Analisis Active learning SVM berbasis Margin Sampling pada Sentimen YouTube MBG

Alvian Putra Hardiadi[1], Minarwati*[2]
[1,2] Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer El Rahma Yogyakarta
*Correspondence author email: * minarwati@stmikelrahma.ac.id*

Abstrak
Program Makan Bergizi Gratis (MBG) memicu diskusi publik yang luas di YouTube, menjadikan analisis sentimen sebagai instrumen penting untuk memahami persepsi masyarakat. Penelitian ini menerapkan Active Learning (AL) berbasis Support Vector Machine (SVM) dengan strategi Margin Sampling dan pendekatan Human-in-the-Loop (HITL), di mana peneliti bertindak langsung sebagai oracle dengan panduan anotasi tiga kelas: Negatif, Netral, dan Positif. Dataset terdiri atas 999 komentar YouTube berlabel, dibagi dengan rasio 60/20/20 menjadi 599 sampel latih awal, 200 sampel oracle pool, dan 200 sampel uji tetap. Representasi teks menggunakan TF-IDF dengan unigram dan bigram (max_features=5.000). Tiga kondisi dibandingkan: AL-HITL, Simulated Active Learning, dan Random Sampling. Setelah 15 iterasi dengan 50 sampel berlabel tambahan, AL-HITL mencapai F1-score makro 0,5389, melampaui Simulated AL (0,4950) dan Random Sampling (0,4977). Mekanisme skip dengan skip rate 33,3% berperan sebagai quality filter yang mencegah negative learning, berbeda dari dua kondisi lainnya yang justru mengalami penurunan dari baseline (0,5154). Kelas Positif sebagai minoritas (20%) memperoleh F1 terendah (0,4286) akibat heterogenitas semantik, sementara kelas Negatif sebagai mayoritas (40,7%) memperoleh F1 tertinggi (0,6098). Penelitian ini diposisikan sebagai studi pendahuluan yang memberikan bukti empiris awal mengenai potensi dan tantangan AL berbasis Margin Sampling untuk analisis sentimen komentar YouTube berbahasa Indonesia.
Kata kunci *— Active Learning, SVM, Sentiment Analysis, YouTube, MBG*

1. PENDAHULUAN

Implementasi kebijakan publik berskala nasional, seperti Program Makan Bergizi Gratis (MBG) yang diluncurkan oleh Presiden Prabowo Subianto, telah menjadi pusat perhatian dan memicu diskusi masif di berbagai platform media sosial, termasuk YouTube [1]. Fenomena ini menghasilkan akumulasi data teks berupa komentar masyarakat yang sangat melimpah, yang mencerminkan berbagai persepsi publik, mulai dari dukungan terhadap pemenuhan gizi anak hingga kritik terhadap tantangan implementasi di lapangan [2] Secara teknis, data tidak terstruktur ini merupakan aset berharga untuk analisis sentimen guna memahami efektivitas kebijakan [3]. Namun, melimpahnya data mentah tersebut menimbulkan tantangan besar dalam pemrosesan informasi, terutama karena ketergantungan model pembelajaran mesin tersupervisi (*supervised learning*) tradisional terhadap ketersediaan dataset berlabel dalam jumlah besar untuk mencapai performa klasifikasi yang optimal [4].

eterbatasan utama sistem pembelajaran *tersupervisi* konvensional terletak pada tingginya kebutuhan sumber daya, waktu, dan biaya untuk proses anotasi manual oleh manusia *(oracle)*. Metode pelabelan pasif yang mengandalkan pengambilan sampel secara acak *(random sampling)* sering kali tidak efisien karena cenderung memilih data *redundan* yang memberikan kontribusi informasi rendah terhadap pembentukan batas keputusan (*decision boundary*) model [5]. Kondisi ini menjadi hambatan serius dalam analisis opini publik yang dinamis seperti komentar YouTube, di mana model dituntut untuk mencapai performa yang baik meskipun hanya memiliki akses terhadap jumlah data berlabel yang sangat terbatas [6].

