SKRIPSI

NORMALISASI TEKS BERBASIS KAMUS SLANG DAN LEVENSHTEIN DISTANCE: IMPLEMENTASI PADA ANALISIS SENTIMEN EKSPEDISI SICEPAT EKSPRES DI TWITTER

Disusun dan diajukan oleh

DIKI SISWANTO

D421 16 316

DEPARTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
MAKASSAR
2022

LEMBAR PENGESAHAN SKRIPSI

NORMALISASI TEKS BERBASIS KAMUS SLANG DAN LEVENSHTEIN DISTANCE: IMPLEMENTASI PADA ANALISIS SENTIMEN EKSPEDISI SICEPAT EKSPRES DI TWITTER

Disusun dan diajukan oleh

DIKI SISWANTO

D42116316

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 25 November 2022 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
A. Ais Prayogi Alimuddin, S.T., M.Eng.	Anugrayani Bustamin, S.T., M.T.
Nip. 198305102014041001	Nip. 199012012018074001

Ketua Program Studi,

Dr. Indrabayu, S.T., M.T., M.Bus.Sys.
Nip. 19730716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Diki Siswanto

NIM : D421 16 316

Departemen : Teknik Informatika

Jenjang : S1

Menyatakan dengan ini karya tulisan saya berjudul:

NORMALISASI TEKS BERBASIS KAMUS SLANG DAN LEVENSHTEIN DISTANCE: IMPLEMENTASI PADA ANALISIS SENTIMEN EKSPEDISI SICEPAT EKSPRES DI TWITTER

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilalihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 14 November 2022

Yang menyatakan,

Diki Siswanto

KATA PENGANTAR

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Puji dan syukur penulis panjatkan kehadirat Allah SWT karena atas berkat dan rahmat-Nya sehingga dapat menyelesaikan tugas akhir dengan judul **"***Normalisasi Teks Berbasis Kamus Slang dan Levenshtein Distance: Implementasi pada Analisis Sentimen Ekspedisi SiCepat Ekspres di Twitter***"** sebagai salah satu persyaratan akademik untuk menyelesaikan program Strata-1 pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Dengan segala kerendahan hati, penulis menyadari bahwa dalam penyusunan dan penulisan laporan tugas akhir ini tidak lepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak, dari masa perkuliahan sampai dengan masa penyusunan tugas akhir ini. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua penulis yang selalu memberikan doa, dukungan semangat, dan motivasi, serta selalu sabar dalam mendidik penulis sejak kecil.
2. Bapak A. Ais Prayogi Alimuddin, S.T., M.Eng., selaku pembimbing I yang senantiasa memberikan saran-saran serta bantuan selama proses pengambilan data hingga selesainya sistem ini dibuat; juga kepada Ibu Anugrayani Bustamin, S.T., M.T., selaku pembimbing II yang senantiasa

menyediakan waktu, tenaga, pikiran, semangat, dan perhatian yang luar biasa dalam membimbing penulis menyusun tugas akhir ini.

3. Bapak Dr. Indrabayu S.T., M.T., M.Bus.Sys. selaku ketua Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang senantiasa memberikan semangat dan motivasi agar tugas akhir ini segera terselesaikan.
4. Segenap Dosen dan Staf Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah membantu dan memberikan banyak ilmu serta dukungan selama masa perkuliahan.
5. Teman-teman Teknik Informatika Angkatan 2016 (IGNITER16) selaku rekan belajar selama masa perkuliahan.
6. Sahabat-sahabat penulis: Muh. Khaeril Syam, Muh Raedi Radifan, Tuti Amalia, S.T., Ica Novita Sari, S.Pd., Lutfi Qadri, Dandi Shoreandi, dan Muh. Agung Alif Hidayat.
7. Seluruh pihak yang turut andil dalam penyusunan tugas akhir ini.

Akhir kata, penulis menyadari bahwa masih banyak kekurangan dalam penulisan tugas akhir ini. Oleh karena itu, penulis menerima segala bentuk masukan, kritik, dan saran untuk kesempurnaan tugas akhir ini. Penulis berharap semoga tugas akhir ini dapat bermanfaat bagi semua pihak yang membacanya.

Makassar, Oktober 2022

Penulis

ABSTRAK

Twitter merupakan salah satu media sosial populer yang dapat dimanfaatkan untuk mengektraksi opini dan sentimen publik terhadap suatu layanan atau jasa, termasuk jasa ekspedisi. Salah satunya adalah SiCepat Ekspres. Opini yang disampaikan masyarakat terhadap SiCepat Ekspres di Twitter dapat dianalisis untuk mengetahui persepsi pengguna terhadap layanan perusahaan tersebut. Akan tetapi, sebagai jejaring sosial dimana satu pengguna dapat secara bebas berkomunikasi atau berpendapat sesuai keinginannya membuat format data teks pada media sosial Twitter menjadi beragam, di antaranya muncul variasi tulisan, seperti kata singkatan, bahasa tidak baku, dan slang. Penelitian ini bertujuan untuk merancang sistem normalisasi teks *tweet* berbasis kamus slang dan algoritma *Levenshtein Distance*. Hasilnya kemudian diimplementasikan pada sistem klasifikasi sentimen menggunakan *pretrained-model* IndoBERT untuk mengetahui pengaruh antara data *tweet* yang dilakukan normalisasi dan tanpa normalisasi. *Dataset* yang digunakan berjumlah 865, dengan rincian 547 data berlabel negatif dan 318 data berlabel positif. *Dataset* dibagi menjadi 80% data latih dan 20% data uji. Pengujian dilakukan sebanyak 3 kali dengan dua skenario, yakni *dataset* yang dinormalisasi dan tanpa normalisasi. Hasilnya didapatkan rata-rata akurasi, *precision*, dan *recall* untuk model dengan *dataset* tanpa normalisasi masing-masing 94.41%, 90.78%, dan 92.75%. Sedangkan pada model dengan *dataset* yang dinormalisasi diperoleh akurasi 97.88%, *precision* 98.95%, dan *recall* 95.30%. Terjadi peningkatan akurasi sebesar 3.47% pada model dengan *dataset* yang dinormalisasi dibandingkan dengan *dataset* tanpa normalisasi.

