

BAB II

TINJAUAN PUSTAKA

Pada bagian ini dijelaskan tentang beberapa penelitian terdahulu yang relevan dengan penelitian ini dari berbagai topik seperti *Natural Language Processing* (NLP) dan model *Named Entity Recognition* (NER).

Dash et al. (2023) menggunakan studi kasus NLP untuk merancang strategi pemasaran berkelanjutan pada penggunaan media sosial. Masalah utama yang dihadapi adalah tantangan lingkungan dan perubahan perilaku konsumen yang mempengaruhi pemasaran, serta media sosial dapat menjadi alat yang efektif dalam menyampaikan pesan keberlanjutan. Penggunaan NLP untuk menganalisis tren pemasaran hijau dan sosial secara mendalam. Penelitian ini menggunakan metodologi berbasis text mining dan Latent Semantic Analysis (LSA) dengan alat KNIME dan VOSviewer. Data yang digunakan mencakup 33 artikel dari database Scopus yang dipublikasikan antara tahun 1991 hingga 2022. Alur penelitian dimulai dari pengumpulan data dengan kata kunci spesifik seperti *green marketing* dan *sustainable marketing*. Selanjutnya menganalisis tekstual dan clustering untuk mengidentifikasi tren utama. Hasil penelitian mengungkap tiga tren utama dalam pemasaran berkelanjutan yaitu green marketing dan perilaku konsumen, pemasaran media sosial berkelanjutan, serta praktik pemasaran *Influencer* di Media Sosial. Media sosial terbukti berperan besar dalam membangun keterlibatan konsumen dan mempengaruhi perilaku mereka melalui kampanye yang emosional dan berorientasi lingkungan. Penelitian ini menegaskan bahwa pemasaran yang menggabungkan keberlanjutan dan media sosial dapat meningkatkan efektivitas kampanye pemasaran, tetapi memerlukan strategi berbasis data yang lebih inovatif dan kreatif. Para pemasar diharapkan memanfaatkan teknologi NLP dan pendekatan pemasaran berbasis media sosial untuk mendorong perubahan perilaku yang lebih hijau dan bertanggung jawab secara sosial.

Hodorog et al. (2022) menggunakan studi kasus Machine Learning (ML) dan NLP untuk mendeteksi kejadian secara real-time di kota pintar dengan menganalisis data media sosial terutama dari Twitter. Masalah utama yang terjadi adalah volume besar data tidak terstruktur yang membutuhkan analisis cepat untuk mendeteksi peristiwa yang mempengaruhi pengambilan keputusan kota. Penggunaan model NLP yang dikombinasikan dengan analisis regresi berganda untuk memahami hubungan antara jenis peristiwa, tingkat kepuasan warga, dan faktor lingkungan. Metodologi melibatkan pembuatan taksonomi risiko berbasis semantik dan pengujian beberapa model bahasa menggunakan data dari kota Cardiff. Model pembelajaran yang diterapkan mencakup AWD-LSTM dan ULMFiT yang dibandingkan dengan model seperti Naive Bayes dan Random Forest untuk klasifikasi peristiwa. Alur penelitian dimulai dengan penambahan data, normalisasi teks, klasifikasi berdasarkan data set risiko, dan penghitungan korelasi antara variabel. Hasil penelitian menunjukkan bahwa kombinasi AWD-LSTM dan ULMFiT mencapai akurasi 88,5% dalam mendeteksi peristiwa. Hubungan signifikan ditemukan antara kemacetan, faktor cuaca, dan kecelakaan lalu lintas, serta antara kepuasan warga dengan acara sosial dan antrian panjang. Analisis regresi juga mengidentifikasi bahwa faktor lingkungan seperti suhu dan kelembaban memiliki *impact* signifikan pada variabilitas peristiwa. Penelitian ini menegaskan bahwa warga dapat berfungsi sebagai sensor aktif melalui media sosial untuk memberikan wawasan yang membantu pengambilan keputusan berbasis data di kota pintar. Model NLP yang dioptimalkan dapat mempercepat deteksi kejadian penting sehingga memungkinkan kota untuk meningkatkan ketahanan dan keberlanjutannya.