Sebagai solusi atas permasalahan tersebut, *Active Learning* (AL) menawarkan pendekatan yang lebih efisien dengan melibatkan model secara aktif dalam memilih sampel data yang paling informatif untuk dilabeli [7]. Berbeda dengan pelabelan pasif, AL menggunakan

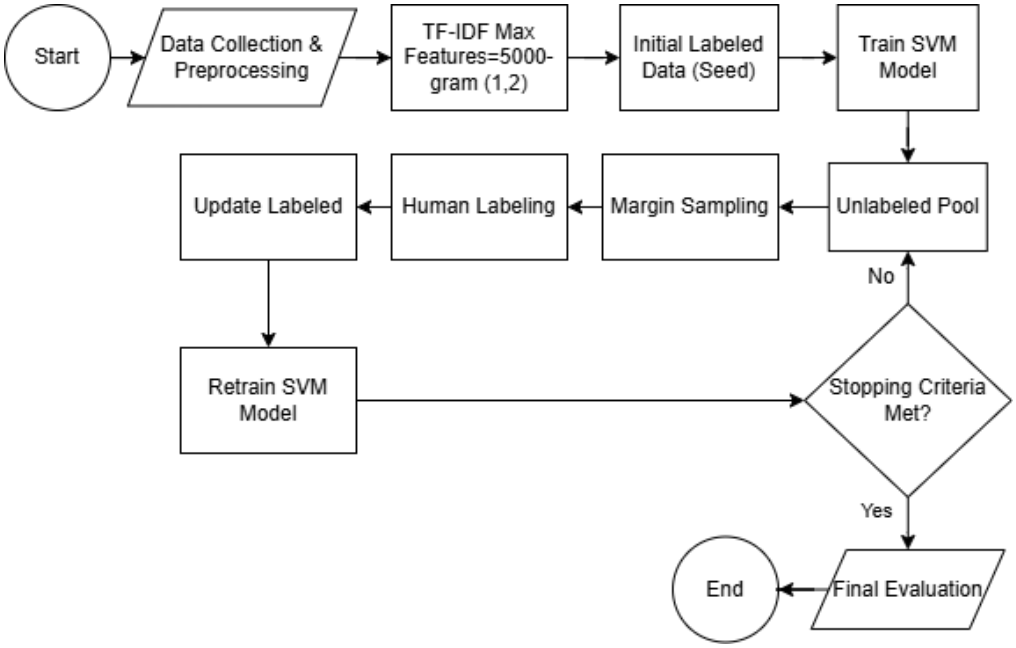
strategi query untuk mendeteksi sampel yang memiliki nilai informasi tinggi bagi model. Di antara berbagai strategi yang ada, *Margin Sampling* salah satu varian *Uncertainty Sampling* sangat cocok diterapkan pada model *Support Vector Machine* (SVM). Strategi ini memprioritaskan sampel yang berada di sekitar *margin hyperplane*, yaitu data dengan tingkat ketidakpastian tertinggi (selisih probabilitas antar kelas paling kecil) [4] [6]. Pemilihan sampel di area margin diharapkan dapat memperbaiki batas keputusan model secara lebih efektif dengan jumlah data berlabel yang minimal.

Terdapat *research gap* yang signifikan dalam literatur analisis sentimen kebijakan publik di Indonesia. Sebagian besar penelitian sebelumnya mengenai Program Makan Bergizi Gratis (MBG) dilakukan pada platform X (Twitter) dengan pendekatan pelabelan pasif menggunakan SVM dan TF-IDF [8] Pendekatan tersebut belum banyak mengeksplorasi potensi *Active Learning* untuk mengatasi keterbatasan data berlabel. Selain itu, penelitian pada komentar YouTube masih sangat terbatas, padahal platform ini memiliki karakteristik yang lebih menantang, seperti bahasa informal, sarkasme, emoji, dan tingkat ambiguitas yang lebih tinggi dibandingkan tweet. Kondisi ini membuka peluang untuk mengeksplorasi efektivitas strategi *Margin Sampling* pada konteks data yang lebih realistis dan kompleks.

Penelitian ini bertujuan mengatasi hambatan utama dalam analisis sentimen berskala besar, yaitu tingginya ketergantungan model tersupervisi terhadap data berlabel, dengan mengimplementasikan kerangka kerja *Active learning*(AL) SVM berbasis *Margin Sampling* pada komentar YouTube mengenai Program Makan Bergizi Gratis (MBG). Melalui pendekatan *eksperimental komputasional*, studi ini mengevaluasi efektivitas strategi *Margin Sampling* dalam mengidentifikasi sampel paling informatif guna mengurangi beban anotasi manual dibandingkan dengan *Random Sampling*, sekaligus menguji ketahanan representasi fitur TF-IDF terhadap kompleksitas bahasa informal pada komentar YouTube. Meskipun peningkatan performa yang diperoleh bersifat *modest*, penelitian ini menunjukkan potensi *Active Learning* dengan *Human-in-the-Loop* (HITL) dalam kondisi sumber daya anotasi yang terbatas. Penelitian ini berkontribusi sebagai studi pendahuluan yang menerapkan teknik *Active Learning* pada diskursus kebijakan nasional di platform YouTube, yang menyajikan tantangan lebih tinggi dibandingkan platform mikroblog seperti X (Twitter) akibat bahasa yang lebih naratif dan ambigu. Secara praktis, hasil penelitian ini diharapkan dapat menjadi referensi awal bagi instansi pemerintah dalam membangun sistem pemantauan opini publik yang lebih efisien terhadap kebijakan strategis [2].