Kata kunci: Normalisasi teks, Twitter, *Levenshtein Distance*, Kamus slang, IndoBERT, *Natural Language Processing.*

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI.....ii

PERNYATAAN KEASLIAN.....iii

KATA PENGANTAR.....iv

ABSTRAK.....vi

DAFTAR ISI.....vii

DAFTAR GAMBAR.....x

DAFTAR TABEL.....xi

BAB 1 PENDAHULUAN.....1

1.1 Latar Belakang.....1

1.2 Rumusan Masalah.....5

1.3 Tujuan Penelitian.....5

1.4 Manfaat Penelitian.....5

1.5 Batasan Masalah Penelitian.....6

1.6 Sistematika Penulisan.....6

BAB 2 TINJAUAN PUSTAKA.....8

2.1 Twitter.....8

2.2 SiCepat Ekspres.....9

2.3 Natural Language Processing.....11

2.4 Normalisasi Teks.....12

2.5 Levenshtein Distance.....13

2.6 Kamus Slang.....15

2.7 IndoCollex.....16
2.8 Analisis Sentimen.....16
2.9 BERT.....17
2.9.1 IndoBERT.....19
2.10 Metriks Evaluasi Sistem.....19
2.10.1 Akurasi.....24
2.10.2 Precision.....24
2.10.3 Recall.....25
2.11 Penelitian Terkait.....25
BAB 3 METODOLOGI PENELITIAN.....29
3.1 Tahapan Penelitian.....29
3.2 Waktu dan Lokasi Penelitian.....31
3.3 Instrumen Penelitian.....31
3.4 Teknik Pengambilan Data.....32
3.5 Perancangan Sistem.....33
3.5.1 Preprocessing.....34
3.5.2 Normalisasi Teks.....36
3.5.3 Klasifikasi Sentimen.....45
3.6 Analisis Kerja Sistem.....45
BAB 4 HASIL DAN PEMBAHASAN.....47
4.1 Dataset.....47
4.2 Implementasi Sistem Normalisasi Teks.....47

4.2.1 Pre-processing data.....47
4.2.2 Normalisasi Berbasis Kamus Slang.....49
4.2.3 Normalisasi Berbasis Levenshtein Distance.....50
4.2.4 Analisis Normalisasi.....56
4.3 Implementasi Sistem Klasifikasi Sentimen.....57
4.4 Pengujian Sistem.....58
BAB 5 PENUTUP.....62
5.1 Kesimpulan.....62
5.2 Saran.....63
DAFTAR PUSTAKA.....64
LAMPIRAN.....67

DAFTAR GAMBAR

Gambar 3.1: Diagram Tahapan Penelitian.....29

Gambar 3.2: Rancangan Sistem.....34

Gambar 4.1: Kode Program Normalisasi Teks Berbasis Kamus Slang.....49

Gambar 4.2: Kode Program Pembobotan Operasi Karakter.....51

Gambar 4.3: Kode Program Algoritma *Levenshtein Distance*.....53

Gambar 4.4: Kode Program Normalisasi Teks Berbasis *Levenshtein Distance*.....54

DAFTAR TABEL

Tabel 2.1: Contoh isi kamus IndoCollex.....16

Tabel 2.2: *Confusion Matrix*.....20

Tabel 2.3: *Multiclass Confusion Matrix*.....21

Tabel 2.4: Rincian nilai TP, FN, FP, dan TN pada kelas positif.....22

Tabel 2.5: Rincian nilai TP, FN, FP, dan TN pada kelas negatif.....23

Tabel 2.6: Rincian nilai TP, FN, FP, dan TN pada kelas netral.....24

Tabel 3.1: Inisialisasi baris dan kolom pada *Levenshtein Distance*.....40

Tabel 3.2: Isian matriks menggunakan *Levenshtein Distance*.....41

Tabel 3.3: Contoh skor kemiripan kandidat kata dengan *string* sumber.....43

Tabel 3.4: Transformasi karakter sumber ke karakter target.....44

Tabel 4.1: Rincian *dataset*.....47

Tabel 4.2: Sampel *tweet* setelah *preprocessing*.....48

Tabel 4.3: Normalisasi *tweet* berbasis kamus IndoCollex.....50

Tabel 4.4: Hasil normalisasi kata berbasis algoritma *Levenshtein Distance*.....55

Tabel 4.5: Jumlah kata yang dinormalisasi berdasarkan metode yang digunakan.56

Tabel 4.6: Kesalahan normalisasi pada algoritma *Levenshtein Distance*.....57

Tabel 4.7: *Hyperparameter* pada IndoBERT.....58

Tabel 4.8: Hasil pengujian sistem klasifikasi sentimen dengan dua skenario.....59

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Pengaruh globalisasi dan pesatnya perkembangan teknologi saat ini membuat platform media sosial dapat dijadikan sebagai medium untuk berkomunikasi, misalnya *chatting*, memberi komentar pada suatu topik, menyampaikan opini dan keluhan, dan lain-lain. Hal ini karena media sosial menawarkan kemudahan dan kecepatan dalam menyebarkan informasi. Twitter termasuk salah satu media sosial populer yang banyak digunakan masyarakat Indonesia. Dari data survei yang dihimpun Statista, Indonesia menduduki peringkat 5 (lima) sebagai negara dengan jumlah pengguna Twitter terbanyak di dunia per Januari 2022, yakni 18,45 juta pengguna.

Twitter memungkinkan para penggunanya mengirim pesan percakapan yang sering disebut dengan istilah *tweet*. Percakapan ini dapat diolah menjadi informasi yang berguna, misalnya untuk mengetahui persepsi atau pandangan masyarakat akan kinerja suatu layanan atau jasa. Ekspedisi pengiriman barang merupakan sektor yang menarik untuk dieksplorasi mengingat menjamurnya industri *e-commerce* saat ini juga berdampak pada peningkatan permintaan jasa ekspedisi yang aman dan cepat. Layanan ekspedisi banyak dibicarakan oleh masyarakat Indonesia di Twitter karena platform ini mendukung diskusi terbuka dan siapapun dapat menanggapi *tweet* yang diunggah oleh pengguna. Selain itu,

pengguna juga dapat melakukan *mention* akun perusahaan yang bersangkutan untuk mendapatkan atensi secara langsung apabila mengalami kendala atau masalah.