Jiang et al. (2022) menggunakan studi kasus NER untuk pengembangan Tweebank-NER, sebuah korpus baru untuk NER pada Twitter. Permasalahan utama dalam pengolahan data Twitter adalah bahasa yang tidak baku, pendek, dan berstruktur tidak formal. Hal ini membuat kinerja model NLP standar menjadi kurang optimal. Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan membuat anotasi baru berbasis Tweebank V2 (TB2) dan melatih model NLP yang

sesuai untuk media sosial. Penelitian ini menggunakan pendekatan anotasi crowdsourcing dengan Amazon Mechanical Turk. Tweet diberi label sesuai dengan data set entitas bernama person (PER), organization (ORG), location (LOC), dan miscellaneous (MISC). Penggunaan skema BIO dan mengevaluasi dengan pairwise *F1-score* untuk memastikan kualitas anotasi. Sehingga menunjukkan bahwa tipe entitas MISC memiliki tingkat kesepakatan antar anotator yang paling rendah. Penelitian ini menghasilkan alur lengkap untuk membangun model NLP, termasuk tokenisasi, lemmatization, part-of-speech tagging, parsing, dan NER menggunakan *framework* Stanza. Model ini dibandingkan dengan berbagai pendekatan lain, seperti *spaCy*, FLAIR, dan BERTweet, untuk mengevaluasi kinerjanya dalam setiap tugas sintaksis. Model Stanza mencapai kinerja terbaik untuk tokenisasi dan lemmatization serta kinerja kompetitif dalam NER dan POS tagging. Hasil penelitian menegaskan bahwa model berbasis transformer seperti BERTweet memberikan performa terbaik untuk NER di Tweepbank-NER, tetapi model berbasis Stanza tetap memberikan alternatif yang ringan dan efektif. Kesimpulannya, anotasi baru dan pipeline NLP yang dikembangkan menyediakan sumber daya yang berharga untuk penelitian NLP di media sosial, memungkinkan penggunaan langsung untuk analisis tweet secara luas. Para peneliti dalam studi ini juga merilis dataset dan modelnya secara publik untuk mendukung pengembangan lebih lanjut.

Kocaman and Talby. (2022) menggunakan studi kasus NER untuk pengembangan berbasis Spark NLP untuk domain klinis dan biomedis. Tujuan untuk menciptakan sistem produksi yang skalabel, efisien, dan akurat untuk ekstraksi entitas dari teks medis tidak terstruktur. Fokus utama adalah mengatasi tantangan dalam menangani data teks besar di domain kesehatan, seperti data rekam medis elektronik yang sering kali tidak terstruktur dan sulit diolah. Permasalahan utama yang diidentifikasi termasuk tingginya kebutuhan akan sumber daya komputasi untuk model berbasis transformer, keterbatasan pendekatan konvensional yang kurang fleksibel, dan minimnya alat produksi yang siap pakai untuk kebutuhan klinis. Penelitian ini menggunakan arsitektur BiLSTM-CNN-Char yang dimodifikasi dan diterapkan dalam lingkungan Apache Spark. Model ini diuji

pada berbagai dataset biomedis publik dan tantangan klinis, seperti 2010 i2b2/VA, 2014 n2c2 de-identification, dan 2018 n2c2 medication extraction. Hasil menunjukkan bahwa model ini mengungguli solusi komersial seperti AWS Medical Comprehend dan Google Cloud Healthcare API dengan peningkatan akurasi sebesar 8,9% dan 6,7%. Selain itu model Spark NLP menunjukkan kemampuan untuk menangani data dalam skala besar dengan waktu perawatan pelatihan yang lebih singkat dibandingkan implementasi pada TensorFlow. Penelitian ini menyimpulkan bahwa pendekatan yang diusulkan tidak hanya efisien tetapi juga dapat diperluas ke bahasa lain tanpa memerlukan perubahan kode yang signifikan, menjadikannya alat yang sangat fleksibel untuk implementasi di lingkungan produksi.

