VOCATECH: Vocational Education and Technology Journal 7,1 (2025): 82-99
Website: http://ojs.aknacehbarat.ac.id/index.php/vocatech/index
ISSN 2716-5183 (online) ISSN 2686-4770 (print)

# LOCATION ENTITY RECOGNITION IN INSTAGRAM CAPTIONS USING SUPPORT VECTOR MACHINE ALGORITHM

**Cut Hilma Arifa[1]**
**Rizal Tjut Adek[1]**
**Yesy Afrillia[1*]**

[1]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Malikussaleh, Jl. Kampus Unimal Bukit Indah, Blang Pulo, Kecamatan Muara Satu, Lhokseumawe, Provinsi Aceh 24335, INDONESIA

**Abstract**

*The rapid advancement of digital technology has significantly influenced productivity and facilitated access to information in daily life, particularly through the widespread use of social media. Instagram is one of the most popular platforms, where text in captions often contains location-related information that can be utilized for spatial analysis. This study aims to identify and classify location entities in Instagram captions using Support Vector Machine algorithm combine with rule-based Named Entity Recognition approach. The method involved linguistic feature extraction based on explicit spatial context, data labeling, model training, and performance evaluation using standard classification metrics: accuracy, precision, recall, and f1-score. Dataset consists of 400 captions primarily written in Indonesian, though some contain mixed-language elements such as foreign term or regional dialect. The dataset is divided into 70% training data ad 30% testing data. Experimental results show that model achieved an accuracy of 90,83%, precision of 97,01%, recall of 87,84%, and f1-score of 92,90%. Evaluation of three NER rules (exact match keyword, prepositional patterns, and descriptive structures) indicates that the combination of all rules yields the highest f1-score (89%), while the best-performing individual rule is the prepositioning pattern (74%). These results demonstrated strong performance in processing varied and unstructured Instagram captions. The combinations of SVM and NER rule-based prove effective in identifying and classifying spatial information into two classes Contains Location and No Location. This approach shows potential for implementation in text-based spatial analysis systems, such as location-based recommendation systems, geographic mapping, and location-based decision support systems.*

**Keywords:**
*Instagram; machine learning; named entity recognition; natural language processing; support vector machine.*

**Abstrak**

Perkembangan teknologi digital yang pesat secara signifikan berpengaruh meningkatkan produktivitas dan kemudahan akses informasi dalam kehidupan sehari-hari, salah satunya penggunaan media sosial yang semakin meluas. Instagram merupakan salah satu platform yang banyak digunakan, dimana teks dalam caption memiliki informasi terkait lokasi yang dapat dimanfaatkan untuk analisis spasial. Penelitian ini bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas lokasi dalam caption Instagram menggunakan algoritma *Support Vector Machine* (SVM) dengan pendekatan *Named Entity Recognition* (NER) *rule-based*. Metode yang digunakan meliputi ekstraksi fitur berbasis linguistik dengan konteks spasial eksplisit, lebelisasi data, pelatihan model, serta evaluasi kinerja model menggunakan matriks klasifikasi: akurasi, presisi, *recall* dan f1-score. Dataset terdiri dari 400 caption umumnya berbahasa Indonesia, namun terdapat unsur bahasa campuran seperti istilah asing atau bahasa daerah. Fokus utama penelitian diarahkan pada pengolahan dan pemahaman teks berbahasa Indonesia. Dataset dibagi menjadi 70% data *training* dan 30% data *testing*. Hasil pengujian menunjukkan bahwa model mendapatkan akurasi sebesar 90,83%, presisi 97,01%, *recall* 87,84% dan f1-score 92,90%. Evaluasi terhadap tiga *rule* NER (*exact match keyword*, pola preposisi, dan struktur deskriptif) menunjukkan bahwa pengenalan entitas berdasarkan gabungan seluruh *rule* memberikan f1-score tertinggi (89%), sementara *rule* individual terbaik adalah pola preposisi (74%). Nilai ini menunjukkan kinerja yang cukup baik dalam pengolahan caption Instagram yang variatif dan tidak terstruktur. Kombinasi metode SVM dan NER *rule-based* terbukti efektif dalam mengidentifikasi dan mengklasifikasi informasi spasial dalam dua kelas *Contain Location* dan *No Location*. Pendekatan ini berpotensi diterapkan pada sistem analisis spasial berbasis teks, seperti sistem rekomendasi lokasi, pemetaan geografis, dan pendukung keputusan berbasis lokasi.

**Kata Kunci:**
*Instagram; machine learning; named entity recognition; natural language processing; support vector machine.*

DOI: 10.38038/vocatech.v7i1.238
Received: 16 Juli 2025; Accepted: 10 Agustus 2025; Published: 13 Agustus 2025

**Citation in APA Style:** *VOCATECH: Vocational Education and Technology Journal, 7*(1), 82-99.

*Corresponding author:
Yesy Afrillia, Jurusan Teknik Informatika, Universitas Malikussaleh, Jl. Kampus Bukit Indah, Lhokseumawe, Provinsi Aceh 24335, INDONESIA
Email: yesyafrillia@unimal.ac.id

## 1. PENDAHULUAN

Teknologi membawa perubahan yang signifikan dalam aktifitas keseharian manusia. Dengan sarana teknologi seperti komputer, *smartphone,* ataupun perangkat digital lainnya semakin memudahkan akses berbagai informasi, seperti informasi pendidikan, kesehatan, hiburan dan lainnya. Teknologi juga mendorong peningkatan produktivitas dalam kegiatan sehari-hari melalui penggunaan aplikasi maupun perangkat lunak yang mendukung kegiatan keseharian (Novian et al., 2024).

Pengenalan entitas bernama atau *Named Entity Recognition* (NER) merupakan salah satu elemen utama dalam pemrosesan teks otomatis (Ashok & Lipton, 2023) , bertujuan untuk mengidentifikasikan entitas bernama dalam teks yang tidak terstruktur seperti atrikel berita, dan postingan media sosial (Budi & Suryono, 2023). Pengertian dari entitas dalam konteks ini dapat berupa objek nyata, meliputi nama individu, lokasi, instansi (Santoso et al., 2020) dan juga objek abstrak seperti konsep dan hubungan. Pengidentifikasian entitas dalam dokumen teks bertujuan untuk menyederhanakan pencarian dan analisis informasi yang didasarkan pada lebelisasi setiap kata entitas terhadap frasa dalam teks sesuai jenis entitasnya.

Media sosial khususnya Instagram, salah satu platform paling popular untuk mengekspresikan opini, berbagi pengalaman maupun berkomunikasi lintas batas. *Caption* postingan Instagram memiliki informasi tekstual yang berpotensi mengandung entitas lokasi. Pengidentifikasian entitas dalam dokumen teks bertujuan untuk memudahkan pencarian informasi yang didasarkan pada pembagian nama entitas terhadap tiap kata yang ada dalam teks (Safrizal, 2019). Informasi ini dapat dimanfaatkan lebih lanjut dalam konteks analisa spasial, pemetaan sosial, maupun pengambilan keputusan berbasis lokasi. Namun demikian, proses indentifikasi entitas lokasi tidaklah mudah. Karakteristik bahasa dari teks yang diperoleh pada media sosial cenderung berbahasa informal, tidak baku, singkat dan sering berupa campuran bahasa atau dialek, hal ini menjadi kendala proses ekstraksi entitas secara akurat. Tidak semua entitas lokasi dapat diperoleh langsung dalam format yang dikenali menggunakan metode konventional berbasis kamus atau *exact matching*.

Oleh karena itu, diperlukan pendekatan yang mampu untuk mengidentifikasi pola linguistik secara fleksible dan kontekstual dalam klasifikasi entitas lokasi. Salah satu pendekatan yang dapat diterapkan adalah algoritma *Support Vector Machine* (SVM) (Ashok & Lipton, 2023), yaitu salah satu algoritma *machine learning* tradisional yang dikenal efektif dalam menangani klasifikasi data berbasis teks.

Studi terdahulu terkait klasifikasi teks ekstraksi entitas lokasi, orang, organisasi dan detail kejadian dari berita online berkaitan dengan kebakaran. Penelitian dilakukan dengan pendekatan *Bidirectional* LSTM-CNNs untuk mengekstraksi entitas lokasi kebakaran pada berita online berbahasa Indonesia. Performa hasil penelitian yaitu sebesar 75% angka akurasi, *recall*, *presicion* dan juga berhasil menampilkan titik persebaran lokasi kebakaran hasil klasifikasi teks berita online dari rentang tanggal 1 Januari 2020 sampai 20 April 2020 (Putra et al., 2021).

Penelitian ini berfokus pada analisis teks postingan Instagram, dengan entitas lokasi sebagai analisa utama pendekatan NER. Setiap teks postingan (*caption*) Instagram akan diklasifikasikan ke dalam dua kelas, yaitu *Contains_Location* dan *No_Location*, berdasarkan keberadaan entitas lokasi dalam teks. SVM merupakan pilihan yang efisien diterapkan pada data berukuran terbatas, karena tidak memerlukan sumber daya komputasi yang besar seperti metode *deep learning* BERT atau LSTM. Meski demikian, SVM tetap mampu menghasilkan teknik ekstraksi fitur berbasis aturan *(rule-based).* Untuk mendukung proses klasifikasi ini, model SVM digunakan dan dilatih menggunakan *caption* yang telah dilabelisasi secara manual. Proses NER dilakukan dengan pendekatan *rule-based* untuk mengekstraksi pola linguistik spesifik yang berhubungan dengan lokasi, seperti pencocokan daftar lokasi, preposisi, dan struktur deskriptif. Hasil NER akan menjadi bagian dari ekstraksi fitur SVM dalam mengklasifikasikan teks caption. Tujuan penelitian ini agar model mampu mengidentifikasikan dan mengklasifikasikan teks secara otomatis berdasarkan ada atau tidaknya informasi lokasi yang terkandung dalam caption.