Salah satu pemain baru di industri jasa ekspedisi adalah SiCepat Ekspres. Perusahaan ini didirikan pada tahun 2014 dan saat ini sudah menjalin kerjasama dengan beberapa platform *e-commerce*. Berkat kerjasama dengan *e-commerce* Tokopedia dalam program 'Bebas Ongkir' membuat SiCepat Ekspres menjadi populer dan dibicarakan oleh masyarakat Indonesia di Twitter karena di awal-awal menawarkan pelayanan yang cepat dibandingkan pendahulunya. Opini yang disampaikan oleh masyarakat di Twitter mengenai pelayanan SiCepat Ekspres dapat diteliti melalui proses analisis sentimen. Analisis sentimen merupakan studi mengenai cara mengekstrak dan mengolah data tekstual guna mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini (Buntoro, 2017). Tujuannya adalah untuk mengetahui apakah positif atau negatif. Klasifikasi tersebut dapat menjadi gambaran seberapa baik atau buruk kualitas kinerja dan layanan dari ekspedisi SiCepat Ekspres. Tentunya hal ini dapat menjadi pertimbangan bagi khalayak sebelum memutuskan untuk menggunakan ekspedisi tersebut.

Namun, dalam upaya analisis sentimen terhadap data di Twitter terdapat beberapa hal yang penting diperhatikan. Di antaranya adalah sebuah *tweet* hanya dapat berisi sebanyak 280 karakter (Gligorić dkk., 2020). Dengan keterbatasan tersebut, teks dalam sebuah *tweet* cenderung mengandung variasi kata tidak baku,

seperti penggunaan singkatan dan bahasa slang (Hidayatullah, 2015). Berdasarkan survei data di Twitter mengenai SiCepat Ekspres ditemukan beberapa variasi kata tidak baku, misalnya ‘ga’, ‘gk’, ‘nyampe’, ‘cepet’, ‘pake’, ‘blm’, ‘kesel’, ‘nganter’, ‘ngirim’, dan ‘dateng’. Variasi tulisan ini kebanyakan berupa *phonetic (sound) alteration*, yaitu pengubahan sedikit pengucapan atau lafal pada kata, seperti ‘nyampe’, ‘dateng’, ‘nyangkut’ dan ‘cepet’. Selain itu, terdapat variasi lain berupa *informal affixation* atau modifikasi imbuhan, contohnya ‘nganter’ dan ‘ngirim’. Penggunaan slang juga ditemukan, misalnya kata ‘ga’, ‘gokil’, ‘pickup’, dan ‘stuck’. Kata-kata tersebut perlu diubah ke bentuk baku agar tidak dianggap sebagai kata yang berbeda dari kata dasar yang sebenarnya. Upaya ini dilakukan agar meminimalisasi kesalahan dalam proses klasifikasi sentimen. Hal ini mengingat model pemrosesan bahasa alami untuk Bahasa Indonesia saat ini kebanyakan dikembangkan dengan dasar bahasa baku (Akbarianto Wibowo dkk., 2020).

Berdasarkan latar belakang yang ada, penelitian ini mengusulkan normalisasi teks berbasis kamus slang dan *Levenshtein Distance*. Algoritma *Levenshtein Distance* ialah metode untuk menghitung jumlah operasi *string* paling sedikit yang diperlukan untuk mentransformasikan suatu *string* menjadi *string* yang lain (Adiwidya, 2009). Jika dibandingkan dengan metode lain yang juga digunakan untuk koreksi teks, yaitu *JaroWinkler Distance*, algoritma *Levenshtein Distance* menawarkan performa yang lebih baik berdasarkan nilai *accuracy, recall* dan *f1score* (Nur, 2021). Penelitian lain oleh (Nugraha & Rizqullah, 2019),

dilakukan skenario pengujian untuk normalisasi teks tidak baku yang tidak disingkat menggunakan *Longest Common Subsequence* (LCS), *Levenshtein Distance*, dan *JaroWinkler Distance*. Ketiganya dibandingkan dan hasilnya *Levenshtein Distance* mengungguli kedua algoritma lainnya dalam hal akurasi untuk dua skenario sekaligus. Oleh karena itu penelitian ini memilih *Levenshtein Distance* dan kamus slang untuk normalisasi teks. Algoritma *Levenshtein Distance* akan digunakan untuk mengoreksi kata singkatan dan variasi *phonetic alteration. S*edangkan kamus slang digunakan sebagai rujukan untuk mengubah kata-kata slang yang memiliki kemiripan dengan kata baku yang tersedia dari korpus.

Teks hasil normalisasi selanjutnya akan diimplementasikan dalam sistem analisis sentimen. Dengan melakukan normalisasi setelah tahap *preprocessing* diharapkan analisis sentimen yang dilakukan dapat memberikan hasil dan akurasi yang optimal. Untuk mengetahui pengaruh normalisasi terhadap akurasi klasifikasi sentimen yang dihasilkan, maka akan dilakukan pengujian sebelum dan sesudah perbaikan. Pada proses analisis sentimen akan menggunakan *Bidirectional Encoder Representations from Transformers* (BERT). Metode ini dipilih karena menawarkan hasil akurasi yang tinggi dan cukup efektif untuk diimplementasi pada analisis sentimen (Atmaja & Yustanti, 2021). Adapun *pretrained-model* BERT yang digunakan pada penelitian ini adalah IndoBERT (Koto dkk., 2020).

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka identifikasi masalah yang ditemukan, yaitu:

1. Bagaimana implementasi kamus slang dan *Levenshtein Distance* untuk normalisasi teks dalam proses analisis sentimen terhadap layanan ekspedisi SiCepat Ekspres dari media sosial Twitter?
2. Bagaimana hasil akurasi dari analisis sentimen layanan ekspedisi SiCepat Ekspres dengan menerapkan normalisasi teks berbasis kamus slang dan *Levenshtein Distance*?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Merancang dan mengimplementasikan kamus slang dan *Levenshtein Distance* untuk normalisasi teks pada proses klasifikasi sentimen terhadap layanan ekspedisi SiCepat Ekspres dari media sosial Twitter.
2. Mengetahui hasil akurasi dari analisis sentimen terhadap layanan ekspedisi SiCepat Ekspres dengan menerapkan normalisasi teks berbasis kamus slang dan *Levenshtein Distance*.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini yaitu dapat mengetahui pengaruh normalisasi teks terhadap tingkat keberhasilan sistem klasifikasi sentimen. Penelitian ini juga dapat digunakan untuk memperoleh gambaran umum mengenai persepsi

pengguna Twitter terhadap layanan jasa ekspedisi SiCepat Ekspres. Selain itu harapannya penelitian ini dapat menjadi bahan bacaan untuk penelitian terkait NLP di masa mendatang.