Narayanan et al. (2020) menggunakan studi kasus NER untuk pengembangan arsitektur neural berbasis karakter pada teks campuran bahasa (code-mixed), khususnya pada tweet berbahasa Hindi-Inggris. Masalah utama yang dihadapi yaitu struktur teks tidak teratur, penggunaan singkatan, dan karakter non-standar. Hal ini menyulitkan algoritma dalam mempelajari pola dari data media sosial. Kebanyakan model NER sebelumnya dirancang untuk teks monolingual yang kaya sumber daya, sehingga kurang efektif untuk teks campuran bahasa. Penggunaan embedding kata dengan Word2Vec atau GloVe untuk representasi tingkat kata, serta model karakter-level bidirectional LSTM untuk menangkap informasi morfologis dan kontekstual di tingkat sub-kata menjadi solusi. Metodologi penelitian mencakup dua jalur yaitu jalur pertama menggunakan embedding kata yang dimasukkan ke model LSTM untuk klasifikasi entitas. Sementara jalur kedua menggunakan embedding berbasis karakter dengan LSTM bidirectional untuk mempertimbangkan konteks dari kedua arah teks. Alur penelitian diawali dengan pre-processing dataset dari tweet Hindi-Inggris yang dianotasi secara manual dengan standar BIO. Model pertama menggunakan embedding kata, sementara model kedua mengekstrak embedding karakter untuk menangkap fitur sintaksis dan morfologis. Hasil penelitian menunjukkan bahwa model berbasis karakter LSTM bidirectional menghasilkan skor F1 rata-rata 0,96, melampaui model berbasis kata dengan peningkatan yang signifikan dalam pengenalan entitas. Kesimpulannya,

model karakter-level lebih efektif untuk NER pada teks campuran bahasa karena mampu menangkap fitur lintas bahasa tanpa memerlukan rekayasa fitur manual yang rumit. Penelitian ini membuka peluang untuk pengembangan lebih lanjut dengan dataset yang lebih besar dan penerapan Conditional Random Fields (CRF) untuk meningkatkan akurasi.

Patton et al. (2020) menggunakan studi kasus NLP untuk pengembangan Contextual Analysis of Social Media (CASM) dalam memahami konteks komunikasi di media sosial. Masalah utama yang dihadapi adalah kurangnya pemahaman model NLP terhadap konteks budaya, geografis, dan sosial, yang dapat mengakibatkan kesalahan dalam mengklasifikasikan data terutama pada komunitas yang terpinggirkan. Contohnya pada penggunaan bahasa African American Vernacular English (AAVE). Metodologi penelitian mencakup penggunaan data Twitter dari remaja yang terlibat dalam jaringan kekerasan di Chicago, dengan metode pelabelan manual yang didukung oleh pakar sosial dan anotator yang dilatih dalam pengenalan konteks lokal. Alur penelitian melibatkan lima langkah yaitu baseline interpretation, proses anotasi (analisis teks, emoji, dan jaringan pengguna), interpretasi konteks, pelabelan data berdasarkan temuan, dan validasi komunitas untuk meningkatkan akurasi. Hasil penelitian menunjukkan bahwa model yang dilatih dengan label berbasis CASM mengungguli model berbasis pelabelan otomatis sederhana, dengan peningkatan f-score dari 52,00 menjadi 64,26. Pendekatan CASM menawarkan metode yang lebih kaya dan kontekstual untuk meningkatkan akurasi prediksi dalam NLP. Namun, penelitian ini juga menyoroti risiko etis terkait penggunaan data media sosial untuk pengawasan, mengingat potensi *impactnya* terhadap komunitas yang dimaksud.

Pitropakis et al. (2020) menggunakan studi kasus NLP untuk mendeteksi ujaran kebencian anti-imigrasi di Twitter. Tujuan penelitian untuk mengembangkan sistem klasifikasi otomatis untuk mendeteksi retorika anti-imigrasi berdasarkan dataset baru yang dibuat secara manual yang terdiri dari tweet terkait imigrasi dari tiga negara berbahasa Inggris (AS, Inggris, dan Kanada). Penelitian ini juga bertujuan untuk mengeksplorasi perbedaan antara penggunaan fitur teks berbasis n-

