



**OPTIMALISASI PRA-PEMROSESAN TEKS DAN
PENGEMBANGAN KORPUS BAHASA INDONESIA UNTUK
MENINGKATKAN KINERJA ANALISIS SENTIMEN PADA
APLIKASI PLN MOBILE**

SEMINAR BIDANG KAJIAN

YESSY ASRI

99223114

ANGKATAN 31 S3 TI

**PROGRAM DOKTOR TEKNOLOGI INFORMASI UNIVERSITAS
GUNADARMA
JUNI 2024**

DAFTAR ISI

Halaman

DAFTAR ISI	ii
1 PENDAHULUAN	3
1.1 Latar Belakang.....	3
1.2 Batasan dan Tujuan	6
1.3 Kontribusi	6
2 TINJAUAN PUSTAKA	7
2.1 <i>PLN Mobile</i>	7
2.2 <i>Deep Learning</i>	8
2.3 Perangkat Natural Language Processing (NLP).....	8
2.4 Analisis Sentimen	10
2.5 <i>Preprocessing</i>	14
2.6 <i>Corpus</i>	15
2.7 Perbandingan Tinjauan	18
3 METODE PENELITIAN.....	31
3.1 Motivasi	31
3.2 Framework Riset.....	32
3.3 Pendekatan.....	34
DAFTAR PUSTAKA.....	36

1 PENDAHULUAN

1.1 Latar Belakang

Consumer Satisfaction atau kepuasan pelanggan adalah indikator penting dalam pengembangan suatu usaha baik untuk berskala besar maupun berskala kecil termasuk UMKM. *Consumer Satisfaction* mengukur bagaimana produk atau jasa yang diberikan mampu memenuhi atau melampaui ekspektasi pelanggan. Semakin tinggi angka *consumer satisfaction* mengindikasikan semakin baik kualitas produk/jasa/layanan yang diberikan perusahaan/UMKM. Pendekatan yang umum digunakan untuk menentukan *consumer satisfaction* adalah dengan Teknik penelitian berbasis survei, seperti yang dilakukan oleh [1][2].

Di era disruptif industri 4.0 dimana banyak aktifitas ekonomi dilakukan berbasis platform digital, teknik penelitian survey dianggap memiliki beberapa kelemahan, yaitu membutuhkan banyak waktu, biaya dan tenaga. Hal ini tentu kurang selaras dengan karakteristik model bisnis di era disruptif yaitu cepat dan mudah dalam layanan. Beberapa penelitian pendahuluan sebenarnya telah melakukan pengembangan teknik ekstraksi *consumer satisfaction* berbasis *big data*. Sayang metode-metode yang dikembangkan umumnya untuk dataset yang berbahasa Inggris, sehingga Bahasa Inggris kaya akan repository dan library. Penelitian yang mengembangkan repository Bahasa Indonesia jumlahnya masih sangat terbatas, bahkan untuk *consumer satisfaction extractor*.

Bahasa Indonesia memiliki aturan ejaan yang kompleks, seperti perubahan akhiran kata dan penggunaan huruf kapital. Kerumitan ini memerlukan pemeriksa ejaan untuk memastikan penggunaan bahasa yang standar dan benar, karena kesalahan pengejaan dapat berdampak signifikan pada pemahaman [3].

Menghadapi aturan ejaan yang sering membingungkan, pemeriksa ejaan otomatis dapat secara substansial meningkatkan kualitas tulisan dalam bahasa Indonesia. Dengan kemajuan teknologi, ketergantungan pada pemeriksa ejaan otomatis meningkat, terutama dalam aplikasi pengolah kata dan pemrosesan bahasa alami [4]. Meski begitu, pemeriksa ejaan untuk bahasa Indonesia masih mengalami kekurangan dalam mendeteksi dan mengoreksi kesalahan pengejaan, belum ada

tinjauan literatur yang luas dalam hal pemeriksa ejaan untuk bahasa Indonesia yang telah selesai [5].

Untuk performa optimal, pembelajaran membutuhkan data pelatihan yang bersih dan terstruktur. Namun, data bahasa alami sering kali tidak terstruktur dan penuh dengan *noise*, seperti tanda baca dan kata-kata yang tidak relevan. Oleh karena itu, penting untuk menerapkan teknik *preprocessing* teks sebelum diproses [6].

Dalam penelitian ini, teknik *preprocessing* teks digunakan untuk mengatasi sifat data teks yang tidak beraturan, dengan tujuan mencapai bias rendah dan akurasi tinggi dalam pemeriksaan ejaan. Teknik ini meliputi tokenisasi, penghilangan *stopword*, *spelling corrector*, *stemming*, dan *lemmatization*. *Tokenisasi* mengubah kalimat menjadi kumpulan token atau kata, sedangkan penghapusan *stopword* menghilangkan kata-kata yang tidak memberikan nilai prediktif. *Spelling corrector* adalah merupakan proses mendekripsi, mengoreksi, dan memberikan saran kata untuk kata-kata yang mengalami kesalahan pada ejaan di dalam suatu teks. *Stemming* menghilangkan awalan atau akhiran untuk mengembalikan kata ke bentuk dasarnya, dan *lemmatization* mengubah kata menjadi bentuk dasarnya untuk membakukan variasi kata dengan arti yang sama [7], [8], [9]. Kualitas model akhir sangat dipengaruhi oleh teknik *preprocessing* yang digunakan [10], [11] dan korpus yang digunakan [12].

Aplikasi *PLN Mobile* adalah aplikasi layanan PLN yang terdaftar di *Google Play Store*. Melalui aplikasi *PLN Mobile*, pelanggan banyak mendapatkan kemudahan diantaranya mengetahui berbagai info mulai dari transaksi token, lokasi pembayaran, melalui banking terdekat, tagihan rekening listrik dan riwayat pemakaian Kwh listrik. Aplikasi *PLN Mobile* juga menyediakan info status atau progress dari layanan pengaduan dan keluhan dari pelanggan kepada pihak PLN melalui *smartphone* yang sekaligus terhubung dengan Aplikasi Pengaduan dan Keluhan Terpadu (APKT) milik PLN. Pada sistem APKT ini tersimpan seluruh pengaduan dan keluhan yang sudah disampaikan oleh pelanggan dan melalui sistem APTK pihak PLN yang akan memberikan info atau progress atas permohonan dan pengaduan yang sudah disampaikan pelanggan (<https://web.pln.co.id/media/siaran-pers/2022/01/lampaui-target-2021-aplikasi-pln-mobile-diunduh-162-juta>).

Google Play atau *Google Play Store* adalah layanan penyedia konten digital Google yang menawarkan berbagai toko produk online untuk aplikasi, game, film, musik dan buku. *Google Play Store* dapat diakses melalui situs web, aplikasi Android, dan Google TV. Terdapat beberapa fitur di *Google Play Store*, salah satunya adalah fitur *rating* dan *review* dari pengguna aplikasi atau layanan yang tersedia. *Review* adalah teks atau kalimat yang berisi evaluasi atau komentar terhadap karya seseorang. Pentingnya ulasan ini sering digunakan sebagai standar untuk aplikasi *PLN Mobile*. Sejumlah perusahaan, termasuk PT Perusahaan Listrik Negara, membuat aplikasi yang memberikan kemudahan informasi kepada pelanggannya. Salah satunya yang baru diperkenalkan pada 20 Desember 2020. *PLN Mobile* telah diunduh lebih dari 10 juta kali dan memiliki rating 4,7 dari 5,0 di situs *Google Play* per 7 Juni 2022. Di kolom komentar ulasan pengguna *PLN Mobile* di situs *Google Play*, juga tercatat 293.519 ulasan pengguna. Aplikasi *PLN Mobile* merupakan aplikasi digital yang dikembangkan oleh PT PLN (Persero) dengan tujuan untuk memberikan pelayanan kelistrikan melalui aplikasi mobile. Pembayaran tagihan listrik, pembelian token, pencatatan nomor meter mandiri, penambahan daya, pengaduan dan pengaduan, monitoring pembelian token, monitoring pemakaian listrik pasca bayar, notifikasi tagihan, notifikasi pemadaman, informasi progres troubleshooting, dan perawatan jaringan listrik semuanya tersedia melalui aplikasi *PLN Mobile* [12]

Berdasarkan uraian di atas, penelitian ini akan merancang dan mengembangkan teknik baru untuk ekstraksi data kepuasan pelanggan dari platform digital menggunakan big data yang sesuai dengan karakteristik Bahasa Indonesia. Membangun repository dan library khusus Bahasa Indonesia yang dapat digunakan untuk analisis kepuasan pelanggan, mengatasi keterbatasan yang ada dibandingkan dengan bahasa Inggris. Mengembangkan algoritma pemeriksa ejaan otomatis yang lebih efektif untuk Bahasa Indonesia, guna meningkatkan kualitas data teks yang digunakan dalam penelitian. Mengaplikasikan teknik *preprocessing* teks seperti *tokenisasi*, penghilangan *stopword*, *spelling corrector*, *stemming*, dan *lemmatization* untuk meningkatkan akurasi dan keandalan model ekstraksi data. Menguji teknik ekstraksi yang dikembangkan dengan menggunakan ulasan pengguna dari aplikasi

PLN Mobile sebagai studi kasus, untuk mengukur kepuasan pelanggan dan menilai efektivitas teknik yang digunakan.

1.2 Batasan dan Tujuan

Batasan masalah pada penulisan ini adalah penelitian ini fokus pada pengembangan teknik ekstraksi data kepuasan pelanggan dalam bahasa Indonesia, sehingga tidak mencakup bahasa lain. Data yang digunakan dalam penelitian ini diambil dari ulasan pengguna aplikasi *PLN Mobile* di *Google Play Store*. Tahap *pre-processing* yang digunakan adalah *casefolding, filtering, spelling corrector, stopword removal, handling emojis, tokenizing, stemming dan lemmatization*. Tujuan dari penelitian ini adalah mengembangkan teknik baru untuk ekstraksi data kepuasan pelanggan dari platform digital menggunakan big data yang sesuai dengan karakteristik bahasa Indonesia, membangun repository dan library khusus bahasa Indonesia untuk analisis kepuasan pelanggan *PLN Mobile*, mengembangkan algoritma pemeriksa ejaan otomatis yang lebih efektif untuk bahasa Indonesia guna meningkatkan kualitas data teks yang digunakan dalam penelitian, mengaplikasikan teknik preprocessing teks seperti *tokenisasi, penghilangan stopword, spelling corrector, stemming, dan lemmatization* untuk meningkatkan akurasi dan keandalan model ekstraksi data serta menguji teknik ekstraksi yang dikembangkan menggunakan ulasan pengguna dari aplikasi *PLN Mobile* sebagai studi kasus untuk mengukur kepuasan pelanggan dan menilai efektivitas teknik yang digunakan.

1.3 Kontribusi

Dari segi keilmuan, usulan penelitian ini memberikan kontribusi keilmuan di bidang analisis kepuasan pelanggan berbasis platform digital, khususnya untuk Bahasa Indonesia, hasil penelitian ini diharapkan dapat memberikan solusi yang lebih efisien dan efektif dalam mengukur kepuasan pelanggan, khususnya dalam konteks bisnis digital di era Industri 4.0.

2 TINJAUAN PUSTAKA

Bab ini merupakan studi literatur tentang materi – materi yang berhubungan dengan proses penelitian.

2.1 *PLN Mobile*

Aplikasi *PLN Mobile* adalah bagian dari upaya transformasi yang dilakukan oleh PLN untuk meningkatkan pelayanan kepada pelanggan (Santikaaristi, 2022, Layanan Kelistrikan Kian Mudah dan Cepat, Ini Kata Pelanggan PLN Tentang PLN Mobile. PT PLN Persero).

Aplikasi PLN Mobile pertama kali diluncurkan pada tahun 2016 dan pada 2020 diluncurkan kembali sebagai New PLN Mobile yang memiliki fitur-fitur baru. Tujuan dari aplikasi ini adalah untuk memudahkan pelanggan dalam bertransaksi soal kelistrikan dan menyampaikan pengaduan.