1.5 Batasan Masalah Penelitian

Berdasarkan rumusan masalah yang ada, maka batasan masalah yang ditetapkan agar penelitian ini lebih terstruktur dan terarah, yakni:

1. Fokus penelitian ini adalah normalisasi teks berbahasa Indonesia dari Twitter terkait ekspedisi SiCepat Ekspres
2. *Dataset* yang digunakan tidak mengandung kalimat sarkasme
3. Hasil normalisasi akan diimplementasikan dalam sistem analisis sentimen
4. Penelitian ini menggunakan *Levenshtein Distance* dan kamus slang dalam proses normalisasi teks, sedangkan pada proses klasifikasi sentimen menggunakan *pretrained-model* IndoBERT.
5. Data yang digunakan pada proses *stemming* menggunakan data dari KBBI
6. Data kamus slang yang akan digunakan yaitu IndoCollex (*Indonesian Colloquial Lexicon*)

1.6 Sistematika Penulisan

Untuk memberikan gambaran umum mengenai isi tulisan secara keseluruhan, berikut diuraikan sistematika penulisan dari laporan tugas akhir ini:

BAB 1 PENDAHULUAN

Bab ini berisi penjelasan tentang latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Bab ini akan menjelaskan landasan teori yang berkaitan dengan penelitian termasuk di dalamnya membahas mengenai Twitter, konsep dasar *Natural Language Processing*, serta metode-metode yang digunakan dalam penelitian ini.

BAB 3 METODOLOGI PENELITIAN

Bab ini berisi tentang tahapan penelitian mulai dari studi literatur, persiapan kebutuhan perancangan sistem, perancangan dan pembuatan sistem, penerapan normalisasi teks untuk analisis sentimen dan skenario pengujian sistem.

BAB 4 HASIL DAN PEMBAHASAN

Bab ini berisi tentang hasil penelitian dan pembahasan terkait sistem yang telah dibuat.

BAB 5 PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan dan mencantumkan saran-saran untuk pengembangan sistem yang lebih lanjut.

BAB 2

TINJAUAN PUSTAKA

2.1 Twitter

Twitter merupakan media sosial yang didirikan pada tahun 2006 oleh Jack Dorsey dengan mengusung konsep *microblogging*. *Microblogging* didefinisikan oleh (Murthy, 2012) sebagai layanan berbasis internet tempat pengguna dapat menyiarkan pesan singkat, pesan tersebut menjadi tersedia secara publik dan dapat dikumpulkan bersama-sama ke seluruh pengguna, kemudian pengguna dapat memutuskan pesan siapa yang ingin mereka terima, tetapi tidak perlu tahu siapa yang dapat menerima pesan mereka.

Pesan yang dikirim oleh pengguna Twitter dikenal dengan istilah *tweet*. *Tweet* yang diunggah dapat berupa teks, gambar, video dan audio. Dengan konsep *microblogging* yang diusung, jumlah karakter dalam sebuah *tweet* dibatasi maksimum sebanyak 140 karakter. Namun, pada tahun 2017 ditingkatkan menjadi dua kali lipat, yakni 280 karakter (Gligorić dkk., 2020).

Twitter menyediakan akses data kepada perusahaan, akademisi, dan *developer* melalui fasilitas *Application Programming Interface* (API). Adanya fasilitas ini memungkinkan siapa saja dapat memanfaatkan *tweet* yang telah diunggah oleh pengguna, misalnya untuk keperluan penelitian bagi perusahaan atau instansi guna memahami opini masyarakat (Aditya, 2015).

2.2 SiCepat Ekspres

SiCepat Ekspres merupakan sebuah perusahaan yang bergerak pada bidang jasa pengiriman barang yang mencakup seluruh wilayah Indonesia. Meskipun terbilang pemain baru, SiCepat telah menjadi perusahaan logistik yang banyak dikenal masyarakat sejak menjamurnya industri *e-commerce* di Indonesia. Perusahaan ini didirikan oleh Rudy Darwin Swigo bersama The Kim Hai pada tahun 2014 dengan nama PT SiCepat Ekspres.

Masyarakat yang sudah terbiasa dengan belanja *online*, membuat perusahaan jasa pengiriman seperti SiCepat berkembang pesat. Saat ini, SiCepat Ekspres telah memiliki lebih dari ratusan cabang yang berada di seluruh Indonesia. Selain itu, perusahaan jasa pengiriman ini juga bermitra dengan banyak perusahaan *e-commerce* dan *online shop* di Indonesia (Media62.ID, 2021).

Adapun *e-commerce* dan *online shop* yang menjadi mitra SiCepat Ekspres di antaranya:

- Tokopedia
- Shopee
- Bukalapak
- Lazada
- Blibli
- Jakmall
- ZALORA
- dan beberapa platform lainnya

Dalam upaya meningkatkan daya saing dengan kompetitor dan mendapatkan tempat terbaik di hati pelanggan, SiCepat menawarkan layanan pengiriman yang variatif yang disesuaikan dengan kebutuhan para pelanggannya. Adapun jenis layanan yang ditawarkan oleh SiCepat Ekspres, di antaranya: (Media62.ID, 2021)

1. SiUntung, yaitu layanan pengiriman dengan waktu sampai 15 jam dengan harga reguler
2. BEST, singkatan dari Besok Sampai Tujuan, yaitu layanan kiriman kilat dengan estimasi waktu satu hari tiba di tujuan.
3. SiCepat Gokil, yakni pengiriman barang minimum *charge* 10kg dengan harga yang ekonomis
4. SiCepat Halu, layanan pengiriman yang tersedia di *e-commerce* dengan harga mulai Rp5.000
5. COD, yakni pengiriman dengan bayar ke kurir saat barang sampai tujuan
6. SiCepat Syariah, merupakan program mengalokasikan nilai potongan ongkir melalui donasi untuk memajukan pendidikan di Indonesia
7. H3LO, yaitu layanan lebih irit ongkos kirim, apabila berat barang 3,3kg cukup bayar 2kg
8. SiCepat GO!, yakni jasa pengiriman internasional dengan biaya lebih hemat
9. SiCepat Klik, yaitu layanan pendukung aktivasi SiCepat dengan teknologi WhatsApp Bot

2.3 Natural Language Processing

Natural Language Processing (NLP) merupakan salah satu cabang ilmu irisan dari bidang linguistik dan Kecerdasan Buatan (*Artificial Intelligence*). *Natural Language Processing* didefinisikan sebagai suatu proses komputasi yang digunakan oleh komputer dalam upaya menganalisa, memahami, dan memperoleh makna dari bahasa manusia yang digunakan dalam kehidupan sehari-hari (Samudro, 2019). NLP diklasifikasikan menjadi dua bagian yaitu *Natural Language Understanding* dan *Natural Language Generation* yang mengembangkan tugas untuk memahami dan menghasilkan teks.