gram kata dan karakter dalam proses klasifikasi. Permasalahan utama yang dihadapi adalah kurangnya dataset teranotasi yang relevan dan kompleksitas bahasa sosial media, seperti penggunaan sarkasme, humor, dan variasi linguistik antar wilayah. Metodologi penelitian meliputi pengumpulan dataset tweet menggunakan API Twitter, preprocessing data seperti tokenisasi dan lemmatization, serta pembuatan representasi teks menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Tiga algoritma seperti Support Vector Machines (SVM), Naïve Bayes (NB), dan Logistic Regression (LR) digunakan untuk klasifikasi dengan evaluasi performa berdasarkan metrik *precision*, *recall*, dan *F1-score*. Hasil menunjukkan bahwa model SVM dengan fitur n-gram kata mencapai kinerja terbaik dengan *F1-score* 84%. Studi ini mengungkapkan bahwa fitur berbasis n-gram kata lebih efektif dibandingkan n-gram karakter dalam mendeteksi ujaran kebencian dengan *impact* signifikan dalam meningkatkan pemahaman terhadap wacana anti-imigrasi di media sosial. *Impact*nya meliputi peningkatan kesadaran privasi pengguna media sosial dan pengembangan metode klasifikasi ujaran kebencian yang lebih canggih untuk aplikasi di masa depan.

Shelar et al. (2020) menggunakan studi kasus NER untuk pengembangan dan perbandingan model yang digunakan untuk mengekstraksi informasi dari teks tidak terstruktur. Tujuan utamanya adalah untuk mengevaluasi kinerja alat dan algoritma yang berbeda, termasuk TensorFlow, *SpaCy*, dan Apache OpenNLP dalam membangun model NER yang dapat disesuaikan. NER penting untuk berbagai aplikasi seperti klasifikasi konten, pencarian informasi, dan chatbot. Masalah utama yang dihadapi adalah bagaimana mengelola data teks tidak terstruktur, seperti memisahkan entitas kompleks (contohnya, "South Africa") menjadi satu kesatuan, bukan entitas terpisah, serta menangani kebutuhan prediksi yang cepat dan akurat dalam berbagai domain. Metodologi penelitian melibatkan pengujian ketiga alat ini menggunakan dataset yang sama untuk mengevaluasi parameter seperti akurasi pelatihan, waktu perawatan prediksi, ukuran model, dan kemudahan pelatihan. Model yang dikembangkan dengan *SpaCy* memberikan hasil terbaik, dengan akurasi pelatihan dan *F1-score* mencapai 100%. Sebaliknya model TensorFlow

menghadapi tantangan seperti pemisahan entitas yang tidak sesuai. Sementara itu Apache OpenNLP memperlihatkan keterbatasan dalam menangani token yang tidak dikenal. Penelitian ini menyimpulkan bahwa *SpaCy* adalah pilihan terbaik untuk membangun model NER yang dapat disesuaikan karena kemampuannya memberikan hasil prediksi yang cepat, akurat, dan mendalam dengan mempertimbangkan kesalahan prediksi (false positives dan false negatives) untuk perbaikan lebih lanjut.

Tohma dan Kutlu. (2020) menggunakan studi kasus NLP untuk pengembangan Bahasa Turki yang memiliki struktur linguistik kompleks. Penelitian ini bertujuan untuk mengidentifikasi hambatan utama dalam pemrosesan Bahasa Turki, seperti ambiguitas morfologis, keragaman tata bahasa, dan kurangnya sumber daya berbasis NLP yang memadai. Studi ini juga mengkaji teknik, sistem, dan sumber daya yang telah dikembangkan untuk menangani tantangan ini, termasuk analisis morfologi, pengenalan entitas bernama, dan analisis semantik. Metodologi yang digunakan dalam penelitian ini melibatkan tinjauan literatur, pengembangan sistem berbasis NLP untuk pengujian linguistik bahasa Turki, dan pendekatan berbasis ML. Beberapa sistem yang diuji termasuk pengenalan entitas bernama berbasis aturan, sistem dialog berbasis bahasa Turki, dan analisis emosi berbasis teks. Hasil menunjukkan keberhasilan yang bervariasi, seperti tingkat akurasi 93,69% dalam pengenalan entitas bernama menggunakan jaringan saraf tiruan dan 95% keberhasilan dalam koreksi ejaan menggunakan algoritma "edit distance metric". Penelitian ini juga menekankan bahwa solusi yang dikembangkan untuk bahasa Turki dapat diterapkan pada bahasa aglutinatif lainnya dengan karakteristik serupa. Namun perlu kolaborasi yang lebih erat antara ahli linguistik dan spesialis teknologi informasi untuk menciptakan solusi yang lebih efektif.