Oleh sebab itu, penelitian ini diharapkan menjadi langkah awal untuk pengembangan NLP pada penelitian platform media sosial berbasis lokasi, dan ikut serta memberikan kontribusi dalam analisis spasial dan pengambilan keputusan berbasis lokasi, terutama dalam konteks perilaku pengguna digital di Indonesia. Diharapkan penelitian ini mampu menjadi langkah awal terhadap pengembangan sistem imformasi yang lebih cerdas, akurat, dan adaptif terhadap dinamika bahasa di ranah dunia maya.

## 2. STUDI PUSTAKA

### *A. Natural Language Processing*

Bidang pembelajaran Kecerdasan Buatan mencakup salah satunya *Natural Language Processing* (NLP), berfokus pada bagaimana komputer dapat memahami dan merespon bahasa manusia. Bertujuan untuk mengubah teks yang dianggap tidak bermakna oleh komputer menjadi representasi numerik yang memiliki nilai dan dapat diproses. Proses NLP dimulai dengan alur *pre-processing* data, dilanjutkan dengan pemberian anotasi untuk menambahkan informasi kebahasaan atau konteks sesuai domain data yang diproses. Selanjutnya, teks akan diformat ke dalam bentuk numerik yang dapat diproses oleh model pembelajaran (Payette et al., 2025).

### *B. Named Entity Recognition*

*Named Entity Recognition* atau sering dikenal dengan Pengenalan Entitas Bernama merupakan topik yang populer dalam bidang ilmu *Natural Language Processing* (NLP) dan *Text Mining*. NER termasuk kedalam tahapan *pre-processing* yang penting bagi tahapan lanjutan seperti peringkasan teks, sistem tanya jawab dan ekstraksi informasi. Dalam NLP, entitas bernama sering didefinisikan sebagai kemunculan entitas dalam teks. Walaupun terdapat beragam definisi, entitas bernama dapat diartikan sebagai kata yang mengidentifikasi satu objek dari kumpulan objek lainnya secara spesifik dengan atribut yang sama (Wang et al., 2022). Pendekatan yang umum digunakan dalam menyelesaikan NER ialah pemodelan *sequence labeling task* (pelebelan urutan), dimana setiap token akan diklasifikasikan berdasarkan jenis entitasnya(Wang et al., 2023).

Pemrosesan entitas bernama sangat bermanfaat dalam bidang *information retrieval* yaitu proses untuk mendapatkan data tertentu dari sekumpulan data berdasarkan *query* (permintaan) yang diberikan. Selain itu, entitas benama sering digunakan dalam studi ektraksi informasi, dimana studi ini mengidentifikasi informasi dalam jumlah besar teks yang tidak terstruktur (Ehrmann et al., 2024). Penelitian Ehrmann mengkaji NER pada dokumen historis yang ditandai dengan struktur bahasa lama, ketidakteraturan ejaan, dan konteks temporal yang kompleks. Tantangan ini sejalan dengan konteks media sosial Instagram, yang menghasilkan teks tidak baku, singkatan, dan gaya bahasa informal. Perbandingan ini menunjukkan pendekatan yang adaptif dalam pengembangan NER dalam lintas domain dan gaya bahasa. Bagian penting dari NER salah satunya yaitu untuk mengidentifikasi entitas seperti nama individu, instansi, organisasi, lokasi ataupun elemen spesifik lainnya dalam suatu data tekstual (Payette et al., 2025).

### *C. Text Mining*

*Text Mining* adalah metode untuk klasifikasi, pengelompokan, ekstraksi, dan pengambilan data dalam bentuk tekstual. Secara umum, *text mining* terdiri dari tiga langkah yaitu: *text pre-processing*, *text mining operation*, *post-processing* (Firdaus & Firdaus, 2021).

### *D. Machine Learning*

*Machine learning* merupakan pengembangan lanjutan dari ilmu data, yang membentuk data menjadi model prediktif. Model ini menghilangkan intervensi manual dari tugas-tugas kompleks, seperti klasifikasi, regresi, pengenalan pola, diagnosis, dan perencanaan. Tahapan *Machine Learning* secara garis besar ada dua tahap, yaitu *training* dan prediksi. Tahapan ini dimulai dari pengumpulan data, *pre-processing* data, serta ekstraksi fitur. Pada tahap *training*, data yang telah diproses menjadi dasar untuk membangun model pembelajaran. Di tahapan ini, fase ekstraksi fitur menjadi elemen penting yang menentukan fitur-fitur pendukung relevan dalam proses analisis, sehingga model bisa mengidentifikasi kesesuaian yang signifikan. Tahapan *training* diakhiri dengan proses evaluasi dengan *test set* yang terpisah untuk menguji performa dan kemampuan model yang akan menghasilkan model final *machine learning* (Tekinerdogan, 2025).

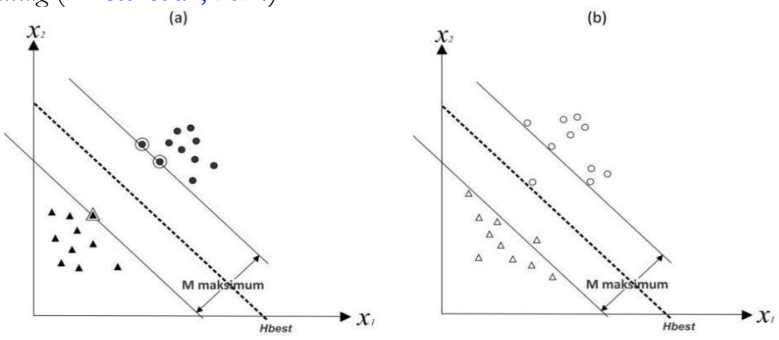
### *E. Pre-processing*

*Pre-processing* adalah tahapan yang berguna untuk merepresentasi data dalam bentuk data baru, meghilangkan noise, mengurangi kosakata dalam teks, melakukan pencocokan kata. *Pre-processing* data merupakan proses penting untuk meningkatkan kinerja model (Anggraeni et al., 2022). *Pre-processing* ini terdiri dari beberapa langkah, yaitu:

1. *Cleaning*
   Proses untuk memperbaiki data yang belum bersih, menyaring data-data yang tidak cocok dari keseluruhan data, dan pengurangan kata tidak penting. *Cleaning* merupakan langkah untuk menghilangkan kumpulan no-abjad, tag, dan tanda baca yang tidak diperlukan.
2. *Case Folding*
   Mengubah atribut teks menjadi huruf kecil dan menghilangkan karakter bukan kata contohnya tanda simbol, penggunaan tanda baca dan angka, hanya menyisakan teks dari abjad "a" sampai dengan "z" (Ma'rifah et al., 2020) dikenal dengan proses *Case Folding*.
3. *Tokenization*
   *Tokenization* merupakan fase untuk pemisahan tipe data *string* teks (kalimat) menjadi bagian token (kata) (Budiman & Widjaja, 2020).
4. *Stopwords Removal*
   Proses ini adalah langkah untuk menghilangkan kata yang tidak penting pada dokumen teks. *Stopword removal* memiliki daftar kata *stopword* yang tidak memiliki makna signifikan, untuk mendapatkan daftar *stopword* ini dibutuhkan pendekatan ilmu tata bahasa (Ferilli, 2021) yang disesuaikan dengan domain studi.

### *F. Support Vectore Machine (SVM)*

Algoritma populer dalam *machine learning* salah satunya adalah *Support Vector Machine*. Memiliki konsep yang sederhana, akan tetapi memiliki karakteristik yang sulit dipahami secara internal (Adek et al., 2021). Selain mampu untuk belajar dari data, SVM merupakan metode yang mampu membuat keputusan (Valkenborg et al., 2023). SVM sebagai metode *machine learning* berbasis teori *statistic*, menawarkan akurasi yang unggul dibandingkan sebagian besar algoritma sejenis dalam berbagai kasus klasifikasi (Nurdin, 2024). SVM efektif dalam membagi kelas menjadi dua kategori. Dalam praktik nyata, banyak penelitian yang melibatkan kasus *multiclass*. Namun, untuk penelitian dengan pembagian klasifikasi sederhana SVM adalah salah satu pendekatan terbaik untuk digunakan (Astrianda, 2020). SVM dapat menangani data yang memiliki banyak fitur-fitur. Alur kerja SVM yaitu dengan menentukan *hyperplane* terbaik, berfungsi untuk menjadi pemisah antara kelas-kelas data dengan margin maksimum dalam ruang fitur, serta performa model yang dijaga agar tidak *overfitting* (Binetti et al., 2024).