NLP dapat dimanfaatkan untuk melakukan tugas-tugas seperti peringkasan otomatis, penerjemahan bahasa, analisis sentimen, pengenalan ucapan, dan segmentasi topik. Namun proses kerja NLP tidaklah mudah, terdapat beberapa kendala terkait bahasa alami dalam berkomunikasi dengan komputer. Ambiguitas adalah salah satu kendala utama bahasa alami yang biasanya dihadapi di tingkat sintaksis yang memiliki sub-tugas sebagai leksikal dan morfologi yang berkaitan dengan studi kata dan pembentukan kata. Masing-masing level ini menyebabkan ambiguitas yang dapat diselesaikan dengan mengetahui kalimat secara utuh dan lengkap. Selain itu, terdapat kendala lain yaitu *non-standard word* yang terkait dengan variasi kata dalam suatu kalimat. Bentuk kalimat yang tidak terstruktur juga dapat menyulitkan dalam fase *preprocessing* teks, misalnya tidak menggunakan tata bahasa yang tepat, mengandung banyak singkatan, kesalahan pengetikan atau pengejaan, dan lain sebagainya. Karena itu, normalisasi

dibutuhkan untuk membuat kalimat yang tidak terstruktur bisa dipahami oleh mesin komputer.

2.4 Normalisasi Teks

Gaya penulisan informal yang digunakan oleh pengguna di media sosial menjadi kendala bagi kebanyakan *tools Natural Language Processing* (NLP). Sifat sosial media sebagai jejaring sosial dimana satu pengguna dapat secara bebas berkomunikasi atau berpendapat sesuai keinginannya membuat format data teks pada media sosial menjadi beragam. Menurut (Wibowo dkk., 2021) terdapat sembilan karakteristik Bahasa Indonesia yang umumnya termasuk dalam ragam tidak baku di media sosial, yakni:

1. *Disemvoweling*: penghilangan huruf vokal. Contoh: jangan → jgn
2. *Shortening/clipping*: pemendekan suku kata. Contoh: internet → inet
3. *Space/dash removal*: pemendekan dengan menghapus spasi atau dash dan bisa diganti dengan huruf lain. Contoh: teman-teman → teman2
4. *Phonetic (sound) alteration*: pengubahan sedikit pengucapan atau lafal pada kata. Contoh: pakai → pake
5. *Informal affixation*: modifikasi imbuhan. Contoh: mengajari → ngajarin
6. *Compunding* atau *acronym*: penyingkatan kata dengan menjadi akronim. Contoh: anak baru gede → abg
7. *Reverse*: pembalikan kata. Contoh: yuk → kuy

8. *Loan words*: pengambilan kata dari bahasa lain. Yang sering di Indonesia yaitu dari bahasa daerah atau Inggris. Contoh: bokap
9. *Jargon: tagline* atau kata yang dipopulerkan, bisa dari suatu *event* atau orang terkenal. Contoh: meneketehe

Adanya variasi-variasi ini menjadi tantangan ketika melakukan tugas terkait NLP. Hal ini karena pada umumnya model bahasa alami dilatih pada teks formal yang bersih seperti data berita. Salah satu solusi yang mungkin untuk masalah ini adalah normalisasi teks.

Normalisasi teks adalah suatu upaya mentransformasikan teks informal dan bersifat *noise* menjadi bentuk yang lebih baku. Dalam beberapa kasus, upaya ini meliputi pemetaan kata tidak baku *out-of-vocabulary* (OOV) ke *in-vocabulary* (IV) baku dengan tetap mempertahankan arti kalimat. Selain itu, normalisasi teks juga dapat mencakup modifikasi yang berada di luar pemetaan, misalnya mengganti, menghapus, atau menambahkan token kata atau tanda baca dan menggunakan huruf besar atau kecil (Lourentzou dkk., 2019).

2.5 Levenshtein Distance

Algoritma *Levenshtein Distance*, atau sering disebut dengan *Edit Distance* merupakan algoritma untuk mencari jumlah perbedaan antara dua buah *string*. Algoritma ini ditemukan pada tahun 1965 oleh seorang ilmuwan Rusia bernama Vladimir Levenshtein. Pada dasarnya, algoritma *Levenshtein Distance* menghitung jumlah minimum pentransformasian suatu *string* menjadi *string* lain yang meliputi penggantian, penghapusan, dan penyisipan (Adiwidya, 2009).

Algoritma ini menggunakan matriks dua dimensi dalam perhitungan nilai jarak edit (*edit distance*). Isian nilai pada matriks tersebut adalah jumlah operasi penghapusan, penyisipan dan penukaran yang dibutuhkan dalam mengubah *string* sumber ke *string* target. Rumus operasi penghapusan, penyisipan, dan penukaran karakter yang digunakan untuk mengisi nilai matriks adalah sebagai berikut: (Rosmala & Risyad, 2017)

$$D(s,t) = min\, D(s-1,t) + 1 \quad \text{(penghapusan)} \tag{2.1}$$

$$D(s,t) = min\, D(s,t-1) + 1 \quad \text{(penyisipan)} \tag{2.2}$$

$$D(s,t) = min\, D(s-1,t-1) + 1, sj \neq ti \quad \text{(penggantian)} \tag{2.3}$$

$$D(s,t) = min\, D(s-1,t-1), sj = ti \quad \text{(tidak ada perubahan)} \tag{2.4}$$