Yilmaz et al. (2020) menggunakan studi kasus NER untuk pengembangan kerangka kerja teks media sosial dalam bahasa Turki, dengan memanfaatkan kombinasi Conditional Random Fields (CRF) dan Bidirectional Long Short-Term Memory (BiLSTM). Masalah utama dalam NER untuk media sosial adalah bahasa informal, penggunaan singkatan, dan kesalahan ejaan yang meningkatkan

tantangan dalam pengenalan entitas bernama dibandingkan teks formal. Solusi yang diusulkan melibatkan penggunaan fitur-fitur seperti representasi karakter, embedding kata, informasi morfologis, serta pendekatan berbasis kamus dan pola untuk memperbaiki akurasi prediksi. Metodologi yang digunakan mencakup pengumpulan data dari Twitter, Facebook, dan Donanimhaber. Hal ini menghasilkan dataset terbesar untuk NER bahasa Turki dengan 16 tipe entitas berbeda. Alur penelitian terdiri dari beberapa tahap yaitu normalisasi teks, ekstraksi fitur berbasis kata dan karakter, serta penggunaan model berbasis BiLSTM-CRF untuk prediksi entitas. Hasil penelitian menunjukkan bahwa kombinasi fitur yang disesuaikan dengan karakteristik teks media sosial dapat meningkatkan kinerja model secara signifikan, dengan *F1-score* mencapai 84.3% dalam pengenalan entitas. Kesimpulannya, penelitian ini menyoroti pentingnya adaptasi model NLP untuk domain spesifik seperti media sosial, di mana bahasa yang digunakan sangat berbeda dari teks formal. Kontribusi utama berupa pengembangan model hybrid dan dataset yang terbuka untuk umum, memberikan peluang lebih luas dalam pengembangan aplikasi NLP di lingkungan berbahasa Turki.

Tabel 2.1 menunjukkan beberapa penelitian yang telah dilakukan sebelumnya terkait dengan topik NLP dan NER :

Tabel 2. 1 Daftar review penelitian terdahulu

No	Fokus	Metode & Data	Hasil	Referensi
1	Merancang strategi pemasaran berkelanjutan melalui media sosial menggunakan NLP untuk mengatasi tantangan lingkungan dan perubahan perilaku konsumen	- Data : 33 artikel dari database Scopus (1991–2022) dengan Kata kunci utama: "green marketing" dan "sustainable marketing" - Metodologi : Menggunakan text mining dan Latent Semantic Analysis (LSA). - Proses: Pengumpulan data, Analisis tekstual, Clustering untuk identifikasi tren.	- Terdapat 3 (tiga) tren utama: a. Green Marketing dan Perilaku Konsumen. b. Pemasaran Media Sosial Berkelanjutan. c. Praktik Pemasaran <i>Influencer</i> di Media Sosial. - Media sosial efektif dalam membangun keterlibatan konsumen melalui kampanye emosional dan berorientasi lingkungan.	(Dash et al., 2023)
2	Menggunakan ML dan NLP untuk mendeteksi kejadian real-time di kota pintar melalui data media sosial terutama Twitter	- Data: Hubungan antara jenis peristiwa, kepuasan warga, dan faktor lingkungan dari media sosial kota cardiff.	- Kombinasi AWD-LSTM dan ULMFiT mencapai akurasi 88,5% dalam mendeteksi peristiwa.	(Hodorog et al., 2022)

		- Metodologi: Penambahan data, Normalisasi teks, Klasifikasi risiko, Korelasi variabel - Pengujian model bahasa: AWD-LSTM, ULMFiT, Naive Bayes, dan Random Forest	- Hubungan signifikan antara: a. Kemacetan, faktor cuaca, dan kecelakaan lalu lintas. b. Kepuasan warga dengan acara sosial dan antrian panjang. - Faktor lingkungan seperti suhu dan kelembaban mempengaruhi variabilitas peristiwa.	
3	Pengembangan Tweepbank-NER untuk NER di Twitter untuk mengatasi tantangan bahasa tidak baku dan struktur tidak formal pada data Twitter.	- Data: Twitter - Metodology: a. Pendekatan anotasi crowdsourcing menggunakan Amazon Mechanical Turk b. Skema BIO untuk pelabelan dan evaluasi kualitas anotasi dengan pairwise <i>F1-score</i>. c. Pengembangan pipeline NLP: tokenisasi, lemmatization, POS tagging, parsing, dan	- Tipe entitas MISC memiliki tingkat kesepakatan antar anotator terendah. - Model Stanza terbaik untuk tokenisasi dan lemmatization, serta kompetitif untuk NER dan POS tagging. - BERTweet memberikan performa terbaik untuk NER di Tweepbank-NER.	(Jiang et al., 2022)