**Gambar 1.** *Hyperplane* terbaik dan Margin Maksimum

Dalam SVM, titik-titik data yang terdapat pada *margin* paling dekat dengan *hyperplane* seperti pada Gambar 1 disebut *support vector*. *Support vector* inilah yang digunakan dalam perhitungan SVM untuk mendapatkan nilai *hyperplane* paling bagus, sedangkan informasi lainnya tidak menjadi tolak ukur dalam perhitungan. Berikut persamaannya:

$$f(x) = w.x + b \quad \text{..... (1)}$$

atau

$$f(x) = \sum_{i=1}^{n} y_i a_i K(\mathrm{x}, x_i) + b \quad \text{..... (2)}$$

Keterangan:

w : parameter *hyperplane* yang dicari
x : titik data masukan *SVM*
b : parameter *hyperplane* yang dicari (nilai bias)
$a_i$ : nilai bobot setiap titik yang data
$y_i$ : kelas data ke-i

$x_i$ : data ke-i
$K(x, x_i)$ : fungsi Kernel RBF

### *G. Confusion Metrics*

*Confusion Metrics,* metode evaluasi yang umum digunakan untuk menentukan hasil uji perilaku model klasifikasi. Struktur *confusion metrics* direpresentasikan dalam bentuk matriks 2 x 2. Terdapat 4 matriks utama yaitu (Hasnain et al., 2020):

1. *True positives* (TP): prediksi benar terhadap kelas actual positif.
2. *False positives* (FP): prediksi salah, dimana hasil klasifikasi menganggap data positif, sedangkan kelas actual negatif.
3. *False negatives* (FN): prediksi salah, dimana hasil klasifikasi menganggap data negatif, sedangkan kelas actual positif.
4. *True negatives (*TN): prediksi benar terhadap kelas negatif.

*Confusion matrix* digunakan untuk mengevaluasi kinerja model klasifikasi terhadap dataset. Nilai yang dihasilkan melalui metode *confusion matrix* adalah berupa evaluasi sebagai berikut:

1. Akurasi
   Merupakan presentase jumlah data yang diprediksi secara benar oleh algoritma, persamaan nilai akurasi yaitu:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{.....(3)}$$

2. Presisi
   Nilai ketepatan dari metode yang digunakan dalam klasifikasi. Nilai ini menunjukkan banyaknya data yang dapat terklasifikasi di kelas yang benar dalam pengujian.

$$Presicion = \frac{TP}{(TP+FP)} \quad \text{.....(4)}$$

3. *Recall*
   *Recall* juga dikenal sebagai *Sensitivity*, nilai matriks yang dapat mengukur hasil presentase data prediksi positif yang benar (*true positive*) dibandingkan dengan keseluruhan kasus prediksi sebenarnya, persamaannya (Valero-Carreras et al., 2023):

$$Recall = \frac{TP}{(TP+FN)} \text{.....(5)}$$

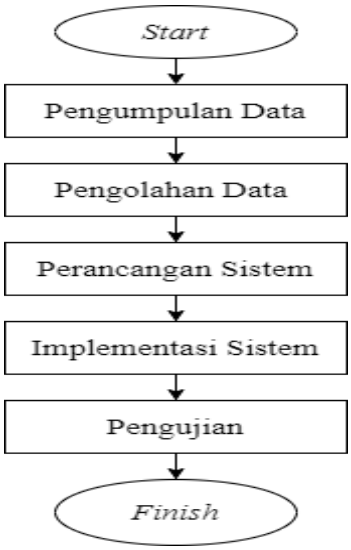
4. F1-Score
   Nilai ini adalah *mean harmonic* (nilai rata-rata harmonik) dari presisi dan *recall*. Persamaanya yaitu (Valero-Carreras et al., 2023):

$$f1 - score = \frac{2 . Precision . Recall}{Precision + Recall} \text{..... (6)}$$

## 3. METODE PENELITIAN

### A. Alur Penelitian

Alur penelitian terdiri dari beberapa tahapan yang dilaksanakan secara berkesinambungan. Adapun tahapan alur penelitian secara garis besar dijelaskan sebagai berikut:




**Gambar 2.** Alur Penelitian

Keterangan Alur Penelitian:

1. Pengumpulan Data
   Penelitian dimulai dengan proses pengumpulan dataset berupa *caption Instagram*. Pengambilan data dilakukan melalui proses *crawling* URL terhadap postingan yang diatur sebagai postingan publik.
2. Pengolahan Data
   Data caption yang dikumpulkan akan melalui tahapan *pre-processing* data, seperti pembersihan teks (*cleaning*), tokenisasi, serta *stopword removal* untuk meningkatkan kualitas teks yang akan dianalisis.
3. Perancangan Sistem
   Setelah data siap digunakan, dilakukan perancangan sistem untuk mempermudah implementasi, termasuk perancangan arsitektur, alur kerja sistem, dan antarmuka pengguna.
4. Implementasi Sistem
   Sistem deteksi entitas lokasi berbasis web dibangun menggunakan bahasa pemograman *Ruby* dan *framework* RoR versi 8. Sistem ini meliputi proses *crawling*, pendeteksian lokasi berbasis NER *rule-based*, pelabelan data, serta klasifikasi algoritma SVM.
5. Pengujian Sistem
   Setelah implementasi selesai, dilakukan proses pengujian sistem untuk mengevaluasi kinerja sistem, baik dari segi fungsionalitas maupun akurasi model. Pengujian menggunakan metriks evaluasi seperti akurasi, presisi, *recall*, dan f1-score.

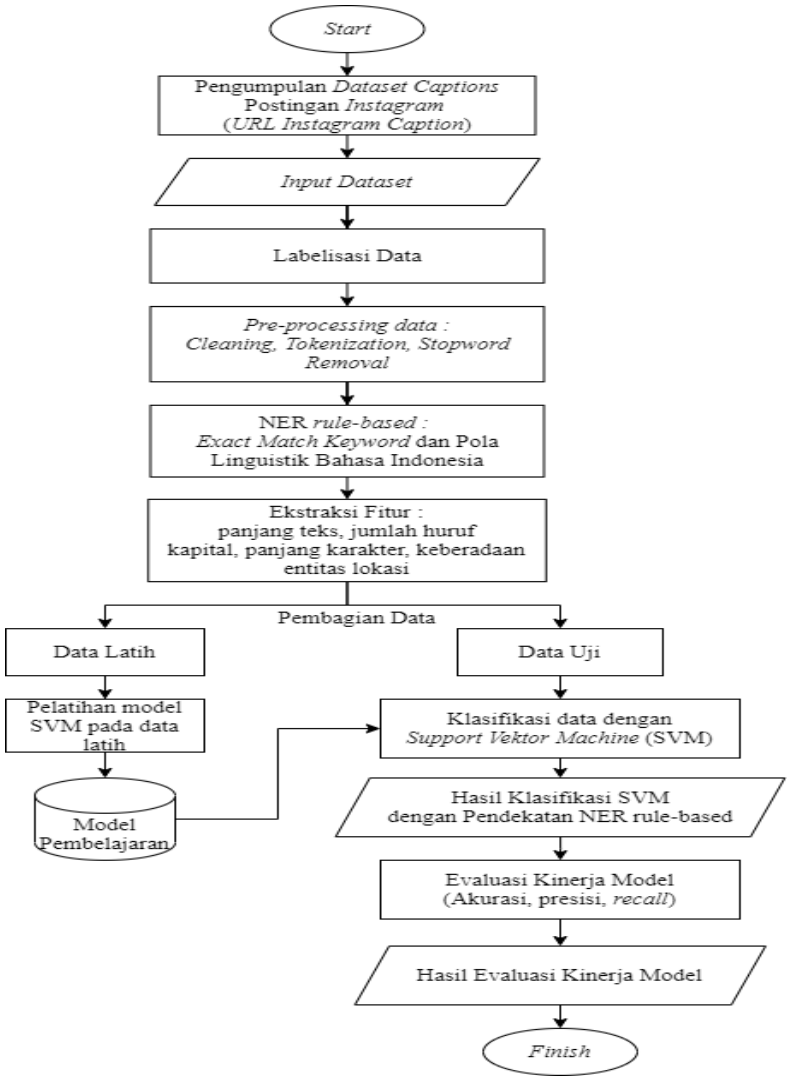
### B. Dataset dan Pengambilan Data

Data penelitian diperoleh melalui metode *crawling caption* Instagram dari postingan publik menggunakan *library (gem) Ruby HTTParty* dan *Nokogiri* untuk parsing HTML. Proses crawling dilakukan secara manual melalui URL postingan publik tanpa melanggar *term of service Instagram*. Sebanyak 400 data caption berhasil dikumpulkan dan dilabelisasi secara manual berdasarkan entitas lokasi dalam teks.

Berbeda dengan penelitian *PromptNER* (Ashok & Lipton, 2023) yang menggunakan dataset publik terstruktur dan evaluasi generalisasi model secara luas dengan lima kali percobaan, penelitian ini bersifat experiment tunggal dengan data tak terstruktur dari media sosial dan fokus pada konteks spasial yang spesifik. Keterbatasan jumlah data disebabkan oleh kebijakan privasi *Instagram*, proses seleksi postingan yang relevan, serta anotasi manual yang memerlukan ketelitian dan waktu. Meski terbatas, jumlah 400 dataset dianggap cukup representatif untuk mengevaluasi efektifitas sistem deteksi entitas lokasi berbasis *rule-based* dan klasifikasi SVM.