Adapun fitur yang saat ini dapat digunakan oleh pengguna adalah sebagai berikut :

1. Kemudahan Pembelian Token & Pembayaran Tagihan
2. Kemudahan Ubah Daya
3. Catat Angka Meter Mandiri
4. Kemudahan Pengaduan Gangguan & Keluhan
5. Kemudahan Memonitor Pemakaian Listrik Pascabayar
6. Kemudahan Memonitor Pembelian Token
7. Notifikasi Tagihan
8. Informasi Progress Penyelesaian Gangguan
9. Notifikasi Padam & Pemeliharaan

Pada September 2023 disampaikan bahwa PLN Mobile mewujudkan transformasi digital menjadi salah satu aplikasi perusahaan pelayanan publik terbaik di Asia dengan hampir 44 juta pengunduh dan mencapai rating 4,8 dari skala 5 (Siaran Pers, 2023). Per tanggal 23 April 2024 diperoleh dari Google Play Store terdapat 841.250 pengguna memberikan ulasan dan mencapai rating sebesar 4,8.

2.2 Deep Learning

Deep learning merupakan subbidang *machine learning* yang algoritmanya terinspirasi dari struktur otak manusia. Struktur tersebut dinamakan *Artificial Neural Networks* atau disingkat ANN. Pada dasarnya, ia merupakan jaringan saraf yang memiliki tiga atau lebih lapisan ANN. Ia mampu belajar dan beradaptasi terhadap sejumlah besar data serta menyelesaikan berbagai permasalahan yang sulit diselesaikan dengan algoritma *machine learning* lainnya.

Deep learning terdiri dari beberapa jaringan saraf tiruan yang saling berhubungan. Adapun beberapa algoritmanya diantaranya *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN), *Long Short-Term Memory Network* (LSTM), dan *Self Organizing Maps* (SOM). *Deep Learning* dapat digunakan untuk memproses *unstructured data* seperti teks dan gambar, kemudian juga dapat mengotomatisasi proses ekstraksi fitur tanpa perlu melakukan proses pelabelan secara manual, memberikan hasil akhir yang berkualitas, mengurangi biaya operasional, bahkan melakukan manipulasi data yang lebih efektif. Adapun penerapannya bisa digunakan untuk pengenalan gambar, suara, *Natural language processing*, dan deteksi anomali (Dicoding, 2021).

2.3 Perangkat Natural Language Processing (NLP)

Natural Language Processing (NLP) atau pemrosesan bahasa alami adalah cabang ilmu komputer yang fokus pada pemahaman bahasa manusia oleh mesin. Tujuan utama dari NLP adalah untuk membuat komputer mampu memproses, menganalisis, dan memahami bahasa manusia dalam bentuk yang sama seperti manusia. NLP dapat diterapkan dalam berbagai bidang seperti pengenalan suara, terjemahan mesin, chatbot, analisis sentimen, dan banyak lagi [13].

Natural language processing (NLP) mengacu pada cabang ilmu komputer dan lebih khususnya, cabang kecerdasan buatan atau AI yang berkaitan dengan memberikan kemampuan pada komputer untuk memahami teks dan kata-kata yang diucapkan dengan cara yang mirip dengan manusia. NLP menggabungkan linguistik komputasional modeling berbasis aturan dari bahasa manusia dengan model

statistik, pembelajaran mesin, dan deep learning. Bersama-sama, teknologi ini memungkinkan komputer untuk memproses bahasa manusia dalam bentuk data teks atau suara dan ‘memahami’ makna lengkapnya, termasuk maksud dan perasaan pembicara atau penulis.

NLP menggerakkan program komputer yang menerjemahkan teks dari satu bahasa ke bahasa lain, merespons perintah suara, dan merangkum volume besar teks dengan cepat bahkan secara real time. Kemungkinan besar peneliti telah berinteraksi dengan NLP dalam bentuk sistem GPS yang dioperasikan suara, asisten digital, perangkat lunak dikte suara ke teks, chatbot layanan pelanggan, dan kenyamanan konsumen lainnya. Namun, NLP juga memainkan peran yang semakin penting dalam solusi enterprise yang membantu menyederhanakan operasi bisnis, meningkatkan produktivitas karyawan, dan menyederhanakan proses bisnis yang kritis. NLP memiliki beberapa sub-bidang, yaitu (IBM, 2021).

1. *Speech Recognition*

Merupakan sub-bidang NLP yang berfokus pada pengenalan suara manusia oleh komputer. Teknologi speech recognition digunakan dalam aplikasi seperti asisten suara, sistem transkripsi, dan pengenalan ucapan.

2. *Text Mining*

Sub-bidang NLP yang berfokus pada ekstraksi informasi yang berguna dari teks yang ada. Teknik text mining digunakan dalam aplikasi seperti analisis sentimen, klasifikasi teks, dan pengenalan entitas.

3. *Machine Translation*

Sub-bidang NLP yang berfokus pada penerjemahan otomatis dari satu bahasa ke bahasa lain. Teknologi machine translation digunakan dalam aplikasi seperti *Google Translate* dan *Microsoft Translator*.

4. *Named Entity Recognition*

Sub-bidang NLP yang berfokus pada pengenalan dan klasifikasi entitas dalam teks, seperti orang, tempat, dan organisasi. Teknik named *entity recognition* digunakan dalam aplikasi seperti pemrosesan dokumen dan analisis data.

5. *Sentiment Analysis*

Sub-bidang NLP yang berfokus pada analisis dan klasifikasi sentimen dalam teks.

Teknik sentimen analysis digunakan dalam aplikasi seperti analisis sosial media dan penilaian produk.

6. *Text Summarization*

Sub-bidang NLP yang berfokus pada pembuatan ringkasan teks yang dapat dipahami oleh manusia. Teknik text *summarization* digunakan dalam aplikasi seperti pengecekan plagiarisme dan pembuatan berita otomatis.

7. *Natural Language Generation*

Sub-bidang NLP yang berfokus pada pembuatan teks otomatis yang dapat dipahami oleh manusia. Teknik natural language generation digunakan dalam aplikasi seperti pembuatan laporan otomatis dan penghasilan konten web, dan beberapa sub bidang lainnya.

2.4 Analisis Sentimen

Deep learning adalah sebuah *artificial intelligence* yang dapat meniru proses kerja otak manusia. *Deep learning* sangat efektif untuk mengolah data mentah dan menciptakan pola untuk keperluan pengambilan keputusan. *Deep learning* sendiri merupakan bagian dari *machine learning* yang memiliki jaringan tersendiri. *Deep learning* juga mampu mengenali pola dan informasi tanpa pengawasan dari data yang tidak terstruktur atau tidak berlabel. Salah satu bagian dari *deep learning* ini yaitu *Sentiment Analysis* (Analisis Sentimen).

Sentiment analysis atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan text mining yang bertujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik , produk, layanan, organisasi, individu, ataupun kegiatan tertentu [14].

Ekspresi atau sentimen mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subject yang berbeda. Oleh karena itu pada beberapa penelitian, terutama pada review

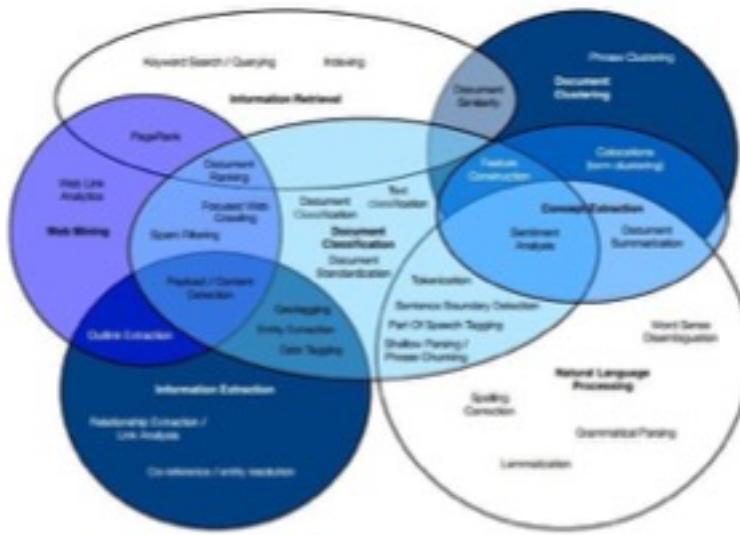
produk, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining*.

Dari kedua penjelasan di atas, kesimpulannya adalah analisis sentimen merupakan sebuah proses untuk memilih sentimen atau pendapat dari seseorang yang diimplementasikan dalam bentuk teks dan bisa dikelompokkan sebagai sentimen *positive* atau *negative* ataupun *neutral*.

Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral. Sentimen analisis juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Kita dapat mencari pendapat tentang produk-produk, merek atau orang-orang dan menentukan apakah mereka dilihat positif atau negatif di web [15].

Fungsi utama dari analisis sentimen pada media sosial adalah untuk menggali informasi, mencari makna, serta opini pengguna dari data posting, baik berupa caption, post, komentar, likes yang diunggah. Dalam bisnis, penggunaan analisis opini dan sentimen sangat dibutuhkan dalam *role business intelligence* untuk mengetahui pendapat pelanggan mengenai produk atau jasa yang ditawarkan, sehingga dapat digunakan sebagai pertimbangan dalam mengembangkan produk, contohnya seperti “ah layanan disini tidak sebagus layanan yang disana” kalimat tersebut merupakan contoh kalimat sentiment negative dalam dunia bisnis. Tidak hanya itu, analisis sentimen juga dapat membantu dalam menganalisa produk kompetitor. Dalam sistem pemerintahan, analisis sentimen dapat membantu pemerintah untuk mengatahui opini yang dikeluarkan masyarakat mengenai kebijakan yang dikeluarkan oleh pemerintah, serta memungkinkan menangkap data secara real-time untuk melihat perubahan preferensi publik dan menggabungkannya dengan sentimen yang dikeluarkan sebelumnya.

Analisis sentimen termasuk kedalam bagian Text Mining, yaitu irisan dari *concept extraction*, *document classification* dan *Natural Language Processing* (NLP) pada klasifikasi sentimen [16]. Pekerjaan text mining dikelompokkan menjadi tujuh daerah praktik yang diilustrasikan pada Gambar 2.1 berikut.



Gambar 1. Diagram Venn Text Mining

Diagram Venn Text Mining adalah diagram yang menggambarkan hubungan antara tiga konsep utama dalam text mining, yaitu "*Text Preprocessing*", "*Text Analysis*", dan "*Text Representation*". Diagram ini digunakan untuk memperjelas bagaimana ketiga konsep ini saling terkait dan saling mempengaruhi dalam proses text mining.

1. Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan penggalian dokumen teks misalnya dalam mesin pencarian (*search engine*) dan pencarian kata kunci (*keywords*).
2. Pengelompokan dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klaster (*clustering*) data mining.
3. Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klasifikasi (*classification*) data mining berdasarkan model terlatih yang sudah memiliki label.
4. *Web mining*, yaitu penggalian informasi dari internet dengan skala fokus yang spesifik.
5. Ekstraksi informasi (*information extraction*), yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.

6. *Natural language processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.
7. Ekstraksi konsep, yaitu pengelompokan kata atau frase ke dalam kelompok yang mirip secara semantik.

Analisis sentimen memiliki beberapa bagian atau sub-bidang yaitu:

1. *Polarity Detection*

Sub-bidang *Sentiment Analysis* yang berfokus pada deteksi polaritas dalam teks. Teknik ini digunakan untuk menentukan apakah sentimen dalam teks positif, negatif, atau netral.

2. *Emotion Detection*

Sub-bidang *Sentiment Analysis* yang berfokus pada deteksi emosi dalam teks. Teknik ini digunakan untuk mengidentifikasi emosi seperti kegembiraan, kecemasan, atau kemarahan yang terkandung dalam teks.