		NER menggunakan <i>framework</i> Stanza. d. Perbandingan kinerja model: Stanza, <i>spaCy</i> , FLAIR, dan BERTweet dalam tugas sintaksis.		
4.	Pengembangan model <i>Named Entity Recognition</i> (NER) berbasis Spark NLP untuk domain klinis dan biomedis untuk menciptakan sistem produksi yang skalabel, efisien, dan akurat untuk ekstraksi entitas dari teks medis tidak terstruktur.	- Data : Dataset biomedis publik: 2010 i2b2/VA, 2014 n2c2 de-identification, dan 2018 n2c2 medication extraction. - Metodologi : a. Menggunakan arsitektur BiLSTM-CNN-Char yang dimodifikasi b. Implementasi dalam lingkungan Apache Spark untuk menangani data skala besar.	- Model mengungguli solusi komersial seperti AWS Medical Comprehend dan Google Cloud Healthcare API, dengan peningkatan akurasi sebesar 8,9% dan 6,7%. - Waktu perawatan pelatihan lebih singkat dibandingkan implementasi pada TensorFlow.	(Kocaman & Talby, 2022)
5.	Pengembangan arsitektur neural berbasis karakter untuk NER pada tweet campuran	- Data: Dataset tweet Hindi-Inggris yang dianotasi secara manual dengan skema BIO	- Model karakter-level bidirectional LSTM mencapai skor F1 rata-rata 0,96.	(Narayan et al., 2020)

	bahasa (code-mixed),terkhususnya Hindi-Inggris untuk mengatasi tantangan teks tidak teratur, singkatan, dan karakter non-standar di media sosial.	- Metodologi: a. Menggunakan embedding kata (Word2Vec, GloVe) dan embedding karakter dengan bidirectional LSTM. - b.Tahapan dilakukan 2 (dua) jalur: Jalur pertama: Embedding kata, LSTM untuk klasifikasi entitas. Jalur kedua: Embedding karakter, LSTM bidirectional untuk konteks dua arah. - Pengujian : Perbandingan kinerja model berbasis kata dan karakter.	- Model berbasis karakter lebih efektif dibandingkan model berbasis kata.	
6.	Pengembangan <i>Contextual Analysis of Social Media</i> (CASM) untuk memahami	- Data: Twitter dari remaja yang terlibat dalam jaringan kekerasan di Chicago.	- Model berbasis CASM meningkatkan f-score dari 52,00 menjadi 64,26	(Patton et al., 2020)