Keterangan:

s = *String* sumber $s(j)$ = Karakter *string* sumber ke-*j*

t = *String* target $t(i)$ = Karakter *string* target ke-*i*

D = Jarak edit *Levenshtein Distance*

Tahap awal dari algoritma *Levenshtein Distance*, yaitu melakukan penyeleksian panjang kedua *string* terlebih dahulu. Jika salah satu atau kedua *string* merupakan *string* kosong, jalannya algoritma ini berhenti dan memberikan hasil *edit distance* yang bernilai nol atau panjang *string* yang tidak kosong. Jika panjang *string* keduanya tidak nol, setiap *string* memiliki sebuah karakter terakhir, misalnya c1 dan c2. Misalnya bagian *string* pertama tanpa c1 adalah s1 dan bagian *string* kedua tanpa c2 adalah s2, dapat dikatakan penghitungan yang dilakukan adalah cara mentransformasikan s1+c1 menjadi s2+c2. Jika c1 sama

dengan c2, dapat diberikan nilai *cost* 0 dan nilai *edit distance*-nya adalah nilai *edit distance* dari pentransformasian s1 menjadi s2. Jika c1 berbeda dengan c2, dibutuhkan pengubahan c1 menjadi c2 sehingga nilai *cost*-nya 1. Akibanya, nilai *edit distance*-nya adalah nilai *edit distance* dari pentransformasian s1 menjadi s2 ditambah 1. Kemungkinan lain adalah dengan menghapus c1 dan mengedit s1 menjadi s2+c2 sehingga nilai *edit distance*-nya dari pentransformasian s1 menjadi s2+c2 ditambah 1. Begitu pula dengan penghapusan c2 dan mengedit s1+c1 menjadi s2. Dari kemungkinan-kemungkinan tersebut, dicarilah nilai minimal sebagai nilai *edit distance* (Adiwidya, 2009).

2.6 Kamus Slang

Slang adalah kata yang tidak memenuhi standar kamus Indonesia (KBBI), biasanya dalam bentuk singkatan atau istilah gaul yang muncul di masyarakat. Istilah slang muncul hampir di setiap kalimat opini di media sosial (Khomsah & Agus Sasmito Aribowo, 2020). Kumpulan istilah ini kemudian dihimpun menjadi satu korpus yang disebut dengan istilah kamus slang.

Konversi kosakata slang menjadi kosakata standar yang benar berdasarkan KBBI bisa menjadi salah satu langkah prapemrosesan yang dapat menghasilkan performa model *classifier* yang lebih baik (Khomsah & Agus Sasmito Aribowo, 2020). Konversi ini akan melibatkan kamus slang sebagai basis data dan rujukan dalam melakukan subtitusi slang ke bentuk bakunya. Jika kalimat yang dikirim menemukan padanan kata dalam kamus slang maka akan diganti dengan kata baku yang sesuai.

2.7 IndoCollex

IndoCollex merupakan korpus yang berisi kamus Bahasa Indonesia tingkat frasa dari bentuk tidak baku ke bentuk baku dalam format TSV (*Tab-Separated Values*). Kamus dibangun dengan menganotasi kata-kata informal yang paling sering muncul dari media sosial Twitter. Kamus IndoCollex terdiri atas 2.624 baris data dan dibagi dalam 2 kolom. Kolom pertama merupakan frasa informal dan kolom kedua adalah frasa formal. Adapun contoh isi kamus ini dapat dilihat pada Tabel 2.1.

Tabel 2.1: Contoh isi kamus IndoCollex

Indeks	Informal	Formal
97	Agan	Juragan
154	Anget	Hangat
211	Bacod	Berisik
240	Baper	Bawa perasaan
483	Cuan	Uang
662	Drakor	Drama Korea
705	Engga	Tidak
775	Fyi	Sekedar informasi
790	Gaje	Tidak jelas
827	Gercep	Gerak cepat

2.8 Analisis Sentimen

Analisis sentimen merupakan studi mengenai cara mengekstrak dan mengolah data tekstual guna mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini (Buntoro, 2017). Analisis sentimen mengklasifikasikan

suatu opini dalam teks ke dalam kategori positif atau negatif. Biasanya suatu opini juga dapat dikategorikan sebagai netral. *Subjectivity analysis* dan *opinion mining* merupakan istilah yang juga dipakai untuk menyebut analisis sentimen.

Tujuan utama dari analisis sentimen adalah untuk memproses, mengekstrak, merangkum, dan menganalisa informasi yang ada dalam teks melalui metode yang berbeda-beda, sehingga dapat menyimpulkan emosi dan sudut pandang yang diberikan oleh penulis dari teks tersebut, dan membagi kecenderungan emosional di teks melalui informasi subjektif yang terkandung di dalamnya (Fimoza, 2021).

Secara umum, analisis sentimen dibagi menjadi tiga tingkatan yaitu tingkat dokumen (*document level*), tingkat kalimat (*sentence level*), dan tingkat berbutirhalus (*fine-grained level*). *Document-level* dan *sentence-level* dapat pula dikategorikan ke dalam *coarse-grained level*. Metode dalam analisis sentimen terbagi menjadi dua jenis, yaitu *learning-based* dan *lexical-based*. *Learning-based* menggunakan data training dan data testing, sedangkan *lexical-based* menggunakan kamus (*opinion lexicon*) (Fimoza, 2021).

2.9 BERT

BERT merupakan kependekan dari *Bidirectional Encoder Representations from Transformers,* yakni model *Deep Learning* untuk NLP yang didasarkan pada *Transformer* di mana setiap elemen output terhubung ke setiap elemen input, dan bobot elemen dihitung secara dinamis berdasarkan hubungan antar elemen. Model ini diusulkan dan diterbitkan oleh Jacob Devlin dan rekan-rekannya dari Google

Research pada tahun 2018. Selanjutnya di tahun 2019, Google mengumumkan bahwa BERT mulai diimplementasikan di mesin pencarinya. Kemudian pada akhir tahun 2020, BERT digunakan di hampir semua kueri pencarian bahasa Inggris milik Google.

BERT dirancang untuk membantu komputer memahami arti bahasa yang ambigu dalam sebuah teks dengan menggunakan teks sekitarnya untuk membangun konteks. Kerangka BERT telah dilatih sebelumnya menggunakan teks dari Wikipedia dan dapat disesuaikan dengan kumpulan data pertanyaan dan jawaban. BERT dapat digunakan pada berbagai macam tugas bahasa, seperti *sentimen analysis* (analisis sentimen), *question answering* (penjawab pertanyaan), *text prediction* (prediksi teks), *text generation* (pembangkitan teks), dan *summarization* (peringkasan teks).