2. METODE PENELITIAN

Analisis sentimen terhadap kebijakan publik seperti Program Makan Bergizi Gratis (MBG) pada platform YouTube menghadapi tantangan utama berupa ketersediaan data teks dalam jumlah besar namun tidak terstruktur. Meskipun data tersebut melimpah, sebagian besar tidak memiliki label sehingga tidak dapat langsung dimanfaatkan oleh model pembelajaran mesin [9]. Permasalahan utama terletak pada ketergantungan model *supervised learning* terhadap data berlabel dalam jumlah besar untuk mencapai performa yang optimal. Proses anotasi manual oleh pakar manusia(*oracle*) menjadi kendala karena membutuhkan waktu, biaya, dan sumber daya yang signifikan, kondisi ini menyebabkan terbatasnya jumlah data latih yang dapat digunakan dalam praktik [10]. Pendekatan konvensional seperti *random sampling* dalam proses pelabelan sering kali tidak efisien, karena data yang dipilih belum tentu memberikan informasi baru yang signifikan bagi model. Akibatnya, model dapat mengalami stagnasi performa meskipun jumlah data latih terus bertambah. Untuk mengatasi permasalahan tersebut, diperlukan pendekatan yang mampu memilih sampel data secara selektif dan informatif. Dalam konteks ini, *Active learning* menawarkan solusi dengan memprioritaskan pelabelan pada data yang paling membingungkan bagi model, khususnya data yang berada di sekitar batas keputusan (*decision boundary*) [11]. Dengan demikian, diharapkan model dapat mencapai performa yang lebih baik dengan jumlah data berlabel yang lebih sedikit [12].



Gambar 1. Tahapan Eksperimen Komputasional

Berdasarkan Gambar 1, penelitian ini mengimplementasikan pendekatan *pool-based active learning* dalam tahapan yang sistematis dan iteratif. Proses dimulai dari pengumpulan data, *preprocessing*, representasi fitur, hingga evaluasi model akhir [13]. Proses diawali pada tahap Start, kemudian dilanjutkan dengan *Data Collection & Preprocessing*. Data komentar YouTube terkait Program Makan Bergizi Gratis (MBG) dibersihkan melalui penghapusan URL, simbol, mention, hashtag, emoji, tanda baca, angka, serta normalisasi huruf kecil. Tahap ini bertujuan untuk mengurangi noise dan meningkatkan kualitas data sebelum masuk ke tahap pemodelan [14].

Selanjutnya, data teks yang telah bersih direpresentasikan ke dalam bentuk numerik melalui proses TF-IDF *Vectorization* dengan konfigurasi maksimum fitur sebanyak 5.000 dan penggunaan *n-gram* (1,2) [8] Secara matematis, TF-IDF didefinisikan pada persamaan 1,2 dan 3.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (1)$$

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

$$TF\ IDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

Keterangan:

- $f_{t,d}$= jumlah kemunculan term t pada dokumen d
- N= total dokumen
- $df(t)$= jumlah dokumen yang mengandung term t

Dalam penelitian ini, TF-IDF menggunakan konfigurasi:

- *max_features* = 5000
- *n-gram* = (1,2)

Representasi ini digunakan untuk menangkap informasi kata tunggal *(unigram)* dan frasa *(bigram)* agar konteks sentimen lebih kaya.

Dari hasil ekstraksi fitur, sebagian kecil data dipilih sebagai *Initial Labeled Data (Seed)* yang digunakan untuk melatih model awal [15]. Model yang digunakan dalam penelitian ini adalah *Support Vector Machine* (SVM), yang dilatih pada tahap *Train* SVM Model untuk membentuk batas keputusan awal (*decision boundary*) [14]. Fungsi keputusan SVM dinyatakan pada persamaan nomer 4 dan optimasinya pada nomer 5.

$$f(x) = w^T x + b \tag{4}$$

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \qquad \text{subject to} \quad y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ = 1, \ldots, \tag{5}$$

di mana Cadalah parameter regularisasi (dalam eksperimen $C = 1.0$) dan $\xi_i$adalah slack variable.

