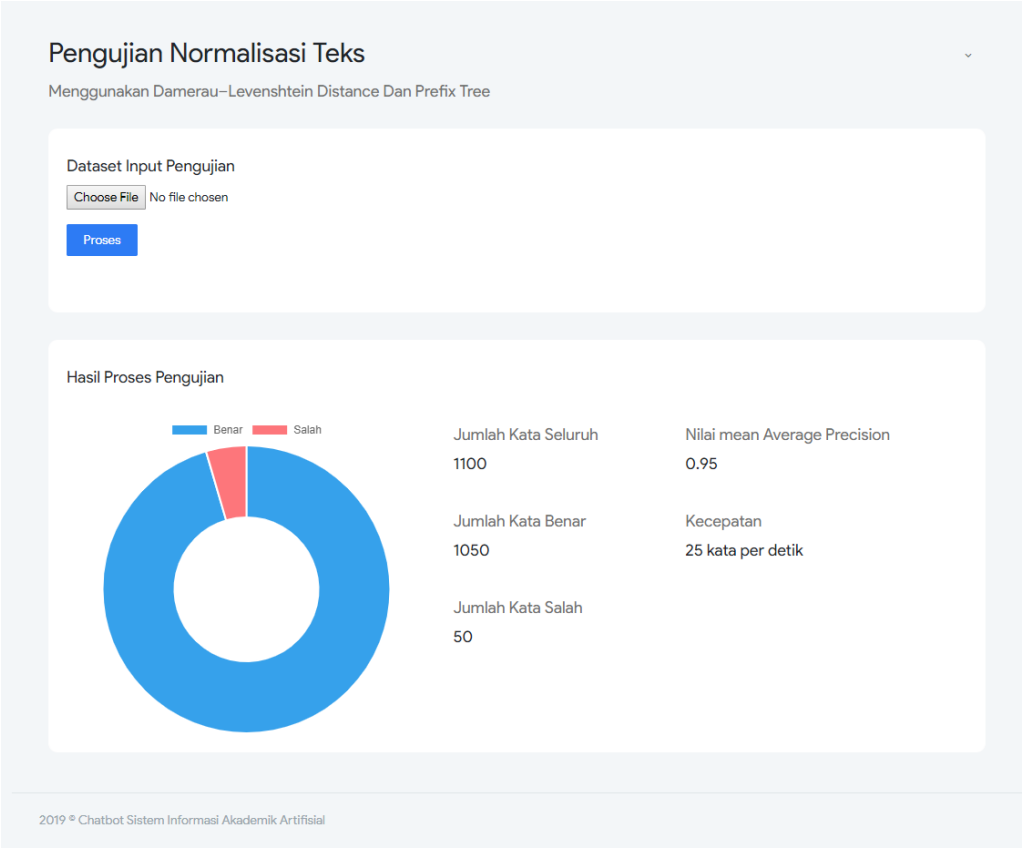
(Nurkholis et al., 2021), (Mandasari, 2017), (Susanto & Puspaningrum, 2020) Program normalisasi teks dibagi dalam dua bagian yaitu bagian *Front-End* dan *Back-End.* Bagian *Front-End* dibuat menggunakan bahasa PHP sedangkan untuk bagian *Back-End* dibuat menggunakan Bahasa *Python.*

(Permata et al., 2020), (Abidin et al., 2018), (Sintaro, 2020) Menu Normalisasi adalah tempat di mana memasukkan kalimat input yang ingin dinormalisasikan, terdapat dua *text box* yaitu *text box* untuk kalimat *input* dan *text box* untuk kalimat output hasil normalisasi serta satu buah tombol untuk melakukan proses normalisasi.



Gambar 3

Menu Pengujian adalah tempat di mana melakukan pengujian algoritme normalisasi teks. Proses pengujian dilakukan dengan memasukkan *file* data *input* berupa *file teks* dengan format “kata benar: kata salah 1, kata salah 2, kata salah n”. Proses pengujian dilakukan dengan menghitung nilai mean average precision dari seluruh proses normalisasi kata *input* yang dimasukkan.