Language model yang sudah ada kebanyakan hanya dapat membaca input teks secara berurutan, baik kiri-ke-kanan atau kanan-ke-kiri, tetapi tidak dapat melakukan keduanya secara bersamaan. BERT dirancang untuk dapat membaca dua arah sekaligus. Kemampuan ini hadir berkat fitur pengenalan *Transformer* yang dikenal sebagai *bidirectionality*. *Transformer* adalah bagian dari model yang memberikan BERT peningkatan kapasitas untuk memahami konteks dan ambiguitas dalam bahasa. *Transformer* akan memproses setiap kata masukan yang memiliki relasi dengan kata lain dalam sebuah kalimat, dibanding memprosesnya satu per satu. Dengan melihat semua kata di sekitarnya, *Transformer*

memungkinkan model BERT untuk memahami konteks penuh dari kata tersebut, dan oleh karena itu lebih memahami maksud dari suatu kalimat.

2.9.1 IndoBERT

IndoBERT adalah *pre-trained model* berbasis *transformer* untuk Bahasa Indonesia yang mengikuti *style* BERT, tetapi dilatih murni sebagai *masked language model* menggunakan *framework* Huggingface. *Transformer encoder* IndoBERT mengikuti konfigurasi *default* dari BERT-Base (uncased), yakni 12 *hidden layer* (dimensi = 768), 12 *attention head*, dan 3 *feed-forward hidden layer* (dimensi 3,072) serta panjang maksimum 512 token per *batch* (Koto dkk., 2020).

IndoBERT dilatih menggunakan lebih dari 220 juta kata Bahasa Indonesia yang dikumpulkan dari tiga sumber utama, yakni (1) Wikipedia bahasa Indonesia (74 juta kata); (2) artikel berita dari Kompas, Tempo, dan Liputan6 (total 55 juta kata); dan (3) Korpus Web Indonesia (90 juta kata) (Koto dkk., 2020).

2.10 Metriks Evaluasi Sistem

Mengukur kinerja suatu model yang telah dibuat merupakan langkah penting dalam bidang *Machine Learning* dan *Natural Language Processing.* Hasil pengukuran yang dilakukan dapat menjadi pertimbangan dalam memilih model terbaik. Salah satu teknik yang dapat digunakan untuk mengukur kinerja suatu model khususnya sistem klasifikasi (misalnya: analisis sentimen) adalah *Confusion Matrix*.

Confusion Matrix merupakan metode evaluasi yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran dari proses klasifikasi. *Confusion Matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada *Confusion Matrix* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Tabel *Confusion Matrix* dapat dilihat pada Tabel 2.2

Tabel 2.2: *Confusion Matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Keterangan :

a) TP (*True Positive*) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas positif.

b) FN (*False Negative*) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas negatif.

c) FP (*False Positive*) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas positif.

d) TN (*True Negative*) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas negatif.

Untuk evaluasi pada sistem klasifikasi dengan *multiclass*, *confusion matrix* juga bisa digunakan, namun dengan beberapa ketentuan tambahan.

Tabel 2.3: *Multiclass Confusion Matrix*

		Prediksi		
		Positif	Negatif	Netral
Aktual	Positif	TPos	FPosNeg	FPosNet
	Negatif	FNegPos	TNeg	FNegNet
	Netral	FNetPos	FNetNeg	TNet

Tidak jauh berbeda dengan klasifikasi biner, *multiclass confusion matrix* juga memiliki elemen TP (*True Positive*), FN (*False Negative*), FP (*False Positive*), dan TN (*True Negative*). Berikut adalah ketentuan dalam menetapkan nilai elemen tersebut:

a) TP (*True Positive*) merupakan banyaknya data yang kelas aktualnya sama dengan kelas prediksinya.

b) FN (*False Negative*) merupakan total dari seluruh baris yang ditunjuk kecuali TP yang dicari.

c) FP (*False Positive*) merupakan total dari seluruh kolom yang ditunjuk kecuali TP yang dicari.

d) TN (*True Negative*) merupakan total dari seluruh kolom dan baris selain yang ditunjuk.

Berdasarkan Tabel 2.3 dapat dijabarkan nilai TP, FN, FP, dan TN untuk masing-masing kelas sebagai berikut:

1. Kelas positif

Rincian nilai TP, FN, FP, TN pada kelas positif sebagaimana ditunjukkan pada Tabel 2.4 adalah sebagai berikut:

- TP terletak pada sel (1,1) berwarna hijau
- FN terletak pada sel (1,2) dan (1,3) berwarna kuning
- FP terletak pada sel (2,1) dan (3,1) berwarna biru
- TN terletak pada sel (2,2), (2,3), (3,2) dan (3,3) berwarna merah

Tabel 2.4: Rincian nilai TP, FN, FP, dan TN pada kelas positif

		Prediksi		
		Positif	Negatif	Netral
Aktual	Positif	TP (1,1)	FN (1,2)	FN (1,3)
	Negatif	FP (2,1)	TN (2,2)	TN (2,3)
	Netral	FP (3,1)	TN (3,2)	TN (3,3)

2. Kelas negatif

Rincian nilai TP, FN, FP, TN pada kelas negatif sebagaimana ditunjukkan pada Tabel 2.5 adalah sebagai berikut:

- TP terletak pada sel (2,2) berwarna hijau
- FN terletak pada sel (2,1) dan (2,3) berwarna kuning
- FP terletak pada sel (1,2) dan (3,2) berwarna biru
- TN terletak pada sel (1,1), (1,3), (3,1) dan (3,3) berwarna merah

Tabel 2.5: Rincian nilai TP, FN, FP, dan TN pada kelas negatif

		Prediksi		
		Positif	Negatif	Netral
Aktual	Positif	TN (1,1)	FP (1,2)	TN (1,3)
	Negatif	FN (2,1)	TP (2,2)	FN (2,3)
	Netral	TN (3,1)	FP (3,2)	TN (3,3)

3. Kelas Netral

Rincian nilai TP, FN, FP, TN pada kelas netral sebagaimana ditunjukkan pada Tabel 2.6 adalah sebagai berikut:

- TP terletak pada sel (3,3) berwarna hijau
- FN terletak pada sel (3,1) dan (3,2) berwarna kuning
- FP terletak pada sel (1,3) dan (2,3) berwarna biru
- TN terletak pada sel (1,1), (1,2), (2,1) dan (2,2) berwarna merah

Tabel 2.6: Rincian nilai TP, FN, FP, dan TN pada kelas netral

		Prediksi		
		Positif	Negatif	Netral
Aktual	Positif	TN (1,1)	TN (1,2)	FP (1,3)
	Negatif	TN (2,1)	TN (2,2)	FP (2,3)
	Netral	FN (3,1)	FN (3,2)	TP (3,3)

Dengan dasar tabel *Confusion Matrix* kemudian dapat dilakukan penghitungan nilai akurasi, *presicion*, dan *recall*. Ketiga metriks tersebut sangat bermanfaat untuk mengukur performa dari *classifier* atau algoritma yang digunakan untuk melakukan prediksi.