3. *Aspect-Based Sentiment Analysis*

Sub-bidang *Sentiment Analysis* yang berfokus pada analisis sentimen terhadap aspek tertentu dari suatu entitas. Teknik ini digunakan untuk menentukan sentimen terhadap aspek tertentu dalam ulasan produk atau layanan, seperti kualitas produk atau harga.

4. *Opinion Mining*

Sub-bidang *Sentiment Analysis* yang berfokus pada pengumpulan dan analisis opini dari sumber-sumber online. Teknik ini digunakan untuk mengetahui opini dan preferensi pengguna terhadap produk, merek, atau topik tertentu.

5. *Sarcasm Detection*

Sub-bidang *Sentiment Analysis* yang berfokus pada deteksi sindiran atau lelucon dalam teks. Teknik ini digunakan untuk mengidentifikasi apakah teks mengandung sindiran atau lelucon yang mungkin mempengaruhi analisis sentimen.

2.5 *Preprocessing*

Preprocessing pada *Natural Language Processing* adalah cara yang melibatkan transformasi data teks mentah menjadi format yang dapat dimengerti. Data mentah sering sekali tidak lengkap, tidak konsisten, dan dipenuhi dengan banyak noise dan kemungkinan besar mengandung banyak kesalahan. *Preprocessing* adalah metode yang terbukti untuk menyelesaikan masalah tersebut, maka dari itu berikut tahapan *preprocessing* yang dilakukan [17].

1. *Lowercasing*

Merupakan cara untuk mengubah huruf kecil pada teks agar semua data memiliki format yang seragam dan untuk memastikan “NLP” dan “nlp” diperlakukan sama.

2. *Punctuation Removal*

Merupakan cara unruk menghapus tanda baca dari data teks. Tujuannya agar tanda baca tidak menambah informasi atau nilai lain dan meningkatkan efisiensi komputasi.

3. *Stop Words Removal*

Merupakan cara untuk menghapus kata henti, yang dimana kata henti ini tidak memiliki arti dengan kata lainnya. Sebagai contoh kata “how” dan “to” pada kalimat berbahasa inggris berikut ,“how to develop chatbot using python”.

4. *Text Standardzitation*

Merupakan cara mengubah kata singkatan menjadi kalimat atau kepanjangan yang dapat dipahami oleh mesin. Sebagai contoh “nlp” diperpanjang menjadi “natural language processing”.

5. *Spelling Correction*

Merupakan cara melakukan koreksi ejaan dan dapat membantu mengurangi beberapa salinan kata yang memiliki maknya yang sama.

6. *Tokenization*

Merupakan cara untuk melakukan tokenisasi yang mengacu pada memecah teks menjadi unit-unit minimal yang bermakna. Tokenization dibagi menjadi dua yaitu tokenizer kalimat dan kata.

7. Stemming

Merupakan cara untuk mengekstraksi sebuah kata menjadi kata dasar.

8. Lemmatization

Merupakan cara atau proses mengekstrak kata dasar dengan mempertimbangkan kosakata. Contohnya, kata “bagus”, “lebih baik”, dan “terbaik” dilakukan lemmatization menjadi “baik”.

9. Exploratory Data Analysis

Merupakan cara yang dilakukan setelah pengumpulan data dan preprocessing teks yaitu melakukan analisis data eksplorasi.

10. End-to-End Processing Pipeline

Merupakan tahapan untuk membangun end-to-end text preprocessing pipeline.

2.6 *Corpus*

Penelitian empiris dapat dilakukan dengan menggunakan teks tertulis atau lisan, seperti teks-teks dasar dari berbagai jenis sastra dan analisis linguistik. Tapi gagasan tentang korpus sebagai dasar untuk sebuah bentuk linguistik empiris berbeda dalam beberapa cara mendasar dari teks-teks tertentu. Pada prinsipnya, setiap koleksi lebih dari satu teks dapat disebut corpus: istilah corpus dalam bahasa latin berarti body, maka corpus dapat didefinisikan sebagai isi setiap teks. Tapi istilah '*corpus*' ketika digunakan dalam konteks linguistik modern memiliki konotasi yang lebih spesifik. Ada empat karakteristik dari corpus [18].

a. *Sampling and Representativeness*

Dalam membangun sebuah korpus dari berbagai bahasa, dapat ditarik dari sebuah sampel yang mewakili dari berbagai pengujian secara maksimal, yaitu menyediakan corpus seakurat mungkin dari kecenderungan yang beragam termasuk proporsi antara *corpus* dan informasi yang dicari. Jadi, tidak semata-mata berdasarkan pada teks sampel yang dipilih, akan tetapi mencari sampel dari berbagai sumber yang diambil dari sumber dokumen aslinya, sehingga akan memberikan gambaran yang cukup akurat dari seluruh informasi yang akan didapatkan.

b. Finite Size

Selain sampling, istilah *corpus* juga cenderung menyiratkan suatu isi teks dengan ukuran yang terbatas, misalnya 1.000.000 kata. Teks dapat terus ditambahkan ke dalamnya, sehingga semakin besar karena lebih banyak sampel yang ditambahkan. Keuntungan utamanya : (1) teks menjadi tidak statis karena teks yang baru akan selalu ditambahkan dan (2) ruang lingkup akan lebih besar dan jauh lebih luas sehingga akan mencakup dari bahasa yang digunakan. Kelemahan utamanya adalah bahwa, karena terus berubah dalam ukuran dan kurang ketatnya sampel, menjadi sumber yang kurang terpercaya dalam segi kuantitatif (sebagai lawan kualitatif). Jadi sebaiknya pada awal pembangunan korpus, rencana riset ditetapkan secara rinci bagaimana berbagai bahasa yang digunakan diambil sampelnya, berapa banyak sampel dan kata harus dikumpulkan sehingga jumlah keseluruhan yang sudah ditetapkan ini dapat digunakan.

c. *Machine-Readable Form*

Corpora yang dapat dibaca oleh mesin memiliki beberapa keunggulan dibandingkan dengan format tertulis atau lisan. Pertama dan paling penting keuntungan dari corpora yang dapat dibaca oleh mesin adalah bahwa dimungkinkan untuk mencari dan memanipulasi dengan cara-cara yang tidak dilakukan dengan format lain. Sebagai contoh, sebuah korpus dalam format buku, akan perlu dibaca dari depan sampai belakang untuk mengambil semua contoh kata, dengan korpus yang dapat dibaca oleh mesin, tugas ini dapat dicapai dalam beberapa menit dengan menggunakan perangkat lunak, atau sedikit lebih lambat, dengan menggunakan fasilitas pencarian di pengolah kata. Keuntungan kedua corpora yang dapat dibaca oleh mesin adalah bahwa dapat dengan cepat dan mudah diperkaya dengan informasi tambahan.

d. Standard Reference

Meskipun tidak termasuk hal yang penting dari definisi suatu korpus, tetapi ada juga pemahaman bahwa korpus merupakan referensi standar untuk berbagai bahasa yang diwakilinya. Hal ini mengandaikan ketersediaan yang luas kepada peneliti lain, keuntungan dari korpus yang tersedia secara luas

adalah bahwa akan memberikan tolok ukur yang dapat digunakan sebagai pembanding dalam studi. Misalnya. secara langsung dibandingkan dengan hasil yang dipublikasikan (selama metodologi sama) tanpa perlu perhitungan ulang. Korpus standar juga berarti penggunaan corpus yang sama digunakan untuk berbagai macam variasi studi dan yang membedakannya yaitu penggunaan data pengujinya dan metodologi yang digunakan dalam studi [18].

2.7 Perbandingan Tinjauan

Berikut beberapa hasil dari penelitian terdahulu. [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31][32], [33], [34], [35], [36], [37], [38], [39], [40], [32], [33], [41], [42], [43], [44]

Tabel 1. Studi Penelitian yang Pernah Dilakukan - Corpus

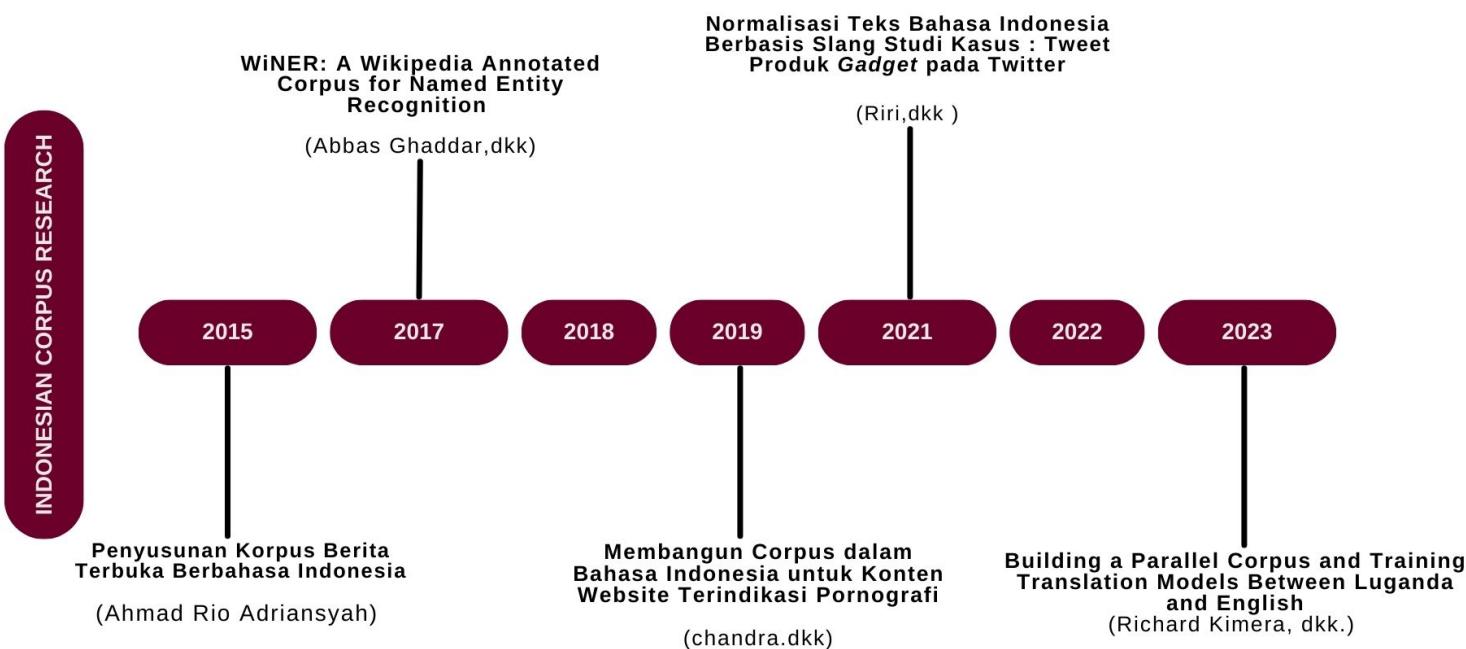
NO	JUDUL	METODE	PROSES	HASIL	AKURASI	KESIMPULAN
1.	Membangun Korpus Bahasa Indonesia untuk Konten Website Terindikasi Pornografi	NLP (Natural Language Processing) dan Vektorisasi Kata	1. Analisis mengenai isu konten pornografi mengidentifikasi konten berindikasi pornografi berdasarkan korpus yang dibangun 2. Melakukan pengecekan data dari aduan masyarakat 3. analisis kebutuhan 4.peranan teknik crawling dan corpus 5. penerapan teknik crawling 6. tokenisasi 7. Stopwords 8. vektorisasi kata 9. uji coba Proses membangun korpus Tahap 1 1. impor perusahaan NLTK terdapat fitur tokenize yaitu fitur pemisahan teks hasil crawling menjadi satuan kecil yang disebut dengan token 2. impor IO Library untuk pembacaan file yang bersifat konten website 3. simpan file yang akan diberi token dalam format.txt 4. membuat blok perulangan Tahap 2 1. menghilangkan kata-kata dari tokenisasi yang tidak memiliki stopwords Tahap 3 1. Mencari kata dasar dari stopwords dan melakukan stemming 2. Vektorisasi kata 3.Terbangun korpus			Mengembangkan korpus konten pornografi menggunakan NLP dan melakukan vektorisasi kata pada hasil korpus yan terindikasi konten pornografi
2.	A NOVEL ARABIC CORPUS FOR TEXT CLASSIFICATION USING DEEP LEARNING AND WORD EMBEDDINGDEEP LEARNING AND WORD	Klasifikasi Teks	1. Pengumpulan data dengan merayapi beberapa portal berita 2. prapemrosesan dasar untuk data yang dikumpulkan memfilter karakter non-arab, menghapus karakter khusus dan tanda baca dan menormalkan teks (ini termasuk, sebagai contoh, mengganti huruf "†" dengan huruf "ا"	1. Ukuran kumpulan data berdasarkan pada pelatihan algoritma pembelajaran mendalam dan kinerja klasifikasi. 2. Word2vec meningkatkan representasi vektor kata dan kinerja klasifikasi.		1. Membangun Korpus serta memproduksi model penyemat kata (menggunakan word2vec, fastText, dll.). 2. Menyelidiki kinerja model pembelajaran mendalam untuk klasifikasi teks Arab
3.	AkuGalau: Korpus Bahasa Indonesia untuk Deteksi Emosi dari Teks	1. Metode berbasis leksikon dan pembelajaran mesin digunakan. 2. Metode Bayes Naive digunakan untuk deteksi emosi.	1. Anotasi manual untuk tweet dengan beberapa tagar emosi. 2. Anotasi otomatis berdasarkan hashtag emosi di tweet. 3. Pemrosesan data awal untuk membersihkan dan menstandarkan format tweet.	1. Korpus teks bahasa Indonesia dikembangkan dengan 500 tweet. 2. Tingkat akurasi deteksi emosi mencapai 82% menggunakan metode Naive Bayes.	1. Eksperimen deteksi emosi mencapai akurasi 82% menggunakan Naive Bayes 2. Eksperimen deteksi emosi dengan akurasi 82% menggunakan metode Naive Bayes.	1. Pengembangan korpus teks emosi Indonesia untuk deteksi emosi. 2. Pentingnya korpus dalam deteksi emosi untuk bahasa yang kekurangan sumber daya.