### C. Skema Sistem

Berikut rancangan dari skema sistem yang dibangun:




**Gambar 3.** Skema Sistem

Keterangan:

1. *Start*: Tahapan dimulai.
2. Pengumpulan Dataset Captions Postingan Instagram (URL Instagram Caption)
   Langkah dimulai dengan *crawling URL Instagram* untuk mendapatkan data caption
3. *Input Dataset:*
   Data yang diperoleh dari langkah *crawling* URL ditambahkan kedalam database dataset.
4. Labesisasi Data:
   Data caption diberikan label kelas secara manual, yaitu teks dengan entitas lokasi (1) dan yang tidak memiliki entitas lokasi (0).
5. *Prepocessing Data (Cleaning, Tokenization, Filtering (Stopwords Removal)):*
   Dilakukan proses *cleaning* dari karakter-karakter, spasi berlebih, *hashtag* kecuali abjad dan angka serta titik (.), kemudian dilakukan *tokenization* untuk memecah teks menjadi *array* kata serta dilakukan *filtering* kata yang tidak penting (*stopwords removal*).
6. *NER rule-based,* Ekstraksi Fitur:
   Pendekatan NER dilakukan untuk mendeteksi entitas lokasi secara eksplisit berdasarkan pola bahasa Indonesia dan pencocokan dengan *database* daftar lokasi *(keyword_location)*. Sistem belum dapat menangani entitas lokasi implisit secara akurat, seperti frasa deskriptif yang merujuk lokasi tanpa menyebutkan nama langsung seperti "kafe dekat Sabang". Penanganan terhadap lokasi implisit direncanakan sebagai

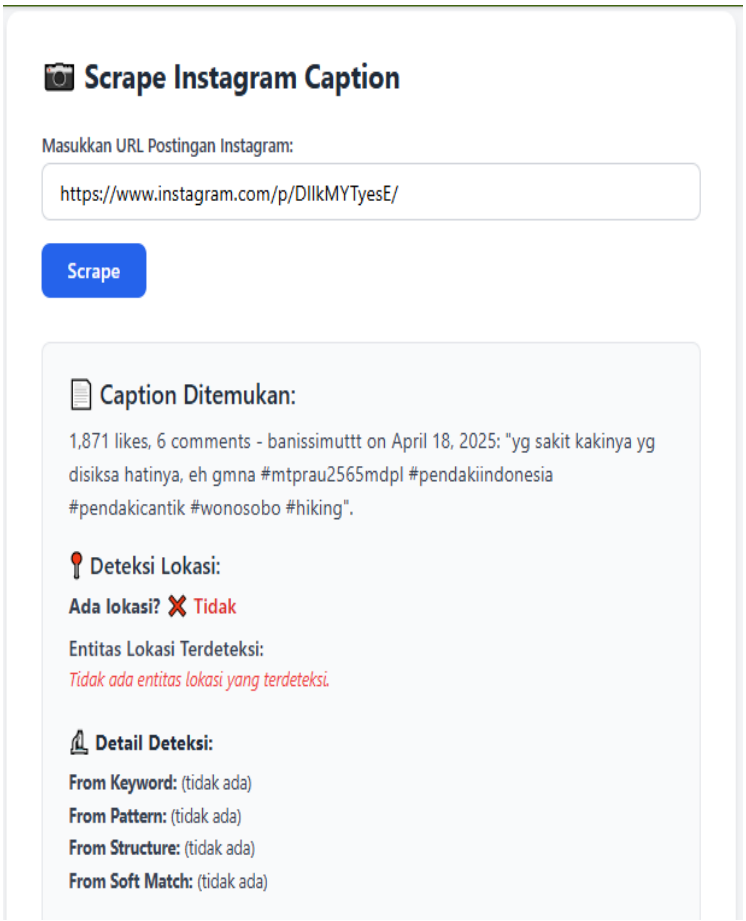
pengembangan lebih lanjut dengan pendekatan berbasis konteks atau model berorientasi kalimat seperti *transformer IndoBERT*. Selain itu, proses ekstraksi fitur diterapkan untuk mengubah informasi tekstual menjadi fitur numerik yang dapat diproses oleh model SVM.

7. Pembagian Data
   Data dibagi dua yaitu 70% pelatihan dan 30 % pengujian
8. Data Latih:
   Data latih akan dilatih dengan model SVM dan disimpan sebagai model pembelajaran.
9. Data Uji:
   Data uji akan dilatih dengan Model Pembelajaran SVM yang telah dilatih.
10. Hasil Klasifikasi:
    Hasil Klasifikasi berupa prediksi teks *Contains_Location* dengan *No_Location* beserta hasil ekstraksi entitas lokasi yang ditemukan dalam teks serta visualisasi hasilnya.
11. Proses Evaluasi Kinerja Model (Akurasi, Presisi, *Recall*, dan *f1-Score*)
    Hasil klasifikasi akan dilakukan evaluasi kinerja model dengan *Confusion Matrix*.
12. Hasil Evaluasi Kinerja Model
    Hasil evaluasi kinerja terdiri dari akurasi, presisi, *recall*, *f1-score*, serta grafik hasil eveluasi.
13. *Finish:* Tahapan selesai.

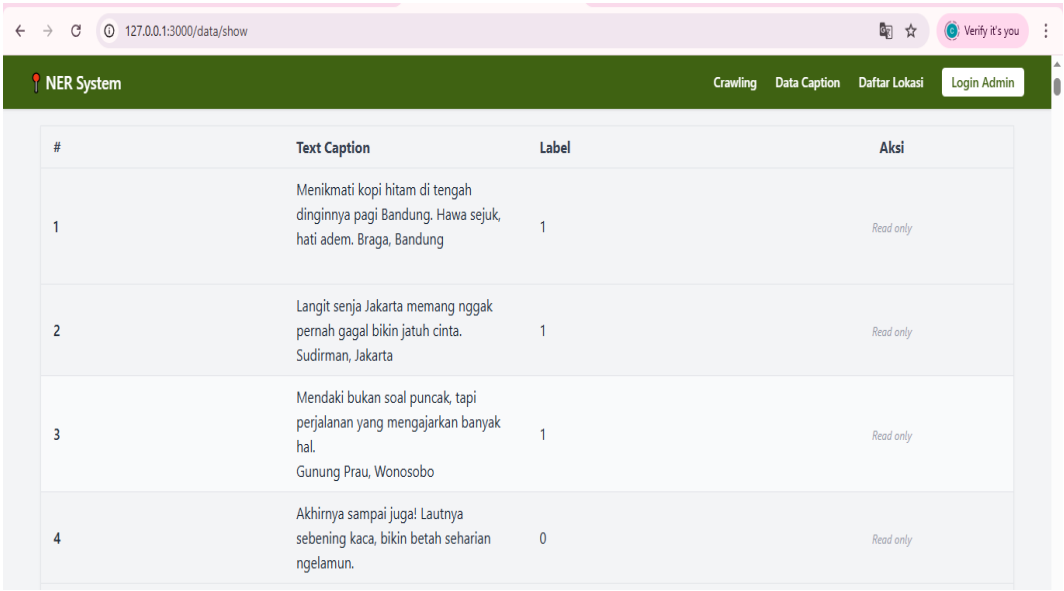
## 4. HASIL DAN PEMBAHASAN

### A. Hasil Implementasi Sistem

Berikut tampilan halaman proses pengambilan data, daftar *caption* dan daftar lokasi:

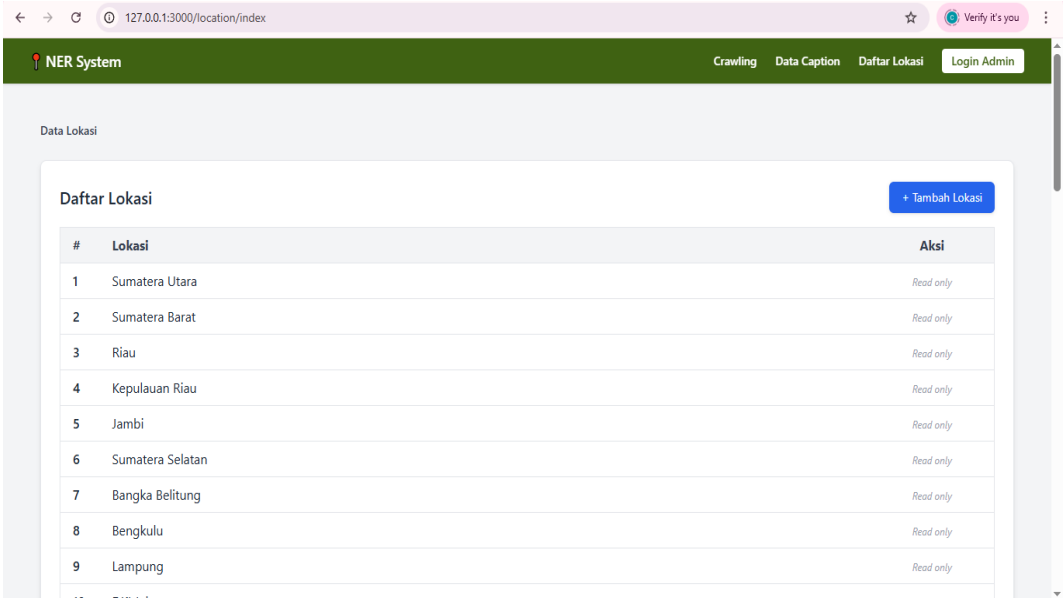



**Gambar 4.** *Interface Crawling* Data




**Gambar 5.** *Interface* Halaman Data *Caption*

Sistem menggunakan daftar lokasi administratif sebagai *keywords* deteksi entitas lokasi, namun daftar ini terbatas karena bersumber dari input manual. Ke depannya, daftar lokasi dapat diintegrasikan dengan *Geographic Information Systems* untuk memperoleh data lokasi yang lebih lengkap dan terstruktur.