Setelah model awal terbentuk, proses berpindah ke tahap *Unlabeled Pool*, yaitu kumpulan data yang belum memiliki label. Pada tahap ini, dilakukan strategi pemilihan data menggunakan Margin Sampling, di mana model memilih sampel dengan tingkat ketidakpastian tertinggi (selisih probabilitas antar kelas paling kecil). Sampel ini dianggap paling informatif karena berada di sekitar batas keputusan model [15]. Ditujukan pada persamaan nomer 6.

$$\phi_{MS}(x) = f^{(1)}(x) - f^{(2)}(x)$$
$$x^* = \arg\min_{x \in U}\left[f^{(1)}(x) - f^{(2)}(x)\right] \tag{6}$$

di mana $f^{(1)}$dan $f^{(2)}$adalah nilai decision function tertinggi dan tertinggi kedua dari seluruh kelas. Sampel dengan margin terkecil adalah yang paling ambigu dan diprioritaskan untuk dilabeli.

Sampel terpilih kemudian dikirim ke tahap *Human Labeling*, di mana peneliti utama bertindak sebagai *Oracle* (*annotator*). Pelabelan dilakukan secara manual dengan mengacu pada panduan anotasi yang telah disusun untuk tiga kelas sentimen (Negatif, Netral, dan Positif). Peneliti memiliki latar belakang pendidikan Informatika dan pengalaman anotasi teks sentimen selama kurang lebih enam bulan. Untuk menjaga konsistensi, setiap sampel yang dianggap ambigu atau sulit diputuskan dicatat dan dilakukan pengecekan ulang. Data yang telah diberi label kemudian digabungkan ke dalam dataset latih melalui tahap *Update Labeled* [9].

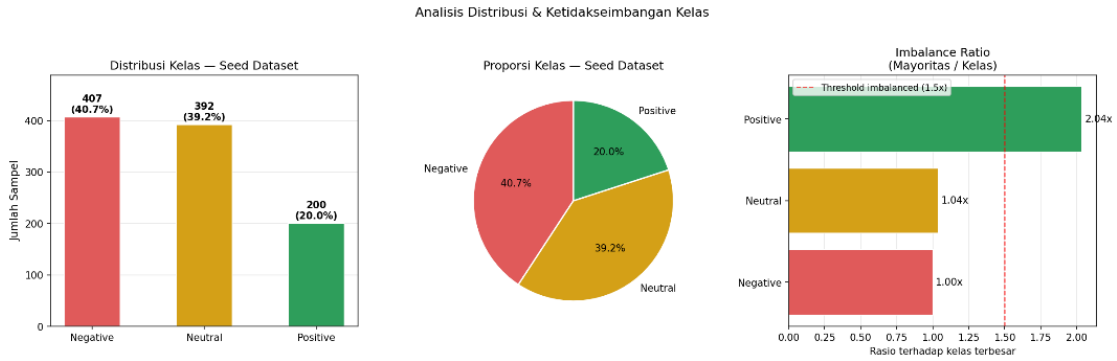
Dataset yang telah diperbarui digunakan kembali pada tahap *Retrain* SVM Model, sehingga model dapat memperbaiki batas keputusannya berdasarkan informasi baru yang lebih relevan. Selanjutnya, sistem melakukan pengecekan pada tahap *Stopping Criteria Met*. Kriteria penghentian dalam penelitian ini ditentukan berdasarkan jumlah iterasi maksimum atau keterbatasan data pada *unlabeled pool*. Jika kriteria belum terpenuhi (*No*), maka proses akan kembali ke tahap *Unlabeled Pool* dan siklus *active learning* diulang [6]. Sebaliknya, jika kriteria penghentian telah terpenuhi (*Yes*), maka proses dilanjutkan ke tahap *Final Evaluation*, di mana model dievaluasi menggunakan metrik performa seperti *accuracy*, *precision*, *recall*, dan *F1-score* [10]. Tahap ini bertujuan untuk mengukur efektivitas pendekatan *active learning* dibandingkan metode *baseline*. Proses diakhiri pada tahap End, yang menandai selesainya seluruh rangkaian eksperimen. Struktur iteratif pada metode ini memungkinkan model untuk belajar secara adaptif dengan memaksimalkan informasi dari data yang paling ambigu, sehingga meningkatkan efisiensi pelabelan dibandingkan pendekatan konvensional berbasis *random sampling*.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan terdiri atas 999 komentar YouTube berlabel yang dikategorikan ke dalam tiga kelas sentimen: Negatif (label 0), Netral (label 1), dan Positif (label 2). Distribusi label pada dataset disajikan pada Tabel 1 dan Analisis Distribusinya disajikan pada gambar 2. Kelas Negatif mendominasi dengan proporsi tertinggi (40,7%), diikuti Netral (39,2%), dan Positif sebagai kelas minoritas (20,0%). Ketidakseimbangan kelas ini berbeda dari asumsi umum analisis sentimen dan mencerminkan karakteristik diskusi publik terhadap MBG yang lebih banyak berisi kritik dan komentar netral daripada dukungan eksplisit. Parameter *class_weight='balanced'* diterapkan pada SVM untuk menangani kondisi ini.