4.	Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis		<p>1. Pengumpulan Data Dokumen menggunakan Crawler</p> <p>Penelitian ini dilakukan dengan tujuan menghasilkan suatu aplikasi yang dapat digunakan untuk membangun corpus dengan berbagai format data yang berasal dari hasil crawling secara otomatis.</p>		<p>1. Berdasarkan hasil uji coba aplikasi yang dilakukan dari proses setting crawler dengan alamat web www.unisbank.ac.id, dengan folder untuk penyimpanan data hasil download pada E:\MyDoku\Penelitian\Corpus\ALWebSpider\1000, dengan max deep level pada nilai satu, image ikut di download, dan proses download pada server yang sama, maka menghasilkan tiga puluh tiga file baik file</p> <p>2. Aplikasi mampu melakukan download dari sebuah alamat web dengan otomatis dengan ketentuan dapat dilakukan oleh user. Semua file yang dapat didownload akan diampli semuanya tanpa terkecuali. Sehingga akan mempermudah user dalam pengumpulan data tanpa harus mendownload satu-per-satu file.</p> <p>3. Aplikasi mampu menampilkan hasil pencarian dokumen dan mengurutkannya berdasarkan urutan dari penemuan dari file data yang dicari, dalam arti dokumen data yang ditemukan pertama kali akan ditempatkan di urutan pertama sedangkan dokumen data yang ditemukan terakhir akan ditempatkan pada urutan paling bawah. Dalam uji implementasi dapat dilihat bahwa urutan pertama adalah file 1.bmp dan yang terakhir adalah 9.htm</p> <p>4. Aplikasi mampu melakukan konversi dari dokumen teks dengan berbagai format data ke dalam bentuk dokumen teks txt, juga dalam melakukan konversi pada semua format file image ke dalam bentuk format bmp. Konversi dilakukan untuk menyamakan format untuk dapat mempermudah dalam penyimpanan dalam database.</p> <p>5. Aplikasi mampu menyimpan dokumen teks dan image dalam tabel teks dan tabel image secara otomatis dari semua hasil pencarian dan konversi yang telah dilakukan pada proses sebelumnya. Sehingga mempermudah user apabila mempunyai data yang besar tanpa harus menginput satu-persatu file data ke dalam database.</p>
			2. Memasukkan Dokumen Teks ke dalam Tabel Corpus		

5.	Peran Korpus Dalam Penyusunan Kamus	<p>Peran korpus dalam penyusunan kamus :</p> <ol style="list-style-type: none"> 1. pada tahap pengumpulan data, korpus dapat membantu pekamus dalam Menyusun senarai kata mulai dari frekuensi yg tertinggi hingga frekuensi yg terendah. 2. pada tahap perentuan lema, korpus dengan program konkordansi dapat membantu pekamus untuk membedakan mana lema/sublema yg berupa majemuk. 3. korpus dapat membantu dalam hal penentuan kelas kata sebuah lema karena korpus memberikan konteks yang berbeda-beda tempat kata itu berada. 4. korpus membantu pekamus dalam mendefinisikan suatu lema 5. korpus memberikan keleluasaan kepada pekamus dalam mencari dan menentukan contoh yg baik bagi pengguna kamus. 6. korpus membantu dalam mengidentifikasi lokasi sebuah kata 7. korpus dapat membantu melacak perubahan kata 	<p>Korpus adalah kumpulan teks alami, baik Bahasa lisan maupun Bahasa tulis yang disusun secara sistematis.</p> <p>"Alami" karena teks yang diproduksi dan digunakan secara wajar dan tidak dibuat-buat. teks-teks tersebut termasuk novel, buku dan kertas akademis, koran, majalah, rekaman siaran pembicaraan dan wawancara, blog, jurnal, daring dan kelompok diskusi.</p> <p>"Sistematik" karena struktur dan isi korpus mengikuti prinsip ekstralinguistik tertentu, khususnya pengambilan sampel, yaitu prinsip dasar dalam pemilihan teks yang akan dimasukkan ke dalam korpus. misalnya, ada korpus yang dibatasi pada jenis teks tertentu, untuk satu atau beberapa variasi bahasa Inggris, atau untuk jangka Waktu tertentu. "Sistematik" juga berarti bahwa informasi tentang komposisi yang tepat dari suatu korpus tersedia bagi peneliti (termasuk jumlah kata dalam setiap kategori dan keseluruhan korpus, bagaimana teks-teks yang termasuk dalam korpus dijadikan sampel, dsb.). Meskipun korpus dapat merujuk pada setiap kumpulan teks yang sistematis, dewasa ini korpus biasanya digunakan dalam arti sempit dan hanya digunakan untuk merujuk pada kumpulan teks sistematis yang telah terkomputerisasi atau yang disajikan dalam bentuk elektronik.</p>	
6.	PENYUSUNAN KORPUS BERITA TERBUKA BERBAHASA INDONESIA	<ol style="list-style-type: none"> 1. mengumpulkan daftar 'seed word' dari beberapa ratus kata berfrekuensi menengah dalam suatu bahasa. 2. ulangi beberapa kali (hingga korpusnya berukuran cukup besar): <ul style="list-style-type: none"> a. pilih 3 kata untuk membuat sebuah query b. kirimkan query tersebut ke search engine popular (Google, Yahoo, Bing) yang mengembalikan halaman 'hasil pencarian'. c. buka halaman pada hasil pencarian tersebut, lalu disimpan. 3. Bersihkan teks dari navigasi bar, iklan, dan skrip lain yang muncul berulang. 4. Hapus duplikat 5. Tokenisasi, lematisasi, dan beri tag POS jika memungkinkan. 6. Masukkan ke dalam perangkat untuk corpus query 	<p>Penelitian dilakukan karna banyaknya redundansi (perulangan) kalimat, kalimat yang dihasilkan website berita cukup banyak dan beragam. Bahasa yang digunakan dalam penyampaian berita biasanya adalah Bahasa formal atau semi formal.</p>	<p>menghasilkan sekumpulan dokumen yang bersifat terbuka untuk komunitas yang meneliti pemrosesan Bahasa natural atau Bahasa Indonesia. Korpus yang diambil dengan metode di atas belum diberikan tag untuk POS (Part-of-Speech). Untuk tagging, dapat digunakan tagger otomatis seperti yang disampaikan oleh [3] atau secara manual. Yang dicantumkan dalam jurnal ini adalah Sebagian kecil data yang sudah berhasil diambil. Ke depannya, korpus ini akan dikembangkan ke website lain dan dengan rentang waktu yang lebih lebar.</p>

7.	PENGENALAN KORPUS DATA BAHASA PADA MAHASISWA PROGRAM STUDI PENDIDIKAN BAHASA DAN SASTRA INDONESIA FKIP UNIVERSITAS MATARAM	<p>1. Kegiatan pengabdian ini dimulai dari persiapan, baik berupa penyusunan proposal hingga pengajuan. Kemudian dilakukan sosialisasi kepada mahasiswa Prodi Pendidikan Bahasa dan Sastra Indonesia.</p> <p>2. Setelah dilakukan sosialisasi, agar program menjadi lebih terarah, dilakukan pendataan terutama mengenai peserta mengenai kelas, minat kepenelitian, asal, dan lain sebagainya yang diperlukan.</p> <p>3. Selanjutnya penyipahan bahan berupa perbaikan materi atau bahan yang disampaikan ke peserta. diperlukan ruang lokasi pelatihan, perangkat komputer dan internet, yang dibutuhkan untuk mengenakan korpus data elektronik.</p> <p>4. Tahap berikutnya adalah tahap pelaksanaan penyajian materi konseptual tentang hakikat korpus data bahasa, jenis korpus data bahasa manual dan digital.</p>	<p>Penelitian ini dilakukan menghasilkan tulisan ilmiah yang berkualitas dari segi substansi sekaligus tata tulis, penyuluhan bentuk kesalan berbahasa dalam skripsi perlu dilakukan</p>	<p>menghasilkan kondisi pertama, menunjukkan bahwa kelemahan mahasiswa kurang memahami korpus data, dapat disebabkan Waktu pelatihan dalam menjelaskan materi korpus data sebagai bagian dari materi pelatihan.</p> <p>Kondisi kedua, kurangnya mahasiswa dalam menemukan data bahasa sesuai topik penelitian kebahasaan, dapat pula disebabkan oleh selain belum memahami tentang hakikat korpus data bahasa, juga minim tentang pemahaman substansi kajian korpus data bahasa yang beragam.</p> <p>Kondisi ketiga, selayaknya mahasiswa telah memperoleh pengetahuan dan kompetensi tentang korpus data bahasa dalam mata kuliah metodologi penelitian bahasa. Namun mengingat, dalam mata kuliah ini tercakup materi/bahan ajar perkuliahan yang padat, maka tidak cukup untuk memaparkan materi korpus data bahasa yang juga cukup luas cakupannya. Maka sebenarnya diperlukan mata kuliah tersendiri yang mengkhususkan kajian tentang korpus data bahasa, yaitu mata kuliah Linguistik Korpus, seperti yang juga telah dilakukan pada perguruan tinggi lainnya.</p>
8.	PENYUSUNAN KORPUS BERITA TERBUKA BERBAHASA INDONESIA	<p>1. Mesin Sketch menggunakan mesin pencari untuk membuat korpus secara efisien.</p> <p>2.Pemberian data, tokenisasi, dan penandaan POS adalah langkah penting.</p>	<p>Makalah penelitian berfokus pada pengembangan korpus untuk bahasa Indonesia.</p> <p>Makalah ini membahas pentingnya korpus bahasa Indonesia yang dapat diakses. Ini menyebut perfungsi korpus bahasa Indonesia terbuka untuk penelitian.</p>	<p>Penelitian ini menyediakan korpus berita Indonesia terbuka untuk NLP. Korpus mencakup data dari situs berita untuk penelitian linguistik. Korpus WinER meningkatkan sistem NER, terutama model LSTM-CRF</p>
9.	WINER: A Wikipedia Annotated Corpus for Named Entity Recognition	<p>1.LSTM-CRF dan model berbasis fitur yang digunakan untuk pengenalan entitas bernama</p> <p>2.Menjelajahi struktur tautan keluar di Wikipedia untuk anotasi entitas bernama</p>	<p>1. Ekstraksi anotasi entitas bernama dari Wikipedia.</p> <p>2. Memanfaatkan struktur tautan keluar dan string permukaan untuk pengenalan entitas.</p>	<p>1. Akurasi anotasi diukur pada 92% .</p> <p>2. Akurasi menurun menjadi 88%</p>
	Normalisasi Teks Bahasa Indonesia Berbasis Slang Studi Kasus : Tweet Produk Gadget pada Twitter	<p>1. Metode normalisasi: Model Word2vec untuk kata-kata gaul di tweet.</p> <p>2. Metode evaluasi: Tugas klasifikasi dengan 3 kelas sentimen.</p> <p>3. Evaluasi kinerja: Peningkatan akurasi dari 88% menjadi 91% pasca-normalisasi.</p>	<p>Dataset dengan normalisasi memiliki akurasi yang lebih tinggi dibandingkan dengan dataset yang tidak dinormalisasi.</p> <p>3. Ekstraksi fitur menggunakan Tf-idf untuk evaluasi dataset.</p>	<p>1. Akurasi meningkat dari 88% menjadi 91% setelah normalisasi teks.</p> <p>2. Normalisasi dengan word2vec menghasilkan akurasi 91%, lebih tinggi daripada tanpa mengurangi kebisingan dalam kumpulan data.</p> <p>1. Metode NB mengungguli SVM dan RFDT dalam akurasi klasifikasi teks.</p> <p>2. Pra-pemrosesan mengurangi kumpulan data tweet menjadi 989 tweet untuk pengembangan model.</p> <p>3. Pembersihan menghilangkan kata-kata yang tidak perlu untuk mengurangi kebisingan dalam kumpulan data.</p>