**Gambar 6.** *Interface* Halaman Daftar Lokasi

## B. Implementasi Klasifikasi dengan *Support Vector Machine* dan *NER Rule-Based*

Berikut tabel data uji hasil prediksi dari caption Instagram dengan ekstraksi entitas lokasi:

**Tabel 1.** Hasil Klasifikasi SVM

| No | Caption | Clean Caption | Label Awal | Label Prediksi | Keyword | Pattern | All Loc |
|---|---|---|---|---|---|---|---|
| 1 | "AURA vintage amat berasa saat memasuki kafe kekinian yang berselimut seni ini. Aura arsitektur kuno menyapa kenangan di | AURA vintage amat berasa memasuki kafe kekinian berselimut seni Aura arsitektur kuno menyapa kenangan | 0 | No Location | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | zaman rumah kayu. Selengkapnya di www.waspada.id #beritawisataaceh". | di zaman rumah kayu Selengkapnya di wwwwaspadaid | | | | |
| 2 | "Sekarang merupakan waktu paling tepat untuk menikmati sunset di Pantai Serang. Matahari terbenam tepat di laut lepas. …….. https://www.panduaji .net/2019/09/liburan-pantai-serang-blitar.html #blitar #pantaiindonesia | Sekarang waktu paling tepat menikmati sunset di Pantai Serang Matahari terbenam tepat di laut lepas Pemandangan seperti hanya ……. httpswwwpanduajin et201909liburanpant aiserangblitarhtml | 1 | Contains Location | Pantai Serang | Pantai Serang |
| 3 | "Menikmati suasana malam di kafe pinggir pantai, ditemani deburan ombak dan sinar bintang yang menyinari malam. Sempurna #DeburanOmbak #MalamBersama". | Menikmati suasana malam di kafe pinggir pantai ditemani deburan ombak sinar bintang menyinari malam Sempurna | 1 | No Location | | |
| 4 | " 📍 Sudut Tenang di Cafe Tujuh G Kadang kamu nggak perlu banyak gaya, cukup berdiri... dan biarkan alam yang bicara. Temukan tempat terbaik untuk rehat dan berpose alami. #SpotFotoHits #AlamIndonesia". | Sudut Tenang di Cafe Tujuh G Kadang kamu nggak perlu banyak gaya cukup berdiri biarkan alam bicara Temukan tempat terbaik rehat berpose alami | 1 | Contains Location | Cafe Tujuh G | Cafe Tujuh G |
| … | | | | | | |
| … | | | | | | |
| 118 | "Baru saja mendaki Gunung Catur dan mencapai Puncak Mangu! Pemandangan alam yang luar biasa dan petualangan yang seru! #MendakiGunung #GunungIndonesia#. | Baru saja mendaki Gunung Catur mencapai Puncak Mangu Pemandangan alam luar biasa petualangan seru | 1 | Contains Location | Gunung Catur | Gunung Catur |
| 119 | "Pesisir Selatan-Kawasan wisata Pantai Penyu yang terdapat di Kampung Pasar Amping Parak, Nagari Amping Parak, | Pesisir Selatan Kawasan wisata Pantai Penyu terdapat di Kampung Pasar Amping Parak | 1 | Contains Location | Kampung Pasar Amping \| Kabupate | Kampung Pasar Amping \| Kabupate |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Kecamatan Sutera, Kabupaten Pesisir Selatan (Pessel) terus ……Baca selengkapnya di www.pesisirselatankab .co.id Via @pesisirselatan24jam #padangjournal. | Nagari Amping Parak Kecamatan Sutera Kabupaten Pesisir Selatan Pessel terus ……Baca selengkapnya di wwwpesisirselatanka bcoid Via pesisirselatan24jam | | | n Pesisir Selatan | n Pesisir Selatan |
| 120 | "Gunungkidul menyimpan sejuta pesona berupa tempat-tempat dengan pemandangan luar biasa indah dan menakjubkan, salah satunya Pantai Watu Lumbung. …… atau menangkap ikan di celah batu karang. Sumber: Kompas.com #panoramaindonesia #infojogja | Gunungkidul menyimpan sejuta pesona berupa tempattempat pemandangan luar biasa indah menakjubkan salah satunya Pantai Watu Lumbung …… atau menangkap ikan di celah batu karang Sumber Kompascom | 1 | Contains Location | Gunung Batur \| Pantai Watu Lumbung | Gunung Batur \| Pantai Watu Lumbung |

Tahapan implementasi sistem klasifikasi caption menggunakan *SVM*, dengan model pembelajaran yang dilatih pada 70 % caption dan 30% data uji untuk pengujian model. Serta untuk meningkatkan performa klasifikasi dalam pengenalan entitas lokasi, *Named Entity Recognition* (NER) berbasis aturan menjadi bagian dari fitur SVM dalam proses klasifikasi.

### 1. *Pre-processing data*

Langkah awal adalah tahap *cleaning data*. Selanjutnya data dipecah dalam bentuk array (*tokenization*) agar memudahkan data *caption* dibersihkan dari simbol serta karakter yang tidak relevan dalam teks (*stopword removal*). Token yang telah bersih kemudian digabungkan kembali menjadi *Clean Caption*.

### 2. Penerapan NER *rule-based*

*Rule* yang diterapkan ada 3 yaitu:

a) Pencocokan *exact match keywords* lokasi dari database.
b) Pola linguistik preposisi (preposisi + nama lokasi).
c) Pola linguistik struktur deskriptif (deskriptor + nama lokasi).

Untuk merealisasikan *proses rule-based* ini, digunakan beberapa algoritma pencarian dalam mengidentifikasi entitas lokasi dari teks, yaitu:

*a) Regex-based search*
Digunakan untuk mencocokkan pola linguistik yang telah ditentukan, seperti pola preposisi (contoh: "di Gunung Rinjani") dan pola struktur deskriptif (contoh: "Jalan Malioboro" atau "Kabupaten Tegal"). Ekspresi regular ini efektif dalam mendeteksi entitas yang tersusun secara eksplisit dalam struktur kalimat.

*b) Hash-based search*
Diterapkan untuk menyaring kata-kata yang tidak relevan atau *exceptional words*, seperti nama platform media sosial atau kata hubung umum. Hal ini memungkinkan sistem mengabaikan entitas yang mirip lokasi namun bukan lokasi geografis yang valid.

*c) Sequential search*
Digunakan untuk mengevaluasi frasa secara berurutan dalam daftar pendek seperti preposisi atau deskriptor, serta untuk menghindari duplikasi hasil deteksi. Meskipun bersifat linier, pendekatan ini tetap efisien karena digunakan pada data berukuran kecil.

Dari beberapa sample data pada Tabel 1, dapat dilihat sistem mampu mengenali lokasi secara akurat apabila:

a) Lokasi termasuk dalam daftar lokasi pada database.
b) Lokasi terdeteksi dalam konteks yang sesuai dengan aturan linguistik yang ditentukan.
c) Kata yang menyerupai nama tempat, akan tetapi bukan lokasi spasial akan diabaikan oleh sistem sesuai dengan pengecualian yang ditentukan dalam pola.

Namun sistem akan gagal mendeteksi lokasi apabila:

a) Lokasi tidak termasuk dalam daftar lokasi pada database.
b) Struktur kalimat tidak eksplisit, contohnya "Lagi di sini aja, nyantai", gagal dideteksi sebagai entitas lokasi, karena lokasi tidak disebutkan secara explisit dalam frasa tempat atau nama geografis tertentu. Meskipun aktivitasnya mengimplikasikan lokasi, namun tidak ada entitas lokasi yang dapat dikenali oleh sistem berdasarkan pola yang ditentukan.
c) Caption terlalu umum, hanya menyampaikan kesan atau pengalaman tanpa menyebutkan lokasi nyata, contohnya "Dibalik lelahnya perjalanan".

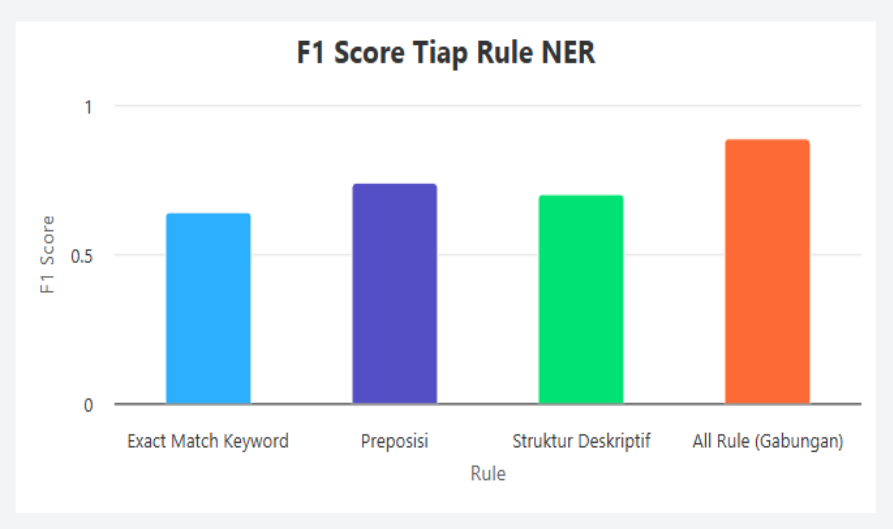
### 3. Evaluasi Kinerja NER *rule-based*

Evaluasi NER *rule-based* dilakukan dengan membandingkan hasil identifikasi lokasi dari sistem dengan label manual menggunakan metriks evaluasi. Dari hasil tersebut dihitung akurasi, presisi, recall dan f1-score untuk setiap rule dan gabungannya.