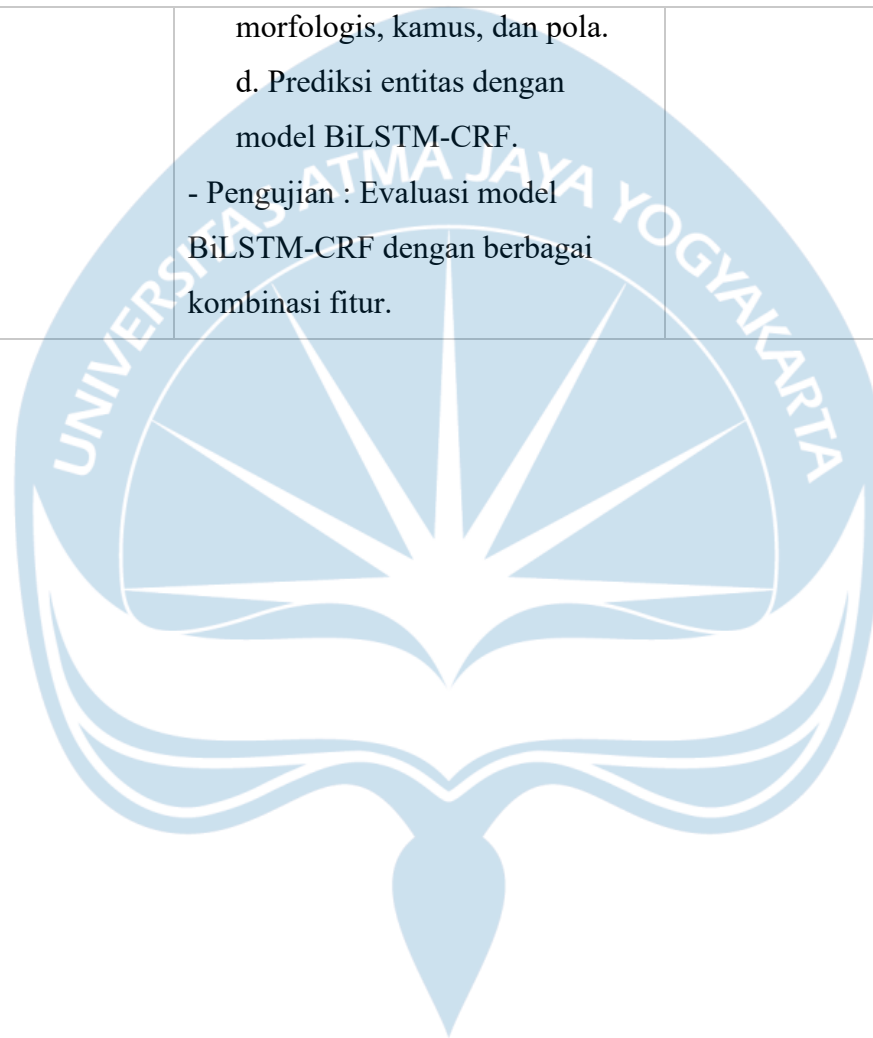
konteks komunikasi di media sosial untuk mengatasi kekurangan model NLP dalam memahami konteks budaya, geografis, dan sosial, terutama untuk komunitas terpinggirkan seperti pengguna AAVE.	- Metodologi: Pelabelan dilakukan secara manual dengan dukungan pakar sosial dan annotator terlatih. - Tahapan utama : a. <i>Baseline interpretation</i>. b. Proses anotasi (analisis teks, emoji, dan jaringan pengguna). c. Interpretasi konteks. d. Pelabelan data berdasarkan temuan. e. Validasi komunitas untuk meningkatkan akurasi. - Pengujian : Perbandingan antara model berbasis CASM dan pelabelan otomatis sederhana.	dibandingkan model pelabelan otomatis. - Pendekatan CASM memberikan analisis lebih kontekstual dan akurat dalam prediksi NLP.	
---	--	--	--

7.	Mengembangkan sistem klasifikasi otomatis untuk mendeteksi ujaran kebencian anti-imigrasi di Twitter serta mengeksplorasi perbedaan efektivitas fitur berbasis n-gram kata dan karakter.	- Data: Dataset manual tweet terkait imigrasi dari tiga negara berbahasa Inggris (AS, Inggris, Kanada). - Metodologi: a. Pengumpulan data menggunakan API Twitter. b. <i>Preprocessing</i> data: tokenisasi, lemmatization. c. Representasi teks menggunakan TF-IDF. d. Klasifikasi menggunakan algoritma: Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR). e. Evaluasi performa dengan metrik <i>Precision</i>, <i>Recall</i>, dan <i>F1-score</i>.	- Model SVM dengan fitur n-gram kata mencapai <i>F1-score</i> tertinggi (84%). - Fitur n-gram kata lebih efektif dibandingkan n-gram karakter dalam mendeteksi ujaran kebencian.	(Pitropakis et al., 2020)
----	--	--	---	---------------------------