10.	Building a Parallel Corpus and Training Translation Models Between Luganda and English	1. Membangun korpus paralel dengan 41.070 kalimat untuk Luganda dan Inggris. 2. Model NMT terlatih menggunakan arsitektur Transformer dengan pencarian hiper-parameter	1. Melatih model NMT dengan pencarian hiper-parameter pada dataset. 2. Membangun korpus paralel untuk terjemahan Luganda dan bahasa Inggris.	Berfokus pada model terjemahan mesin saraf Luganda-Inggris	Skor BLEU: 17,47 untuk En2Lu, 21,28 untuk Lu2En. 1. Model NMT mencapai skor BLEU 17,47 dan 21,28. 2. Model NMT Luganda-Inggris dikembangkan, dataset menjadi publik
-----	--	---	---	--	---

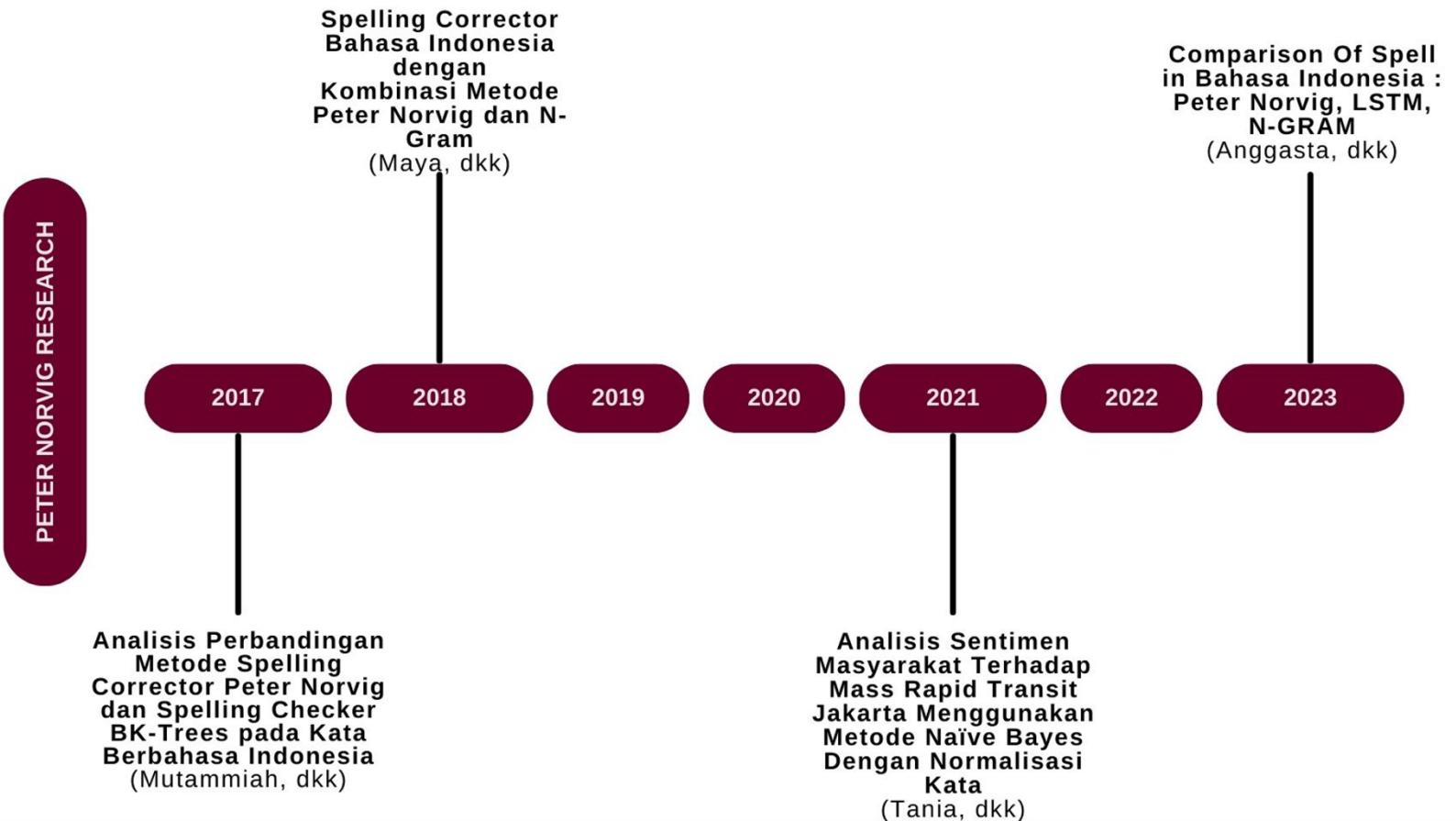


Gambar 2. Indonesian Corpus Research

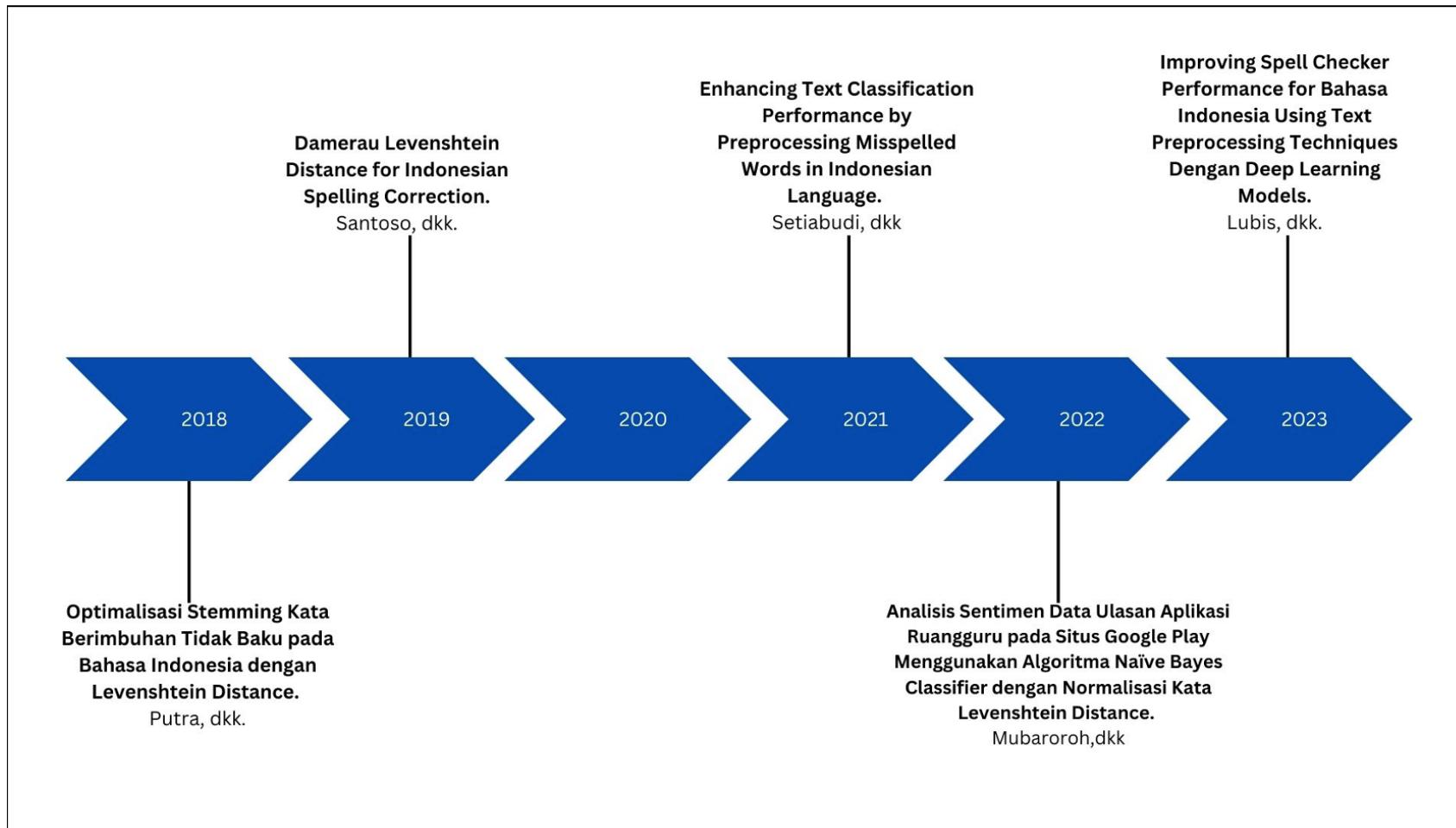
Tabel 2. Studi Penelitian yang pernah dilakukan - Text Preprocessing