Gambar 3

Dalam analisis hasil perbaikan kata salah yang dengan kondisi *False Positive* akan menggunakan hasil perbaikan kata dengan sumber kedua yaitu data log pesan SIAA Chatbot yang berjumlah 164 kata tidak baku berasal dari 50 pengguna. Daftar kata hasil perbaikan yang memenuhi kondisi False Positive disajikan.

No	Kata Uji	Perbaikan	Yang Diharapkan
1	cma	cema	**cuma**
2	brp	bos	**berapa**
3	brg	bos	**barang**
4	uas	buas	**uas**
5	soalnya	solnya	**soal**
6	kbar	akbar	**kabar**
7	dsn	dan	**di sana**

No	Kata Uji	Perbaikan	Yang Diharapkan
28	nanya	hanya	**tanya**
29	diem	dism	**diam**
30	kyanya	khan	**kayaknya**
31	msh	masah	**masih**
32	kasian	kasiat	**kasihan**
33	ttp	tatap	**tetap**
34	kmna	kuna	**ke mana**

8	sma	asma	**sama**
9	disana	disapa	**di sana**
10	ukm	ukoma	**ukm**
11	pastinya	patinya	**pasti**
12	verkas	cerkas	**berkas**
13	pke	ke	**pakai**
14	pkl	pakal	**pkl**
15	besk	besek	**besok**
16	septem	septet	**september**
17	udah	idah	**sudah**
18	disini	disigi	**di sini**
19	ayoklah	aduklah	**ayo**
20	nyampe	nyam	**sampai**
21	pake	pale	**pakai**
22	dmn	dina	**di mana**
23	kbr	akbar	**kabar**
24	rmah	ramah	**rumah**
25	udh	adeh	**sudah**
26	urt	urat	**urut**
27	tau	atau	**tahu**

35	gtu	getu	**gitu**
36	tauu	taju	**tahu**
37	disni	diisi	**di sini**
38	bln	blen	**bulan**
39	bli	bali	**beli**
40	brng	bung	**barang**
41	slmat	semat	**selamat**
42	pdes	pres	**pedas**
43	gmna	guna	**bagaimana**
44	kerjaan	kerajaan	**pekerjaan**
45	bener	beber	**benar**
46	makasih	makas	**terima kasih**
47	dapet	bapet	**dapat**
48	pingin	pingi	**ingin**
49	tlng	tong	**tolong**
50	cepet	ceper	**cepat**
51	pengen	penge	**ingin**
52	sdh	sadah	**sudah**
53	ngbrol	jebrol	**mengobrol**

Gambar 4

SIMPULAN

Berdasarkan rumusan masalah, hasil penelitian dan pembahasan terhadap normalisasi *teks* pada chatbot sistem informasi akademik menggunakan *algoritme Damerau–Levenshtein Distance* dan *Prefix Tree* dapat disimpulkan bahwa:

Ada tujuh tahap normalisasi dalam program normalisasi teks yaitu normalisasi garis baru, normalisasi huruf kecil, normalisasi karakter berulang, normalisasi spasi, tokenisasi, normalisasi kata dasar, dan pengecekan ejaan

Dari hasil pengujian, *algoritme Damerau–Levenshtein Distance* dengan fungsi Perhitungan Kedekatan Huruf menghasilkan nilai *mean average precision* sebesar 0,86 dengan nilai *average precision* tertinggi pada kata dengan 8 karakter atau lebih di atas 0,90 sedangkan tanpa fungsi Perhitungan Kedekatan Huruf menghasilkan nilai *mean average precision* sebesar 0,77

Data korpus yang berasal dari Kamus Besar Bahasa Indonesia berjumlah 59.093 kata dibentuk ke dalam bentuk *prefix tree* atau *trie* menjadi 161.828 *node*. Waktu proses yang dibutuhkan program normalisasi teks untuk melakukan perbaikan kata dengan data korpus berbentuk *prefix tree* sebesar 0.004 detik untuk kata dengan panjang mulai dari 3 karakter, bertambah 0.002 detik untuk setiap karakter. Untuk kata dengan panjang mulai dari 11 karakter, peningkatan waktu proses berubah menjadi 0.003 detik untuk setiap karakter. Hal ini lebih cepat dibandingkan dengan data korpus konvensional yaitu 0,2 detik untuk kata dengan 3 sampai 12 karakter.