**Tabel 2.** Evaluasi Kinerja Rules NER Lokasi

| ***Rule*** | **TP** | **FP** | **FN** | **TN** | **Akurasi** | **Presisi** | ***Recall*** | **F1-Score** |
|---|---|---|---|---|---|---|---|---|
| *Exact Match Keywords* | 127 | 19 | 125 | 129 | 64% | 87% | 5% | 64% |
| Preposisi | 147 | 1 | 105 | 147 | 74% | 99% | 58% | 74% |
| Struktur Deskriptif | 136 | 2 | 116 | 146 | 71% | 99% | 54% | 70% |
| *All Rules* (Gabungan) | 217 | 20 | 35 | 128 | 86% | 92% | 86% | 89% |

a) *Rule Exact Match Keywords*, menghasilkan presisi tinggi (87%) namun *recall* rendah (50%). Meskipun pencocokan lokasi akurat, sistem gagal mendeteksi banyak lokasi karena keterbatasan pada daftar lokasi. *Rule* ini sangat bergantung pada kelengkapan basis data lokasi yang menyebabkan *recall* rendah.
b) *Rule* Preposisi, menggunakan pola seperti "di Wonosobo". *Rule* ini mencatat presisi sangat tinggi (99%) dengan *recall* 58%. Rule ini efektif dalam mengenali pola linguistik eksplisit yang mengarah ke lokasi, meskipun banyak entitas terlewatkan karena tidak sesuai dengan pola preposisi.
c) *Rule* Struktur Deskriptif, mendeteksi entitas lokasi berdasarkan gabungan deskriptor dan nama tempat. Dengan presisi tinggi (99%) dan *recall* 54%, rule ini dapat diandalkan dalam mengenali lokasi administratif, namun terbatas pada struktur pengenalan deskriptor tertentu.
d) Gabungan *rule* memberikan performa terbaik, dengan akurasi 86%, presisi 92%, *recall* 86% dan f1-score tertinggi sebesar 89%. Kombinasi *rule* mampu menyeimbangkan kelebihan dan kekurangan masing-masing pendekatan, menghasilkan deteksi lokasi yang lebih valid dan optimal secara keseluruhan.




**Gambar 7.** Grafik F1-Score *Rules* NER

## 4. Klasifikasi SVM

Model SVM yang digunakan adalah Kernel RBF, dilatih dengan label 1 (*Contains_location*) dan 0 (*No_location*). Pada sisi *backend Ruby* diimplemetasikan dengan *gem lib-svm*, menggunakan *train_model* untuk pelatihannya dan *predict* untuk proses klasifikasi. Prediksi dilakukan terhadap 30% caption berdasarkan *timestamps created_at: :desc*. Berikut perhitungan manual SVM:

Persamaan utama yang digunakan:

$f(x) = \sum_{i=1}^{n} y_i a_i K(x, x_i) + b$ ..... (7)

Dengan kernel RBF:

$K(x, x_i) = \exp(-\gamma . |x - x_i|^2)$ ..... (8)

*Support vector*, alpha, y, fitur (X), kernel, *decision function*, dan nilai bias yang akan digunakan untuk perhitungan manual data training (data diperoleh dari pemodelan SVM dalam sistem):

**Tabel 3.** *Support Vector, Alpha, y* dan fitur (X)

| *SV* | *a* | *y* | *X1* | *X2* | *X3* | *X4* | *K(xi,x)* | *fi* | *b* |
|---|---|---|---|---|---|---|---|---|---|
| SV1 (1) | 0,5 | 1 | 0,4835 | 0,7672 | 0,4136 | 1 | 0,1570 | 0,0785 | 0,6999 |
| SV2 (2) | 0,5 | -1 | 0,0455 | 0,0517 | 0,0467 | 0 | 0,9958 | -0,4979 | -1,2981 |
| SV3 (3) | 0,5 | 1 | 0,1736 | 0,0948 | 0,1845 | 1 | 0,3579 | 0,1789 | 0,5000 |

Nilai bias yang digunakan adalah nilai rata-rata bias:

b = (0,6999 + (-1,2981) + 0,5000)/3
b = -0,0327

Berikut 10 *sample* hasil prediksi data uji yang telah dinormalisasikan dalam proses ekstraksi fitur dalam sistem:

**Tabel 4.** Hasil Prediksi Manual

| X Test | K(SV1) | K(SV2) | K(SV3) | fi SV1 | fi SV2 | fi SV3 | DF | Predict |
|---|---|---|---|---|---|---|---|---|
| **1** | 0,1570 | 0,9958 | 0,3579 | 0,0000 | 0,0000 | 0,0000 | -0,0671 | 0 |
| **2** | 0,1417 | 0,9961 | 0,3457 | 0,0000 | 0,0000 | 0,0000 | -0,0671 | 0 |
| **3** | 0,7021 | 0,3161 | 0,9590 | 0,3510 | 0,1581 | 0,4795 | 0,9215 | 1 |
| **4** | 0,4747 | 0,3665 | 0,9819 | 0,2373 | 0,1833 | 0,4909 | 0,8444 | 1 |
| **5** | 0,4864 | 0,3654 | 0,9875 | 0,2432 | 0,1827 | 0,4938 | 0,8525 | 1 |
| **6** | 0,5421 | 0,3607 | 0,9736 | 0,2711 | 0,1803 | 0,4868 | 0,8711 | 1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **7** | 0,6383 | 0,3426 | 0,9887 | 0,3192 | 0,1713 | 0,4944 | 0,9177 | 1 |
| **8** | 0,1596 | 0,9988 | 0,3573 | 0,0000 | 0,0000 | 0,0000 | -0,0671 | 0 |
| **9** | 0,4473 | 0,3677 | 0,9692 | 0,2237 | 0,1839 | 0,4846 | 0,8250 | 1 |
| **10** | 0,4767 | 0,3667 | 0,9799 | 0,2383 | 0,1833 | 0,4900 | 0,8445 | 1 |

Contoh perhitungan manual dari satu data uji dengan dataset dan fitur yang telah dinormalisasikan dalam sistem:

**Tabel 5.** X Test Perhitungan Matematis

| X Test | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| **4** | 0,0892 | 0,0752 | 0,08067 | 1 |

Langkah awal menghitung jarak $||x_i\text{-}x||^2$:

Jarak ke SV1:

$||x_i\text{-}x||^2 = (0{,}4835 - 0{,}0892)^2 + (0{,}7672 - 0{,}07526)^2 + (0{,}4136 - 0{,}0806)^2 + (1 - 1)^2$

$= 0,\ 1554 + 0{,}4786 + 0{,}1108 = 0{,}7451$

Jarak ke SV2:

$||x_i\text{-}x||^2 = (0{,}0455 - 0{,}0892)^2 + (0{,}0517 - 0{,}07526)^2 + (0{,}0467 - 0{,}0806)^2 + (0 - 1)^2$

$= 0,\ 0019 + 0{,}0005 + 0{,}0011 + 1 = 1{,}0036$

Jarak ke SV3:

$||x_i\text{-}x||^2 = (0{,}1736 - 0{,}0892)^2 + (0{,}0948 - 0{,}07526)^2 + (0{,}1845 - 0{,}0806)^2 + (1 - 1)^2$

$= 0,\ 0071 + 0{,}0003 + 0{,}0107 = 0{,}0183$

Selanjutnya menghitung kernel RBF dengan *gamma* = 0,5 yang ditentukan secara manual dengan persamaan $\exp(-\gamma \,.\, |x - x_i|^2)$:

$K_{SV1} = \exp(-0{,}5 \,.\, |0{,}7451|^2$

$= \exp(-0{,}3726) = 0{,}6888$

$K_{SV2} = \exp(-0{,}5 \,.\, |1{,}0036|^2$

$= \exp(-0{,}5018) = 0{,}6056$

$K_{SV3} = \exp(-0{,}5 \,.\, |0{,}0183|^2$

$= \exp(-0{,}0091) = 0{,}9909$

Lalu menghitung nilai kontribusi tiap SV:

$fi = y_i a_i K(\mathrm{x}, x_i)$

SV1 = 1 x 0,5 x 0,6888

= 0, 3444

SV2 = -1 x 0,5 x 0,6056

= -0, 3028

SV3 = 1 x 0,5 x 0,9909

= 0, 4954

Jumlah Nilai Kontribusi:

$\Sigma fi = f_1 + f_2 + f_3$

$= 0,\ 3444 + (-0,\ 3028) + 0,\ 4954$

$= 0{,}5370$

Berikutnya menghitung nilai *decision function*:

$f(x) = \Sigma fi + b$

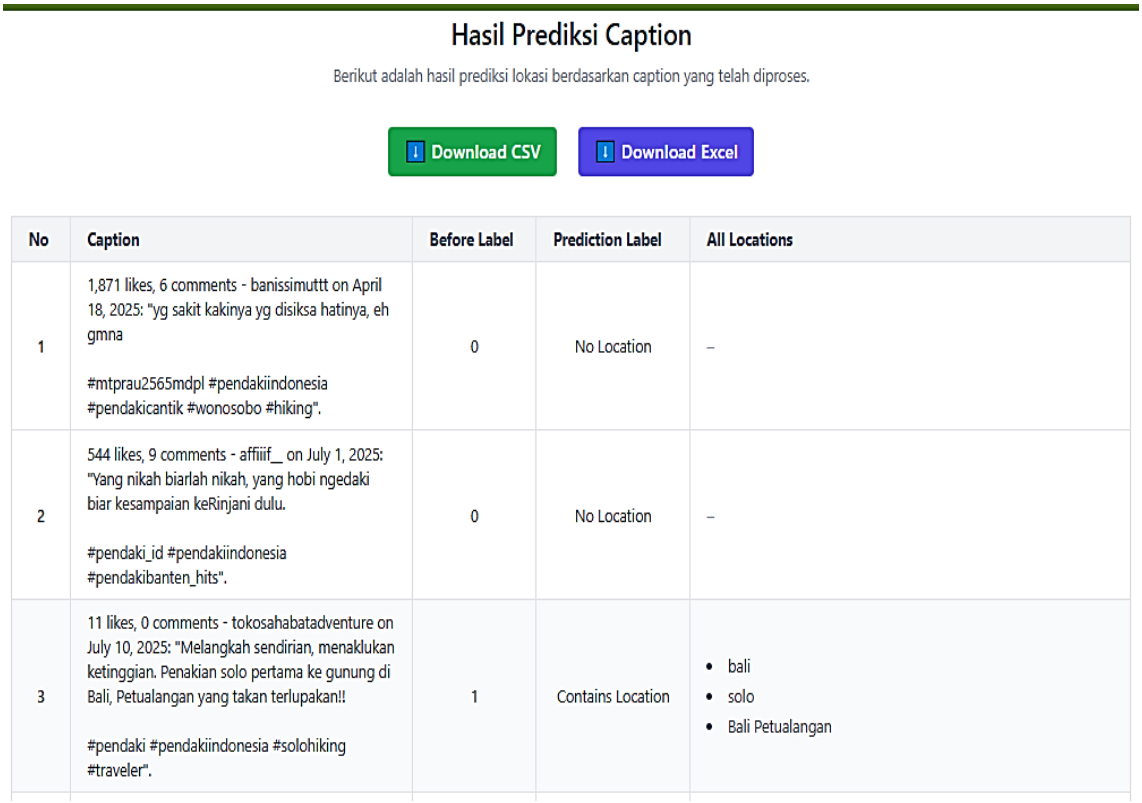
$= 0{,}5370 + (-0{,}0327) = 0{,}5043$

$f(x) > 0$

0,5043 > 0, maka hasil prediksi = 1 (*Contains_location*).

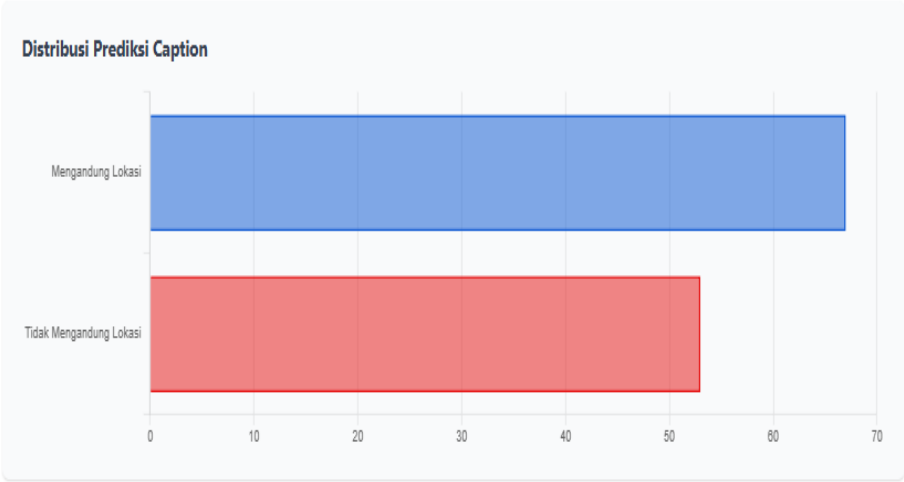
Jika $f(x) < 0$, maka diklasifikasikan sebagai 0 (*No_location*).

Hasil perhitungan manual ini menandakan bahwa implementasi model SVM telah bekerja dengan baik dan efektif dalam memberikan keputusan dan telah sesuai dengan sistem yang berjalan.




| No | Caption | Before Label | Prediction Label | All Locations |
|---|---|---|---|---|
| 1 | 1,871 likes, 6 comments - banissimuttt on April 18, 2025: "yg sakit kakinya yg disiksa hatinya, eh gmna<br><br>#mtprau2565mdpl #pendakiindonesia #pendakicantik #wonosobo #hiking". | 0 | No Location | – |
| 2 | 544 likes, 9 comments - affiiif_ on July 1, 2025: "Yang nikah biarlah nikah, yang hobi ngedaki biar kesampaian keRinjani dulu.<br><br>#pendaki_id #pendakiindonesia #pendakibanten_hits". | 0 | No Location | – |
| 3 | 11 likes, 0 comments - tokosahabatadventure on July 10, 2025: "Melangkah sendirian, menaklukan ketinggian. Penakian solo pertama ke gunung di Bali, Petualangan yang takan terlupakan!!<br><br>#pendaki #pendakiindonesia #solohiking #traveler". | 1 | Contains Location | • bali<br>• solo<br>• Bali Petualangan |

**Gambar 8.** *Interface* Hasil Prediksi SVM




**Gambar 9.** Grafik Distribusi Prediksi Klasifikasi

## C. Evaluasi Kinerja Model SVM

**Tabel 6.** Hasil Prediksi Klasifikasi SVM

| Confusion Matrics | Prediksi: Contain_Location | Prediksi: No_Location |
|---|---|---|
| Actual: 1 | 65 | 9 |
| Actual: 0 | 2 | 44 |

Tabel diatas menunjukkan nilai hasil klasifikasi berdarkan prediksi, adapun penjelasannya sebagai berikut:

1. TP (*True Positive*) berjumlah 65, jumlah data label actual 1 (*Contains_location*) yang berhasil diprediksi dengan benar oleh model sebagai *Contains_location*.
2. FN (*False Negative*) berjumlah 9, jumlah data lebel actual 1 (*Contains_location*) yang salah diprediksi oleh model sebagai *No_location*.

3. FP (*False Positive*) berjumlah 2, jumlah data label actual 0 (*No_location*) yang salah diprediksi oleh model sebagai *Contains_location*.
4. TN (*False Negative*) berjumlah 44, data label actual 0 (*No_location*) yang berhasil diprediksi dengan benar oleh model sebagai *No_location*.

FN sebanyak 9 menunjukkan model gagal dalam mengenali entitas lokasi yang sebenarnya ada, menandakan model kesulitan dalam menangkap lokasi yang tersirat (implisit). FP sebanyak 2 menunjukkan model salah mendeteksi lokasi pada data yang tidak dimilikinya. Hal ini menunjukkan model masih kesulitan membedakan ekspresi lokasi implisit, terutama yang tidak sesuai pola yang ditentukan. Untuk mengurangi kesalahan ini, pendekatan NER perlu diperkuat dengan fitur yang lebih kontekstual dan representatif yang lebih luas terhadap penyebutan lokasi.