Aini, Y., Hakim, A. R., & Mulyono. (2020). Sentiment analysis of Bank BNI user comments using the support vector machine method. Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020, 202-207. https://doi.org/10.1109/iSemantic50169.2020.9260010	
Arid Ridho, Mahyuddin K.M. Nasution, Opim Salim Shompul, Elvawaty Muisa Zanzami	2020 Metode penelitian yang digunakan dalam studi ini adalah analisis sentimen menggunakan metode Support Vector Machine. Tahapan penelitian meliputi: 1. Pengumpulan Data: Komentar dari pengguna layanan BNI berbasis Aplikasi Perbankan Mobile. 2. Pelabelan Data: Data komentar dielabel secara manual menjadi sentimen positif dan negatif. 3. Preprocessing: Tahapan awal dalam pengolahan teks, meliputi: a. Case Folding: Mengubah semua kata menjadi huruf kecil. b. Data Cleaning: Membersihkan kata-kata dengan menghapus tanda baca. c. Stopword Removal: Menghapus kata-kata umum yang tidak memberikan makna. d. Stemming: Mengubah kata-kata dengan akhiran menjadi bentuk dasar. e. Tokenization: Memecah dokumen menjadi bagian kecil seperti kata-kata. 4. Analisis Sentimen: Mengklasifikasi komentar ke dalam sentimen positif dan negatif menggunakan metode Support Vector Machine.
Fatihah Rahmadayana, & Yulianti Sibaroni. (2021). Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(5), 936-942. https://doi.org/10.29207/resti.v5i5.3457	
Fatihah Rahmadayana, Yulianti Sibaroni	2021 1. Pengumpulan Data: Data tweet dari Twitter dikumpulkan menggunakan Twitter API dengan bahasa pemrograman Python. Pemberian Label pada Data: 2. Data Labelling (Manual Label) 3. Pre-processing Data: Ekspansi Akronim, case folding, data cleansing, slang word, emoji translation, stopword removal, stemming. 4. Pembobotan Istilah (TF-IDF) 5. Pemodelan menggunakan SVM 6. Optimasi dengan Randomized (Untuk mengatur hiperparameter SVM) 7. Evaluasi (Pengukuran Kinerja menggunakan F1-Score)
Anam, M. K., Fitri, T. A., Agustin, A., Lusiana, L., Firdaus, M. B., & Nurhuda, A. T. (2023). Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm. IJKOM Jurnal Ilmiah, 15(2), 290-302. https://doi.org/10.33098/ijkom.v15i2.1590.290-302	
M. Khairul Anam, Triyani Arita Fitri, Agustin, Lusiana, Muhammad Bambang Firdaus, Agus Tri Nurhuda	2023 1. Data Collection 2. Pelabelan dengan Lexicon Based (Using InSet Lexicon) 3. Preprocessing (case folding and cleaning, tokenization, word normalization, stop words removal, stemming) 4. Pembobotan dengan TF-IDF 5. Melakukan pemodelan SVM dengan Linear Kernel 5. Evaluation dengan Confusion Matrix
Aulia, B., Utomo, P. E. P., Khaira, U., & Suratno, T. (2021). ANALISIS SENTIMEN TAGAR #INDONESIA TERSERAH DI MASA COVID-19 MENGGUNAKAN METODE SENTISTRENGTH. Jurnal Komputer dan Informatika, 9(2), 207-213. https://doi.org/10.35508/jcon.v9i2.4275	
Bisma Aulia, Pradita Eko Prasetyo Utomo, Ulfa Khaira, dan Tri Suratno	2021 Tahapan penelitian meliputi : 1. Crawling data, 2. preprocessing data, - Case Folding - Tokenization - Stopword - Stemming 3. pembobotan kata menggunakan TF-IDF 4. Pemodelan menggunakan SentiStrength 5. Evaluasi menggunakan confusion matrix

		Malik Iryana, T., & Pandu Adikara, P. (2021). Analisis Sentimen Mayarakat Terhadap Mass Rapid Transit Jakarta Menggunakan Metode Naive Bayes Dengan Normalisasi Kata. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 5(6), 2548-2964. http://j-ptik.uin.ac.id																			
Tania Malik Iryana, Indriati, Dan Putra Purba Adikara	2021	Metode yang digunakan adalah kombinasi metode Peter Norvig dan N-Gram untuk spelling corrector pada dalam Bahasa Indonesia. Metode ini diimplementasikan dalam sebuah aplikasi yang menguji kesalahan kata dalam kalimat input. Data yang digunakan adalah dokumen ARPA yang berisi nilai-nilai probabilitas dari serangkaian kata n, yang didapatkan dari proses pembangunan language model dari suatu dokumen menggunakan script SRILM.																			
		-Tahapan dimulai dari : 1. Perbaikan Data 2. Peningkatan Algoritma : - Menginput data b. Preprocessing : case folding, tokenization, slangword, peter norvig, stopword, stemming c. Hasil 3. Pembentukan Kata (TF-IDF) 4. Evaluasi 5. Model : Naive Bayes	✓		✓		✓	✓	✓						✓					(Han ya TF)	90,03% 90,03% 94,4% 92,2%
Luthfi, A. R., Lawi, Y. V., Rahman, D. A., & Witaryah, D. (2023). Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models. <i>Ingenierie des Systemes d'Information</i>, 28(5), 1335–1342. https://doi.org/10.18280/ist.280522																					
Arif Ridho Lubis, Mahyuddi K. M. Nasution, Opini Sintopul, Elviantawaty Muissa Zamzami	2022	1. Pra-Pemrosesan (Preprocessing): Data tweet yang diperoleh dari media sosial Twitter -Pembersihan Data: Menghilangkan karakter khusus, tanda baca, dan emotikon dari teks. -Tokenisasi: Memecah teks menjadi token atau kata-kata individuul. -Normalisasi: Teksi Mengubah huruf kapital menjadi huruf kecil untuk konsistensi. -Penghapusan Stopwords: Menghapus kata-kata umum yang tidak memberikan nilai tambah dalam analisis. -Stemming atau Lemmatisasi: Mengubah kata-kata ke bentuk dasarannya untuk memungkinkan variasi kata. -Pembentukan Data Tambahan: Menghapus URL, tag HTML, atau informasi yang tidak relevan.) 2. Pengembangan Model (Deep Learning): Model deep learning menggunakan Bi-LSTM digunakan untuk melakukan perbaikan kata-kata pada data tweet yang telah diproses. Model ini bertujuan untuk meningkatkan akurasi ejakan dalam kata-kata bahasa Indonesia. 3. Klasifikasi Kata Formal dan Non-Formal: Data yang telah dieck ejajanya kemudian diklasifikasikan menjadi kata-kata formal dan non-formal menggunakan model deep learning. Proses klasifikasi ini membantu dalam membedakan antara kata-kata standar dan non-standar dalam data tweet.				✓		✓		✓	✓					✓	✓				82,5%
Kusuma, A. T. A., & Ratnasari, C. I. (2023). COMPARISON OF SPELL CORRECTION IN BAHASA INDONESIA: PETER NORVIG, LSTM, AND N-GRAM. <i>JIKO (Jurnal Informatika dan Komputer)</i>, 6(3), 214–220. https://doi.org/10.33387/jiko.v6i3.7072																					
Anggasta TA Kusuma, Chaniyah I. Ratnasari	2023	Metode yang digunakan dalam studi ini adalah perbandingan tiga pendekatan dalam koreksi ejakan Bahasa Indonesia, yaitu metode Peter Norvig, Long Short-Term Memory (LSTM), dan N-gram. Penelitian ini menggunakan data SPECIL (Spell Error Corpus for Indonesian Language) yang mencakup dokumen dengan berbagai jenis kesalahan seperti penyimpangan, penghapusan, transposisi, dan substitusi. Dataset pengujian terdiri dari 150 kata, sejalan dengan referensi korpus 150 kata dari Leipzig Corpus Collection® yang digunakan untuk metode Peter Norvig dan N-gram. Metode Peter Norvig menunjukkan sebagai yang paling kuat, mencapai tingkat akurasi yang mengesankan sebesar 89%. Metode N-gram menghasilkan akurasi 75%, sementara LSTM, meskipun masih memberikan akurasi yang masuk akal sebesar 74%, tetapi dibandingkan dengan pendekatan lainnya. Studi ini juga menggunakan dataset referensi 10.000 kata untuk pengujian, menunjukkan kekokohan ketiga metode terhadap dataset yang lebih besar. Meskipun metode Norvig tampak sebagai yang terbaik, baik n-gram maupun LSTM juga memberikan kontribusi signifikan.																			Peter Norvig: Akurasi: 89% Kecepatan perhitungan: 35 kata per detik Kata yang berhasil dikoreksi: 387 kata Kata yang tetap tidak diketahui: 5%
Yanfi, Yanfi, Ford Lumban Gaol, Benfano Soewito, Harczo Leslie Hendric Spits Warnars																					
		Dalam konteks pengembangan korектор ejakan untuk bahasa Indonesia, beberapa permasalahan yang sering dihadapi termasuk variasi ejakan kata, kesalahan ketik, dan kompleksitas struktur bahasa Indonesia yang hanya akan prefiks, suffiks, dan imbuhan. Selain itu, kurangnya sumber daya dan data yang memadai untuk melatih model korektor ejakan juga menjadi tantangan dalam pengembangan spell checker untuk bahasa Indonesia. Dari berbagai metode yang telah dibahas dalam artikel, spell checker yang menggunakan kombinasi algoritma Levenshtein Distance, Jaro-Winkler, Finite state automata, Peter Norvig, BK-Trees, N-gram, lemmatisasi, dan kombinasi Minimum Edit Distance dan bigram dapat dianggap sebagai metode yang baik. Kombinasi metode ini memungkinkan untuk mendekripsi dan menganalisa kesalahan ejakan dengan tingkat akurasi yang lebih baik, mengingat variasi ejakan kata, dan memperbaiki struktur bahasa Indonesia yang kompleks. Dengan demikian, spell checker yang menggunakan beberapa metode tersebut dapat memberikan solusi yang efektif dalam mendekripsi dan memperbaiki kesalahan ejakan dalam bahasa Indonesia. Namun, penting untuk tetap melakukan penelitian dan pengembangan lebih lanjut untuk meningkatkan kualitas dan efisiensi spell checker dalam konteks bahasa Indonesia yang kaya akan variasi dan kompleksitas linguistik.																			N-gram: Akurasi: 75% Kecepatan perhitungan: 21 detik per kata Kata yang berhasil dikoreksi: 150 kata Kata yang tetap tidak diketahui: 11%
Ari, Y., Sulyanti, W. N., Kuwardiani, D., & Fajri, M. (2022). Pelajaran Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. <i>PETIR</i>, 15(2), 254–275. https://doi.org/10.3332/petir.v15i2.1733																					
Yessy Ari, Widya Nita Sulyanti, Dwina Kuswardan i, Muhammad Fajri	2022	Metode penelitian yang digunakan dalam studi analisis sentimen opini pelanggan terhadap aplikasi PLN Mobile meliputi beberapa tahapan. Tahap awal penelitian menseleksi pengumpulan data ulasan menggunakan teknik web scraping, penerjemahan mesin, pelabelan data, preprosesing teks (case folding, slang word, tokenizing, filtering, stemming), analisis TF-IDF, klasifikasi teks menggunakan pendekatan Lexicon Vader, dan evaluasi model dengan metode Naive Bayes . Selain itu, penelitian ini juga melibatkan proses tagging data yang menghasilkan 489 pendapat positif, 145 pendapat negatif, dan 365 pendapat netral. Hasil analisis sentimen ini kemudian dibandingkan dengan ulasan berdasarkan peringkat pengguna untuk memperoleh pemahaman yang lebih baik mengenai opini pelanggan terhadap aplikasi PLN Mobile.	✓		✓						✓						✓	✓		70% neutral (50%) pos (47%) neg (23%) neg92 %) neg67 %) neg7 7%)	



Gambar 3. Spelling Corrector Research



Gambar 4. Research Position of Levenshtein Distance as Spelling Corrector

2017

- A Non-Word Error Spell Checker for Patient Complaints in Bahasa Indonesia. (Ratnasari dkk)
- Pengoreksian Ejaan Kata Berbahasa Indonesia Menggunakan Algoritma Levenshtein Distance (Bradley, dkk)

2018

- Arabic Sentiment Analysis Using a Levenshtein Distance Based Representation Approach (Khamar, dkk)
- Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levenshtein Distance Method (Naga, dkk)

2019

- Auto-correction of English to Bengali Transliteration System using Levenshtein Distance (Hossain, dkk)
- Damerau Levenshtein Distance for Indonesian Spelling Correction (Santoso, dkk)