REFERENSI

Abidin, Z., Sucipto, A., & Budiman, A. (2018). Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention Translation of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based. *J. Kelitbangan*, *6*(02), 191–206.

Ahmad, I., Borman, R. I., Caksana, G. G., & Fakhrurozi, J. (2021). IMPLEMENTASI STRING MATCHING DENGAN ALGORITMA BOYER-MOORE UNTUK MENENTUKAN TINGKAT KEMIRIPAN PADA PENGAJUAN JUDUL SKRIPSI/TA MAHASISWA (STUDI KASUS: UNIVERSITAS XYZ). *SINTECH (Science and Information Technology) Journal*, *4*(1), 53–58.

Amanda, D. (2017). *PENGUJIAN KEPUASAN SEBAGAI VARIABEL INTERVENING ANTARA PENGARUH KEPERCAYAAN DAN ATRIBUT PRODUK TABUNGAN BATARA IB TERHADAP LOYALITAS NASABAH (STUDI PADA PT. BANK TABUNGAN NEGARA (PERSERO) TBK, KANTOR CABANG SYARIAH PALEMBANG).[SKRIPSI]*. UIN RADEN FATAH PALEMBANG.

Ashari, D. P. (2019). *SISTEM PENDUKUNG KEPUTUSAN PENGUJIAN KELAYAKAN ANGKUTAN UMUM MENGGUNAKAN METODE ANALITYCAL HIERARCHY PROCESS (Decision Support System For Testing Feasibility Of Public Transport Using Analitycal Hierarchy Process Method)*. Universitas Teknokrat Indonesia.

Ayu, M., Sari, F. M., & Muhaqiqin, M. (2021). Pelatihan Guru dalam Penggunaan Website Grammar Sebagai Media Pembelajaran selama Pandemi. *Al-Mu'awanah: Jurnal Pengabdian Kepada Masyarakat*, *2*(1), 49–55.

Bakri, M. (2017). Penerapan Data Mining untuk Clustering Kualitas Batu Bara dalam Proses Pembakaran di PLTU Sebalang Menggunakan Metode K-Means. *Vol*, *11*, 1–4.

Borman, R. I. (2016). Penerapan String Matching Dengan Algoritma Boyer Moore Pada Aplikasi Font Italic Untuk Deteksi Kata Asing. *Jurnal Teknoinfo*, *10*(2), 39–43.

Damayanti, D., & Sumiati, S. (2018). Sistem Informasi Daya Tarik Pembelian Produk UMKM Home Industri Berbasis WEB. *Konferensi Nasional Sistem Informasi (KNSI) 2018*.

Darwis, D. (2017). Teknik Steganografi untuk Penyembunyian Pesan Teks Menggunakan Algoritma GIFSHUFFLE. *Jurnal Teknoinfo*, *11*(1), 19–24.

Darwis, D., Wamiliana, W., & Junaidi, A. (2017). Proses Pengamanan Data Menggunakan Kombinasi Metode Kriptografi Data Encryption Standard dan Steganografi End Of File. *Prosiding Seminar Nasional METODE KUANTITATIF 2017*, *1*(1), 228–240.

Hana, P., Rusliyawati, R., & Damayanti, D. (2019). Pengaruh Media Richness Dan Frequently Update Terhadap Loyali Tas Civitas Akademika Perguruan Tinggi. *Jurnal Tekno Kompak*, *13*(2), 7–10.

Kuswoyo, H. (2016). Thematic structure in Barack Obama's press conference: A systemic functional grammar study. *Advances in Language and Literary Studies*, *7*(2), 257–267.

Kuswoyo, H., & Susardi, S. (2016). Thematic progression in EFL students' academic writings: A systemic functional grammar study. *Teknosastik*, *14*(2), 39–45.

Manalu, N. J., & Setyadi, M. A. (2010). Analisa Nilai Guna Teknologi Informasi Dalam Perbaikan Proses Penyediaan Barang Pada PT Xyz. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*.

Mandasari, B. (2017). *Role Playing Activity in English for Business Class for Non-English Study Program*.