Distribusi nilai ini menjadi landasan perhitungan matriks evaluasi akurasi, presisi, *recall*, dan f1-score:

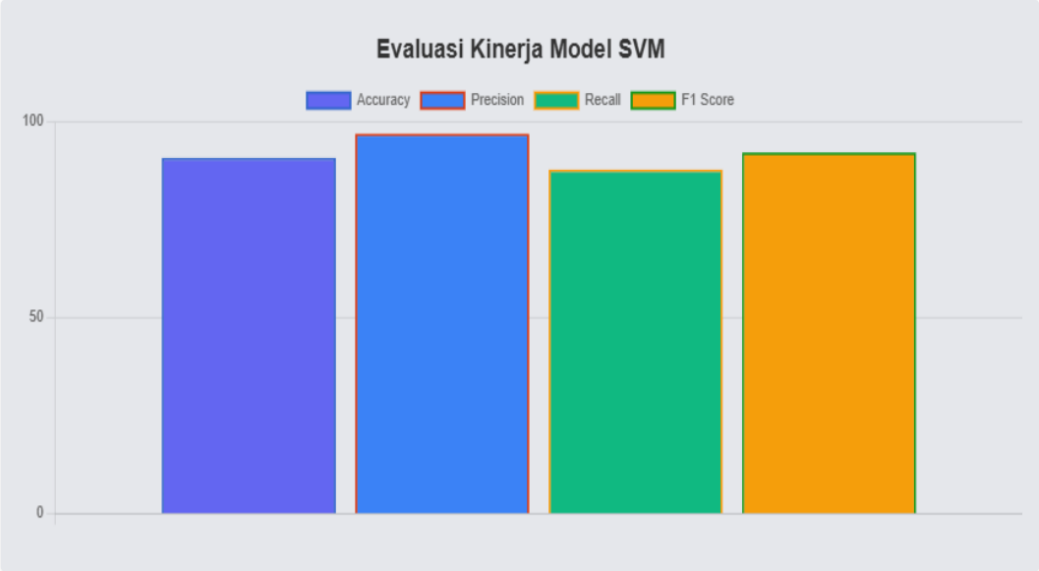
$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \text{ x } 100 = \frac{(65+44)}{(65+44+2+9)} \text{ x } 100 = 90{,}83\%$$

$$Recall = \frac{TP}{(TP+FP)} \text{ x } 100 = \frac{65}{(65+9)} \text{ x } 100 = 87{,}84\%$$

$$Presicion = \frac{TP}{(TP+FP)} \text{ x } 100 = \frac{65}{(65+2)} \text{ x } 100 = 97{,}01\%$$

$$F1-score = \frac{2\,.\,precision\,.\,recall}{precision+recall} \text{ x } 100 = \frac{2\,.\,97{,}01\,.\,87{,}84}{97{,}01+87{,}84} \text{ x } 100 = 92{,}20\%$$

Berikut ini adalah grafik dari hasil evaluasi kinerja model SVM dan pendekatan NER *rule-based*:




**Gambar 10.** Grafik Evaluasi Model Klasifikasi

## D. SIMPULAN

Penelitian ini menunjukkan bahwa klasifikasi teks postingan Instagram menggunakan algoritma SVM dan pendekatan NER *rule-based* dapat mengidentifikasi dan mengklasifikasikan postingan teks Instagram dengan efektif ke dalam dua kelas, yaitu *Contains_location* yang memiliki entitas lokasi, dan *No_location* yang tidak memiliki entitas lokasi. Evaluasi hasil kinerja model menunjukkan capaian yang cukup baik dengan nilai relatif tinggi, yaitu akurasi sebesar 90,83%, presisi 97,01%, *recall* 87,84% serta f1-score 92,20% dari 120 data testing. Hal ini menunjukkan bahwa pendekatan NER *rule-based* dapat meningkatkan akurasi klasifikasi teks berbasis lokasi di media sosial dengan evaluasi kinerja dari gabungan *rules* mendapatkan nilai f1-score

sebesar 89%. Namun, model pembelajaran ini masih memiliki keterbatasan dalam mengidentifikasi lokasi implisit, yaitu ketika entitas lokasi ditulis dalm bentuk yang tidak baku, menggunakan bahasa informal, atau bercampur dengan bahasa asing atau dialek yang tidak bisa teridentifikasi dengan pola yang ditentukan. Selain itu, pendekatan *rule-based* sangat terikat dengan kelengkapan daftar lokasi dan pengenalan pola yang diberikan pada sistem, sehingga masih memiliki kemungkinan error dalam mengidentifikasi ragam tulisan tertentu.

Pada pengembangan selanjutnya, penelitian ini dapat dikembangkan dengan memperkuat fitur kebahasaan yang digunakan, menggunakan pembelajaran berbasis korpus yang dapat menambah ruang lingkup daftar entitas lokasi serta mengkaji lebih lanjut model klasifikasi teks berbasis *deep learning* yang lebih mudah beradaptasi terhadap ragam bahasa informal di media sosial. Hasil penelitian ini diharapkan dapat menjadi rujukan awal terhadap pengembangan sisten analisa spasial berbasis teks yang efisien dan responsif serta dapat diterapkan dalam berbagai aplikasi seperti sistem pengambilan keputusan dan sistem rekomendasi berbasis lokasi, hingga pemetaan geografis pengguna media sosial.

## DAFTAR PUSTAKA


Adek, R. T., Fikry, M., & Khalil, U. (2021). News Opinion Classification Application with Support Vector Machine algorithm using framework Codeigniter. *Journal Of Informatics And Telecommunication Engineering*, *5*(1), 160–166. https://doi.org/10.31289/jite.v5i1.5189

Anggraeni, S. R., Ranggianto, N. A., Ghozali, I., Fatichah, C., & Purwitasari, D. (2022). Deep Learning approaches for multi-label incidents classification from Twitter textual information. *Journal of Information Systems Engineering and Business Intelligence*, *8*(1), 31–41. https://doi.org/10.20473/jisebi.8.1.31-41

Ashok, D., & Lipton, Z. C. (2023). *PromptNER: Prompting for Named Entity Recognition*. http://arxiv.org/abs/2305.15444

Astrianda, N. (2020). Klasifikasi kematangan buah tomat dengan variasi model warna menggunakan Support Vector Machine. *VOCATECH: Vocational Education and Technology Journal*, *1*(2), 45–52. https://doi.org/10.38038/vocatech.v1i2.27

Binetti, M. S., Massarelli, C., & Uricchio, V. F. (2024). Machine Learning in Geosciences: a review of complex environmental monitoring applications. *Machine Learning and Knowledge Extraction, 6*(2), 1263–1280. https://doi.org/10.3390/make6020059

Budi, I., & Suryono, R. R. (2023). Application of Named Entity Recognition method for Indonesian datasets: a review. *Bulletin of Electrical Engineering and Informatics 12*(2), 969–978. https://doi.org/10.11591/eei.v12i2.4529

Budiman, A. E., & Widjaja, A. (2020). Analisis pengaruh teks Preprocessing terhadap deteksi plagiarisme pada dokumen tugas akhir. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(3). https://doi.org/10.28932/jutisi.v6i3.2892

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2024). Named Entity Recognition and Classification in historical documents: a survey. *ACM Computing Surveys*, *56*(2). https://doi.org/10.1145/3604931

Ferilli, S. (2021). Automatic multilingual stopwords identification from very small corpora. *Electronics (Switzerland)*, *10*(17). https://doi.org/10.3390/electronics10172169

Firdaus, A., & Firdaus, W. I. (2021). Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan). *Jurnal Jupiter, 13*(1).

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix measures for Web Services ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Ma'rifah, H., Prasetya Wibawa, A., & Akbar, M. I. (2020). Klasifikasi artikel ilmiah dengan berbagai skenario Preprocessing. *SAKTI: Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, *2*(2), 70–78.

Novian, D., A, H., & Sudirman, R. (2024). Analisis penggunaan teknologi AI ChatGPT terhadap kualitas tugas siswa kelas x di SMA Negeri 1 Gorontalo. *VOCATECH: Vocational Education and Technology Journal*, *6*(1), 62-70. https://doi.org/10.38038/vocatech.v6i1.178

Nurdin, N. (2024). Klasifikasi penerima bantuan dari kepemilikan kartu pelaku utama sektor kelautan dan perikanan dengan metode Support Vector Machine. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3). https://doi.org/10.23960/jitet.v12i3.4507

Payette, M., Abdul-Nour, G., Meango, T. J.-M., Diago, M., & Côté, A. (2025). Leveraging Failure Modes and Effect Analysis for Technical Language Processing. *Machine Learning and Knowledge Extraction*, *7*(2), 42. https://doi.org/10.3390/make7020042

Putra, A. A., Kurniawan, R., & Statistika STIS, P. (2021). Bidirectional LSTM-CNNs untuk ekstraksi entity lokasi kebakaran pada berita online berbahasa Indonesia (Bidirectional LSTM-CNNs for entity extraction of fire location in Indonesian online news) studi kasus di provinsi DKI jakarta (case study in DKI Jakarta province). *Seminar Nasional Official Statistics*. https://doi.org/https://doi.org/10.34123/semnasoffstat.v2020i1.601

Safrizal, S. (2019). Pengenalan Karakter Jawi Tulisan Tangan Menggunakan Fitur Sudut. *VOCATECH: Vocational Education and Technology Journal*, *1*(1), 1-14. https://doi.org/10.38038/vocatech.v1i0.1

Santoso, J., Setiawan, E. I., Yuniarno, E. M., Hariadi, M., & Purnomo, M. H. (2020). Hybrid Conditional Random Fields and K-Means for Named Entity Recognition on Indonesian news documents. *International Journal of Intelligent Engineering and Systems*, *13*(3), 233–245. https://doi.org/10.22266/IJIES2020.0630.22

Tekinerdogan, B. (2025). Machine Learning product line engineering: a systematic reuse framework. *Machine Learning and Knowledge Extraction*, *7*(3), 58. https://doi.org/10.3390/make7030058

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the Confusion Matrix. *Computers and Operations Research*, *152*. https://doi.org/10.1016/j.cor.2022.106131

Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Support Vector Machines. *American Journal of Orthodontics and Dentofacial Orthopedics 164*(5), 754–757). https://doi.org/10.1016/j.ajodo.2023.08.003

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). *GPT-NER: Named Entity Recognition via Large Language Models*. http://arxiv.org/abs/2304.10428

Wang, Y., Tong, H., Zhu, Z., & Li, Y. (2022). Nested Named Entity Recognition: a survey. *ACM Transactions on Knowledge Discovery from Data*, *16*(6). https://doi.org/10.1145/3522593