2020

- Spelling Checker using Algorithm Damerau Levenshtein Distance and Cosine Similarity (Hamidah, dkk)

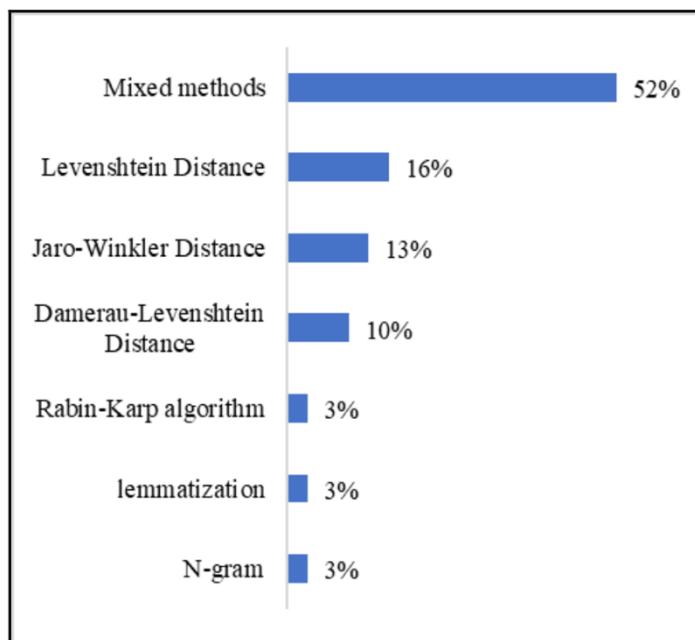
2021

- Analisis sentimen data ulasan aplikasi ruangguru pada situs google play menggunakan algoritma naive bayes classifier dengan normalisasi kata levenshtein distance (Mubaroh, dkk)

2022

- Improving spell checker performance for bahasa indonesia using tet preprocessing technique dengan deep learning models. (Lubis, dkk)

2023



Gambar 5. Distribusi Makalah dalam hal Metode Penelitian

Gambar 5 menggambarkan distribusi studi yang dianalisis berdasarkan metode penelitian yang diterapkan [5]. Terlihat bahwa 52% peneliti menggunakan metode campuran untuk memeriksa kesalahan pengejaan (16 makalah). Selain itu, penelitian lain menggunakan Levenshtein Distance (5 makalah), Jaro-Winkler Distance (4 makalah), dan algoritma Damerau-Levenshtein Distance (3 makalah).

3 METODE PENELITIAN

3.1 Motivasi

Kepuasan pelanggan adalah indikator utama dalam pengembangan bisnis, baik untuk perusahaan besar maupun usaha kecil dan menengah (UMKM). Kepuasan pelanggan mengukur sejauh mana produk atau jasa memenuhi atau melampaui ekspektasi pelanggan. Pendekatan umum untuk menentukan kepuasan pelanggan adalah melalui teknik penelitian berbasis survei. Namun, di era disruptif industri 4.0, teknik ini memiliki kelemahan signifikan dalam hal waktu, biaya, dan tenaga.

Saat ini, banyak aktivitas ekonomi dilakukan melalui platform digital, sehingga pendekatan survei tradisional menjadi kurang efektif. Beberapa penelitian awal telah mencoba mengembangkan teknik ekstraksi kepuasan pelanggan berbasis *big data*, tetapi metode yang ada umumnya hanya tersedia dalam bahasa Inggris, sementara pengembangan repository dan *library* untuk bahasa Indonesia masih sangat terbatas. Bahasa Indonesia memiliki aturan ejaan yang kompleks, yang memerlukan pemeriksa ejaan otomatis untuk memastikan standar dan keakuratan penulisan.

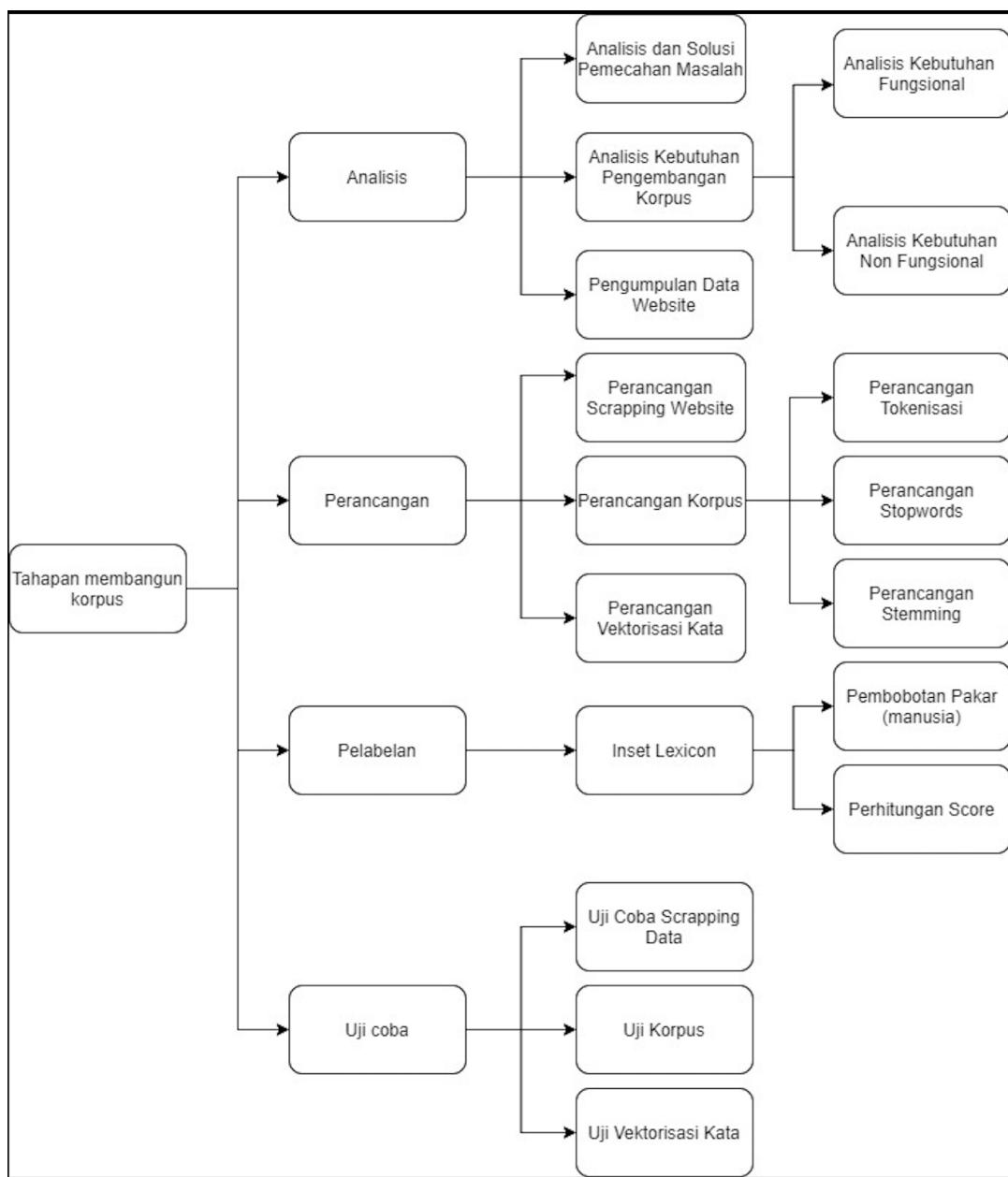
Motivasi utama penelitian ini adalah mengatasi keterbatasan teknik penelitian berbasis survei dengan merancang dan mengembangkan metode baru untuk ekstraksi data kepuasan pelanggan dari platform digital menggunakan *big data*, yang sesuai dengan karakteristik bahasa Indonesia. Hasil dari penelitian ini bertujuan untuk membangun *repository* dan *library* khusus bahasa Indonesia untuk analisis kepuasan pelanggan. Mengembangkan algoritma pemeriksa ejaan otomatis

yang lebih efektif untuk bahasa Indonesia. Menerapkan teknik *preprocessing* teks seperti *tokenisasi*, penghilangan *stopword*, *spelling corrector*, *stemming*, dan *lemmatization* untuk meningkatkan kualitas data. Menguji teknik yang dikembangkan dengan menggunakan ulasan pengguna dari aplikasi PLN Mobile sebagai studi kasus untuk mengukur kepuasan pelanggan.

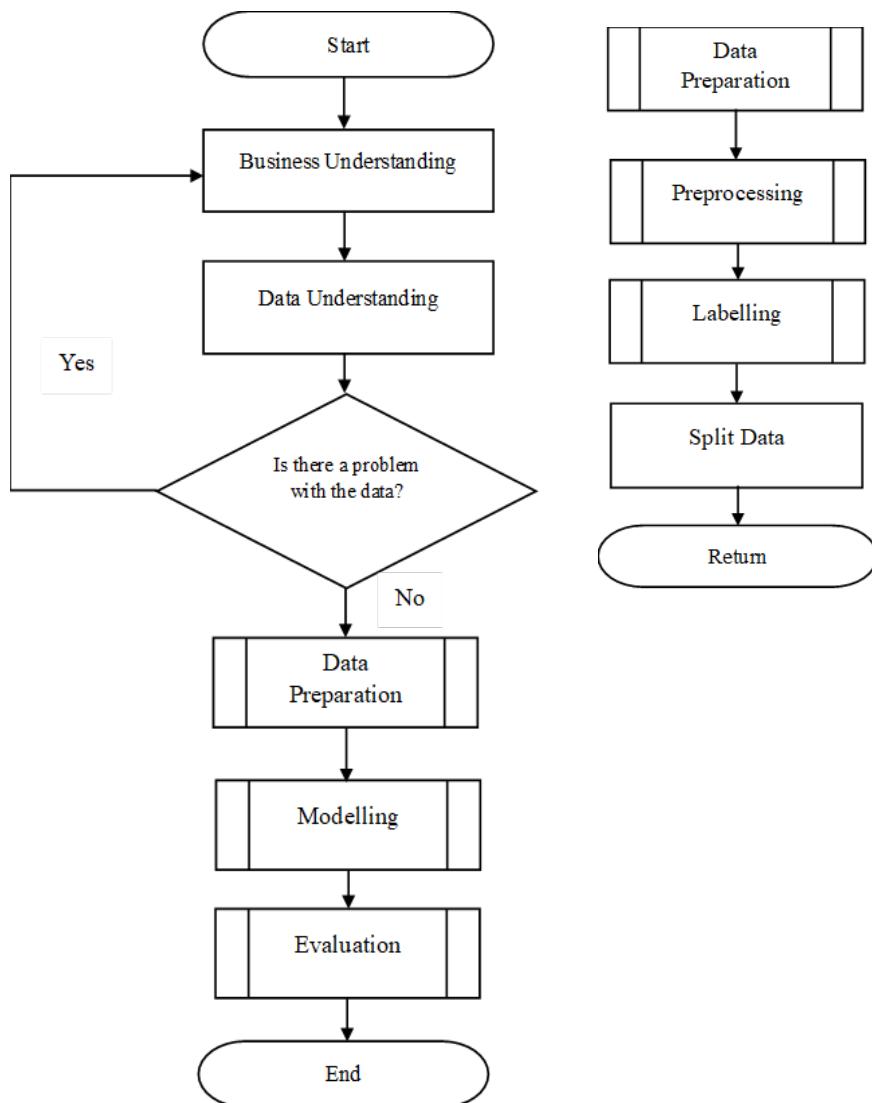
Dengan demikian, penelitian ini tidak hanya bertujuan untuk meningkatkan akurasi dan keandalan model ekstraksi data tetapi juga menyediakan alat yang bermanfaat untuk analisis kepuasan pelanggan di era digital, khususnya bagi penutur bahasa Indonesia.