		- Pengujian : Perbandingan performa antara model SVM, NB, dan LR dengan berbagai fitur teks.		
8.	Pengembangan dan perbandingan model <i>Named Entity Recognition</i> (NER) untuk mengekstraksi informasi dari teks tidak terstruktur.	- Data: Dataset yang sama digunakan untuk menguji ketiga model. - Pengujian alat berdasarkan parameter: a. Akurasi pelatihan. b. Waktu perawatan prediksi. c. Ukuran model. d. Kemudahan pelatihan.	- <i>SpaCy</i> menunjukkan hasil terbaik dengan akurasi pelatihan dan skor F1 mencapai 100%. - TensorFlow menghadapi masalah pemisahan entitas yang tidak sesuai. - Apache OpenNLP memiliki keterbatasan dalam menangani token tidak dikenal.	(Shelar et al., 2020)
9.	Mengidentifikasi hambatan dalam pemrosesan bahasa Turki yang memiliki struktur linguistik kompleks untuk menangani tantangan seperti	- Data: Bahasa Turki - Metodologi: a. Tinjauan literatur.	- Tingkat akurasi 93,69% untuk pengenalan entitas bernama menggunakan jaringan saraf tiruan.	(Tohma & Kutlu, 2020)

	ambiguitas morfologis, keragaman tata bahasa, dan kurangnya sumber daya NLP.	b. Pengembangan dan pengujian sistem berbasis NLP untuk analisis linguistik bahasa Turki. c. Pendekatan berbasis ML. - Pengujian : Pengujian berbagai sistem dan algoritma untuk tugas NLP pada bahasa Turki.	- Keberhasilan 95% dalam koreksi ejaan dengan algoritma "edit distance metric". - Solusi yang dikembangkan dapat diterapkan pada bahasa aglutinatif lain.	
10.	Pengembangan kerangka kerja NER untuk teks media sosial dalam bahasa Turki untuk mengatasi tantangan bahasa informal, singkatan, dan kesalahan ejaan di media sosial.	- Data: Dataset dari Twitter, Facebook, dan Donanimhaber. - Metodologi: Kombinasi Conditional Random Fields (CRF) dan Bidirectional Long Short-Term Memory (BiLSTM). - Tahapan : a. Normalisasi teks. b. Ekstraksi fitur berbasis kata dan karakter. c. Penggunaan informasi	- Model hybrid mencapai skor F1 sebesar 84,3% dalam pengenalan entitas. - Kombinasi fitur yang disesuaikan meningkatkan kinerja model secara signifikan.	(Yilmaz et al., 2020)

		morfologis, kamus, dan pola. d. Prediksi entitas dengan model BiLSTM-CRF. - Pengujian : Evaluasi model BiLSTM-CRF dengan berbagai kombinasi fitur.		
--	--	---	--	--



Berdasarkan hasil studi literatur yang telah dilakukan, konteks analisis data menggunakan NLP dan NER banyak berfokus pada analisis sentimen atau pengelompokan sederhana berdasarkan kata kunci. Namun, penelitian yang secara khusus mengintegrasikan NER untuk mengekstraksi dan mengelompokkan label spesifik, seperti kandungan *skincare* yang relevan dengan tipe kulit pengguna dari data tidak terstruktur, seperti komentar aplikasi *TikTok* masih belum ditemukan. Kekosongan ini menunjukkan bahwa belum ada penelitian yang secara eksplisit mengembangkan korpus dan model NER yang dapat menangani data spesifik ini.

Penelitian ini bertujuan untuk mengisi *gap* tersebut dengan membangun korpus dari data komentar aplikasi *TikTok* yang berisi ulasan *skincare* dan mengembangkan model NER untuk mengekstraksi dan mengidentifikasi informasi penting, seperti kandungan *skincare*, jenis kulit, permasalahan kulit, tips, jenis produk, waktu perawatan, perawatan, dan *impact*. Fokus ini memberikan kontribusi baru dibandingkan penelitian sebelumnya yang hanya mengandalkan analisis berbasis sentimen atau kata kunci. Selain itu, penelitian ini bertujuan untuk mengolah data tidak terstruktur menjadi informasi yang terorganisir, yang nantinya dapat digunakan untuk mendukung pemahaman konsumen dalam memilih produk *skincare* yang sesuai untuk penelitian selanjutnya. Penelitian ini tidak hanya berkontribusi pada literatur NLP dan NER, tetapi juga memberikan manfaat praktis bagi industri kecantikan dalam memahami preferensi dan kebutuhan konsumen secara lebih mendalam.