Nabila, Z., Isnain, A. R., Permata, P., & Abidin, Z. (2021). ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS. *Jurnal Teknologi Dan Sistem Informasi*, *2*(2), 100–108.

Napianto, R., Utami, E., & Sudarmawan, S. (2017). VIRTUAL PRIVATE NETWORK (VPN) PADA SISTEM OPERASI WINDOWS SERVER SEBAGAI SISTEM PENGIRIMAN DATA PERUSAHAAN MELALUI JARINGAN PUBLIK (STUDI KASUS: JARINGAN TOMATO DIGITAL PRINTING). *Respati*, *7*(20).

Nurkholis, A., Susanto, E. R., & Wijaya, S. (2021). Penerapan Extreme Programming dalam Pengembangan Sistem Informasi Manajemen Pelayanan Publik. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *5*(1), 124–134.

Pasha, D. (2017). *Pengembangan Model Rantai Pasok Industri CPO Untuk Meningkatkan Produktifitas Dan Efisiensi Rantai Pasok Menggunakan Sistem Dinamik (Studi Kasus: Minyak Goreng di PT Tunas Baru Lampung)*. Institut Teknologi Sepuluh Nopember.

Permata, P., Abidin, Z., & Ariyani, F. (2020). Efek Peningkatan Jumlah Paralel Korpus Pada Penerjemahan Kalimat Bahasa Indonesia ke Bahasa Lampung Dialek Api. *Jurnal Komputasi*, *8*(2), 41–49.

Riskiono, S. D., Pasha, D., & Trianto, M. (2018). Analisis Kinerja Metode Routing OSPF dan RIP Pada Model Arsitektur Jaringan di SMKN XYZ. *SEMNASTEKNOMEDIA ONLINE*, *6*(1), 1.

Romdhoni, A. H., Tho'in, M., & Wahyudi, A. (2012). Sistem Ekonomi Perbankan Berlandaskan Bunga (Analisis Perdebatan Bunga Bank Termasuk Riba Atau Tidak). *Jurnal Akuntansi Dan Pajak*, *13*(01).

Rulyana, D., & Borman, R. I. (2014). Aplikasi Simulasi Tes Potensi Akademik Berbasis Mobile Platform Android. *Seminar Nasional FMIPA-Universitas Terbuka. DKI Jakarta*.

Sintaro, S. (2020). RANCANG BANGUN GAME EDUKASI TEMPAT BERSEJARAH DI INDONESIA. *Jurnal Informatika Dan Rekayasa Perangkat Lunak*, *1*(1), 51–57.

Suaidah, S., & Sidni, I. (2018). Perancangan Monitoring Prestasi Akademik dan Aktivitas Siswa Menggunakan Pendekatan Key Performance Indicator (Studi Kasus SMA N 1 Kalirejo). *Jurnal Tekno Kompak*, *12*(2), 62–67.

Suryono, R. R., Darwis, D., & Gunawan, S. I. (2018). Audit Tata Kelola Teknologi Informasi Menggunakan Framework Cobit 5 (Studi Kasus: Balai Besar Perikanan Budidaya Laut Lampung). *Jurnal Teknoinfo*, *12*(1), 16–22.

Suryono, R. R., Nurhuda, Y. A., & Ridwan, M. (2019). Analisis Perilaku Pengguna Sistem Informasi Pengetahuan Obat Buatan Untuk Kebutuhan Swamedikasi. *Jurnal Teknoinfo*, *13*(1), 1–4.

Susanto, E. R., & Puspaningrum, A. S. (2020). Model Prioritas Program Pemerataan Ipm Di Provinsi Lampung Menggunakan Metode Analytic Hierarchy Process. *Jurnal Teknoinfo*, *14*(1), 9–14.

Tenriawali, A. Y. (2018). Representasi Korban Kekerasan dalam Teks Berita Daring Tribun Timur: Analisis Wacana Kritis. *Jurnal Totobuang*, *6*(1), 1–15.

Wibisono, A. D., Rizkiono, S. D., & Wantoro, A. (2020). Filtering Spam Email Menggunakan Metode Naive Bayes. *Telefortech: Journal Of Telematics And Information Technology*, *1*(1), 9–17.