2.10.1 Akurasi

Akurasi merupakan metode pengujian berdasarkan tingkat kedekatan antara nilai prediksi dengan nilai aktual. Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi.

Persamaan akurasi ditunjukkan pada persamaan 2.5 berikut.

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\,\% \tag{2.5}$$

2.10.2 Precision

Precision merupakan metode pengujian dengan melakukan perbandingan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh

informasi yang terambil oleh sistem baik yang relevan maupun tidak. Persamaan *precision* ditunjukkan pada persamaan 2.6 berikut.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2.6)$$

2.10.3 Recall

Recall merupakan metode pengujian yang membandingkan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi relevan yang ada dalam koleksi informasi (baik yang terambil atau tidak terambil oleh sistem).

Persamaan *recall* ditunjukkan pada persamaan 2.7 berikut.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2.7)$$

2.11 Penelitian Terkait

1. **Kusuma, R. M. R. W. P., & Yustanti, W. (2021). Analisis Sentimen Customer Review Aplikasi Ruang Guru Dengan Metode BERT (Bidirectional Encoder Representations from Transformers)**

   Penelitian ini dilakukan untuk analisa sentimen terhadap aplikasi Ruang Guru di Google Play Store. Data *review* komentar diambil dari fitur komentar yang ada di Google Play Store menggunakan teknik *scrapping*. Data yang digunakan berjumlah 5437 *records*. Model dan metode yang digunakan adalah model *pretrained* BERT. Pada model ini diperoleh nilai F1 Score adalah 98.9% dengan proporsi data latih dan data uji 70:30.

Kemudian, dilakukan evaluasi terhadap model dan menghasilkan nilai akurasi sebesar 99%, presisi sebesar 64.13%, *recall* sebesar 60.51%.

2. **Riyaddulloh, R., & Romadhony, A. (2021). Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk Gadget Pada Twitter.**

   Tujuan dari penelitian adalah untuk melakukan normalisasi kata slang dari *tweets gadget* serta melakukan analisa terhadap pengaruh normalisasi dan terhadap kinerja proses selanjutnya, yaitu klasifikasi teks. *Dataset* yang digunakan berisi 1000 *tweet* bertema *gadget*. Normalisasi kata non-baku pada *dataset* menggunakan model *Word2vec* dan kamus slang. Setelah itu diseleksi fitur dengan pembobotan TF-IDF. Hasil akurasi yang diperoleh dalam pengujian dengan menggunakan variabel *Train Data Size* sebesar 80%, *Test Data Size* sebesar 20% diperoleh hasil akurasi *dataset* tanpa normalisasi sebesar 88% dan diperoleh hasil akurasi *dataset* dengan normalisasi menggunakan *Word2vec* sebesar 91%. Sehingga dapat disimpulkan bahwa *dataset* setelah dilakukan proses normalisasi akan memiliki performansi klasifikasi yang lebih baik dibanding *dataset* yang belum dinormalisasi.

3. **Riady, A. W. R. (2019). Normalisasi Mikroteks Berdasarkan Phonetic pada Twitter Berbahasa Indonesia Menggunakan Algoritma Jaro-Winkler Distance dan Rule Based.**

Penelitian ini memadukan algoritma *Jaro-Winkler Distance* dengan *Rule-based* untuk menemukan solusi dalam menormalisasikan teks Twitter berbahasa Indonesia. Algoritma dan metode yang digunakan diimplementasikan ke dalam sistem dengan menggunakan bahasa pemrograman Python. Data yang digunakan sebanyak 400 *tweets* dalam format json. Output yang dihasilkan berupa kalimat yang telah dinormalisasi menggunakan algoritma *Jaro-Winkler Distance* dan *rule based* dengan acuan kamus data dari kateglo sebagai kata baku yang sebenarnya. Tetapi, ada beberapa hasil dari kata masukan yang tidak sesuai, karena dalam penelitian ini belum memperhatikan struktur kalimat. Penelitian ini mendapatkan akurasi sebesar 94% dengan nilai presisi 96%, *recall* 96% dan *f-score* 0.96.

4. **Saniyah, Z. (2019). Normalisasi Mikroteks Berbentuk Singkatan pada Teks Twitter Berbahasa Indonesia Menggunakan Algoritma Longest Common Subsequences.**

   Tujuan dari penelitian ini adalah memperbaiki kata dari *tweet* yang berbentuk singkatan menjadi bentuk kata baku yang sebenarnya sesuai kamus Bahasa Indonesia dengan menggunakan algoritma *Longest Common Subsequences*. Pada penelitian ini, sistem hanya dapat memberikan saran kata jika terdapat lebih dari satu pilihan saran kata dengan jumlah karakter terpendek yang sama. Penelitian ini tidak dapat menentukan satu dari beberapa pilihan saran kata, karena tidak melihat

hubungan atau struktur kalimat. Algoritma *Longest Common Subsequences* yang diterapkan untuk normalisasi menghasilkan tingkat akurasi 83%, presisi 90%, *recall* 87%, dan *f1-Score* 0.88 dengan jumlah data uji sebanyak 400 *tweet* berbahasa Indonesia.

5. **Antinasari, P., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku.**

   Metode klasifikasi *Naive Bayes* dengan perbaikan kata tidak baku diimplementasikan dalam penelitian ini guna memperoleh sentimen tentang opini film pada dokumen Twitter berbahasa Indonesia. Perbaikan kata tidak baku menggunakan normalisasi *Levenshtein Distance* dilakukan setelah tahap *pre-processing*. Normalisasi *Levenshtein Distance* memberikan pengaruh akurasi yang lebih baik terhadap hasil klasifikasi dibandingkan dengan *pre-processing* yang hanya dilakukan dengan perbaikan kata tidak baku menggunakan kamus_katabaku. Berdasarkan pengujian yang telah dilakukan, diperoleh hasil akurasi terbaik dari analisis sentimen tentang opini film pada dokumen Twitter berbahasa Indonesia menggunakan *Naive Bayes* dengan perbaikan kata tidak baku adalah sebesar 98.33% dan untuk *precision*, *recall*, dan *f-measure* adalah 96.77%, 100%, dan 98.36%.