3.2 Framework Riset

Penelitian dilakukan dalam beberapa tahapan. Mulai dari tahapan pencarian literatur hingga pemodelan dan perhitungan. Hasil pengujian diharapkan dapat memenuhi persyaratan tertentu guna menjawab tujuan dari penelitian ini. Tahapan metode penelitian dalam bentuk diagram alir dapat dilihat pada Gambar 2.



Gambar 6. Tahapan Pembangunan Corpus



Gambar 7. Tahapan Preprocessing

3.3 Pendekatan

Pendekatan penelitian yang digunakan dalam proposal disertasi ini adalah pendekatan kuantitatif dengan fokus pada teknik eksperimental dan *big data*. Berikut beberapa pendekatan yang akan dilakukan:

- Kuantitatif:** Penelitian ini bertujuan untuk mengukur kepuasan pelanggan menggunakan data numerik yang diekstraksi dari ulasan pengguna di platform

- digital *PLN Mobile*. Pengukuran ini akan dilakukan melalui pengembangan teknik baru yang mengandalkan data besar (big data).
2. **Eksperimental:** Penelitian akan mengembangkan dan menguji algoritma baru untuk ekstraksi data kepuasan pelanggan yang sesuai dengan karakteristik Bahasa Indonesia. Teknik ini melibatkan proses eksperimen untuk menguji efektivitas teknik *preprocessing teks*, seperti *tokenisasi*, penghilangan *stopword*, *spelling corrector*, *stemming*, dan *lemmatization*.
 3. **Big Data:** Penelitian ini memanfaatkan data besar dari ulasan pengguna aplikasi *PLN Mobile* yang akan diolah dan dianalisis untuk mengukur kepuasan pelanggan. Teknik ekstraksi yang dikembangkan akan diuji menggunakan data ulasan pengguna untuk memastikan akurasi dan keandalannya.

Pendekatan ini berfokus pada pengembangan teknik baru yang mengatasi keterbatasan teknik survei tradisional dan menyesuaikan dengan kebutuhan era disruptif industri 4.0, di mana efisiensi dan kecepatan sangat penting. Jenis data yang digunakan merupakan data sekunder, yaitu data pada ulasan pengguna aplikasi *PLN Mobile*, yang memiliki lebih dari 293.519 ulasan pengguna di Google Play Store (per Juni 2022). Aplikasi ini menyediakan berbagai layanan kelistrikan yang digunakan oleh jutaan pelanggan, sehingga menjadi sumber data yang kaya untuk analisis kepuasan pelanggan.

DAFTAR PUSTAKA

- [1] H. Kartika, M. Hertian Ranova, and C. Setia Bakti, ‘SURVEI KEPUASAN PELANGGAN UNTUK PENINGKATAN KUALITAS JASA PERAWATAN MESIN ATM DENGAN METODE CSI DAN IPA’.
- [2] R. Darwas, A. Saputra Sistem Informasi, S. Indonesia Padang Jalan Khatib Sulaiman Dalam No, and K. Padang, ‘SISTEMASI: Jurnal Sistem Informasi Implementasi Sistem Informasi Kepuasan Mahasiswa Terhadap Layanan Tugas Akhir Implementation of Student Satisfaction Information System to the Final Project Services’. [Online]. Available: <http://sistemas.ftik.unisi.ac.id>
- [3] A. R. Lubis, Y. Y. Lase, D. A. Rahman, and D. Witarsyah, ‘Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models’, *Ingenierie des Systemes d'Information*, vol. 28, no. 5, pp. 1335–1342, 2023, doi: 10.18280/isi.280522.
- [4] M. Rivera-Acosta, J. M. Ruiz-Varela, S. Ortega-Cisneros, J. Rivera, R. Parra-Michel, and P. Mejia-Alvarez, ‘Spelling correction real-time american sign language alphabet translation system based on yolo network and LSTM’, *Electronics (Switzerland)*, vol. 10, no. 9, May 2021, doi: 10.3390/electronics10091035.
- [5] Y. Yanfi, F. L. Gaol, B. Soewito, and H. L. H. S. Warnars, ‘Spell Checker for the Indonesian Language: ExtensiveReview’, *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 5, pp. 1–7, May 2022, doi: 10.46338/ijetae0522_01.
- [6] V. Christanti Mawardi, N. Susanto, and D. Santun Naga, ‘Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levenshtein Distance Method’, in *MATEC Web of Conferences*, EDP Sciences, Apr. 2018. doi: 10.1051/matecconf/201816401047.

- [7] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, ‘Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya’, *Information (Switzerland)*, vol. 12, no. 2, pp. 1–17, Feb. 2021, doi: 10.3390/info12020052.
- [8] A. Ayedh, G. TAN, K. Alwesabi, and H. Rajeh, ‘The Effect of Preprocessing on Arabic Document Categorization’, *Algorithms*, vol. 9, no. 2, Apr. 2016, doi: 10.3390/a9020027.
- [9] P. E. Ltrc, M. Chinnakotla, and R. Mamidi, ‘Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning’. [Online]. Available: <https://github.com/PravallikaRao/SpellChecker>
- [10] P. H. Santoso, E. Istiyono, Haryanto, and W. Hidayatulloh, ‘Thematic Analysis of Indonesian Physics Education Research Literature Using Machine Learning’, *Data (Basel)*, vol. 7, no. 11, Nov. 2022, doi: 10.3390/data7110147.
- [11] J. G. Shim, K. H. Ryu, S. H. Lee, E. A. Cho, Y. J. Lee, and J. H. Ahn, ‘Text mining approaches to analyze public sentiment changes regarding covid-19 vaccines on social media in korea’, *Int J Environ Res Public Health*, vol. 18, no. 12, Jun. 2021, doi: 10.3390/ijerph18126549.
- [12] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, ‘Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile’, *PETIR*, vol. 15, no. 2, pp. 264–275, Nov. 2022, doi: 10.33322/petir.v15i2.1733.
- [13] D. Khurana, A. Koli, K. Khatter, and S. Singh, ‘Natural Language Processing: State of The Art, Current Trends and Challenges’.
- [14] H. Taherdoost and M. Madanchian, ‘Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research’, *Computers*, vol. 12, no. 2. MDPI, Feb. 01, 2023. doi: 10.3390/computers12020037.
- [15] F. Fridom Mailo *et al.*, ‘Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia’, 2019.

- [16] M. D. Devika, C. Sunitha, and A. Ganesh, ‘Sentiment Analysis: A Comparative Study on Different Approaches’, in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 44–49. doi: 10.1016/j.procs.2016.05.124.
- [17] A. Kulkarni and A. Shivananda, *Natural language processing recipes: Unlocking text data with machine learning and deep learning using python*. Apress Media LLC, 2019. doi: 10.1007/978-1-4842-4267-4.
- [18] ‘tony_mcenery_andrew_wilson_corpus_linguisticsbook4you-org’.
- [19] A. T. A. Kusuma and C. I. Ratnasari, ‘COMPARISON OF SPELL CORRECTION IN BAHASA INDONESIA: PETER NORVIG, LSTM, AND N-GRAM’, *JIKO (Jurnal Informatika dan Komputer)*, vol. 6, no. 3, pp. 214–220, Dec. 2023, doi: 10.33387/jiko.v6i3.7072.
- [20] R. Sianipar and E. B. Setiawan, ‘PENDETEKSIAN KEKUATAN SENTIMEN PADA TEKS TWEET BERBAHASA INDONESIA MENGGUNAKAN SENTISTRENGTH’.
- [21] A. Kulkarni and A. Shivananda, *Natural language processing recipes: Unlocking text data with machine learning and deep learning using python*. Apress Media LLC, 2019. doi: 10.1007/978-1-4842-4267-4.
- [22] B. Aulia, P. E. P. Utomo, U. Khaira, and T. Suratno, ‘ANALISIS SENTIMEN TAGAR #INDONESIATERSERAH DI MASA COVID-19 MENGGUNAKAN METODE SENTISTRENGTH’, *Jurnal Komputer dan Informatika*, vol. 9, no. 2, pp. 207–213, Oct. 2021, doi: 10.35508/jicon.v9i2.4275.
- [23] D. Haryalesmana Wahid, ‘Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity’, *IJCCS*, vol. 10, no. 2, pp. 207–218, 2016.
- [24] ‘JEPIN (Jurnal Edukasi dan Penelitian Informatika)’.
- [25] ‘8. Analisis Perbandingan Metode Spelling Corrector Peter Norvig dan Spelling Checker BK-Trees pada (sinta 3)’.
- [26] ‘How to Write a Spelling Corrector’, 2007.
- [27] L. Ilmknun and Q. · Follow, ‘NLP Task: Penjelasan singkat tentang Spelling Correction’.

- [28] R. A. Abou Khachfeh, I. El Kabani, and Z. Osman, ‘A NOVEL ARABIC CORPUS FOR TEXT CLASSIFICATION USING DEEP LEARNING AND WORD EMBEDDING’, *BAU Journal - Science and Technology*, vol. 3, no. 1, Dec. 2021, doi: 10.54729/2959-331x.1014.
- [29] ‘122 (1)’.
- [30] ‘245437-aplikasi-untuk-membangun-corpus-dari-dat-a260333f’.
- [31] ‘Hingga Januari 2019, laman https’. [Online]. Available: <https://trustpositif.kominfo>
- [32] R. Chandra, A. Suhendra, M. Agung Sucipta Iskandar, and L. Yuniar Banowosari, ‘Building Corpus in Bahasa Indonesia for Pornographic Indicated Website Content’, 2019. [Online]. Available: <https://trustpositif.kominfo.go.id/>.
- [33] S. Jafar, S. Rohana Hariana Intiana, B. Wahidah, M. Khairussibyan, and P. Bahasa dan Sastra Indonesia, ‘PENGENALAN KORPUS DATA BAHASA PADA MAHASISWA PROGRAM STUDI PENDIDIKAN BAHASA DAN SASTRA INDONESIA FKIP UNIVERSITAS MATARAM’, 2023. [Online]. Available: <http://journal.unram.ac.id/index.php/darmadiksani>
- [34] Bangladesh. F. of E. and E. E. Khulna University of Engineering & Technology, Institute of Electrical and Electronics Engineers. Bangladesh Section, and Institute of Electrical and Electronics Engineers, 2019 *4th International Conference on Electrical Information and Communication Technology (EICT)*.
- [35] ‘Artikel Corpus BhsIndonesia (2)’.
- [36] J. Bata, ‘#AkuGalau: Korpus Bahasa Indonesia untuk Deteksi Emosi dari Teks’. [Online]. Available: www.youtube.com.
- [37] G. R. Bennett, ‘An Introduction to corpus Linguistics Part 1 Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers’, 2010. [Online]. Available: <http://www.press.umich.edu/titleDetailDesc.do?id=371534>
- [38] ‘Corpus_157Language (1)’.
- [39] ‘Text Book Corpus (1)’.

- [40] A. R. Adriansyah, ‘PENYUSUNAN KORPUS BERITA TERBUKA BERBAHASA INDONESIA’. [Online]. Available: <http://sketchengine.co.uk/>
- [41] V. Christanti Mawardi, N. Susanto, and D. Santun Naga, ‘Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levenshtein Distance Method’, in *MATEC Web of Conferences*, EDP Sciences, Apr. 2018. doi: 10.1051/matecconf/201816401047.
- [42] E. Erwina, T. Tommy, and M. Mayasari, ‘Indonesian Spelling Error Detection and Type Identification Using Bigram Vector and Minimum Edit Distance Based Probabilities’, *SinkrOn*, vol. 6, no. 1, pp. 183–190, Nov. 2021, doi: 10.33395/sinkron.v6i1.11224.
- [43] P. Gupta, ‘A context sensitive real-time Spell Checker with language adaptability’, Oct. 2019, [Online]. Available: <http://arxiv.org/abs/1910.11242>
- [44] P. E. Ltrc, M. Chinnakotla, and R. Mamidi, ‘Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning’. [Online]. Available: <https://github.com/PravallikaRao/SpellChecker>