Tahap prapemrosesan diterapkan secara seragam meliputi konversi huruf kecil, penghapusan URL, *mention* (@), tagar (#), emoji, tanda baca, dan angka, serta normalisasi spasi. Statistik panjang komentar menunjukkan rata-rata 30,9 kata per komentar dengan standar deviasi tinggi (σ=40,4 kata), mencerminkan heterogenitas konten yang umum pada data komentar

YouTube. Proses prapemrosesan hanya mereduksi rata-rata panjang komentar sebesar 1,1%, mengindikasikan bahwa sebagian besar konten komentar bersifat substantif. Dataset kemudian dibagi menggunakan *stratified split* dengan rasio 60/20/20, menghasilkan 599 sampel *initial train*, 200 sampel *oracle pool*, dan 200 sampel *fixed test set* yang tidak berubah selama seluruh eksperimen disajikan pada gambar 3. Distribusi kelas pada ketiga subset konsisten yakni, Negatif =40%, Netral =39%, Positif =20% memvalidasi keberhasilan stratifikasi.

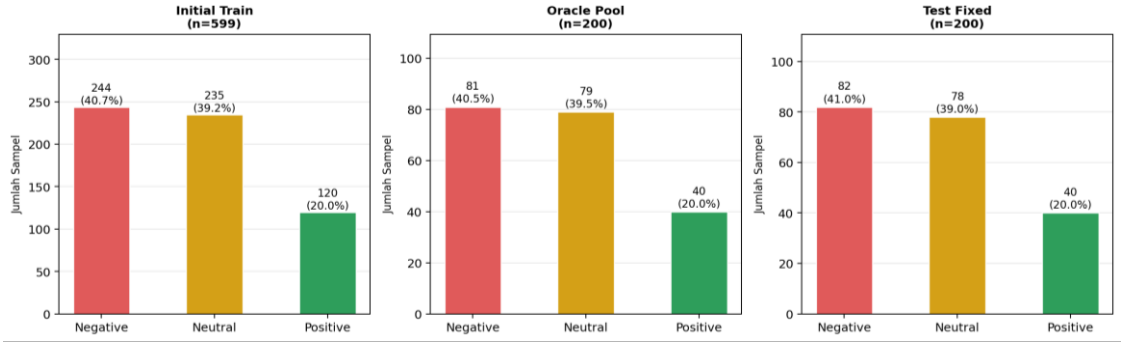


Gambar 2. Analisis Distribusi & Ketidakseimbangan Kelas

Tabel 1. Distribusi Kelas pada Seed Dataset

Kelas Sentimen	Jumlah Sampel	Persentase (%)
Negatif (0)	407	40,7
Netral (1)	392	39,2
Positif (2)	200	20
Total	999	100

Ekstraksi fitur dilakukan menggunakan TF-IDF dengan konfigurasi yang disajikan pada Tabel 2. *Vectorizer* di-*fit* hanya pada 599 sampel *initial train* untuk mencegah *data leakage*. Analisis *out-of-vocabulary* (OOV) menunjukkan nol *zero-vector* pada ketiga subset, mengkonfirmasi tidak adanya *covariate shift* linguistik yang signifikan.



Gambar 3. Distribusi Label per Subset Split (Stratified 60/20/20)

Tabel 2. Konfigurasi TF-IDF

Parameter	Nilai
max_features	5.000
ngram_range	(1, 2) — *unigram & bigram*
sublinear_tf	True
Fit pada	*Seed dataset* (800 sampel train)

.

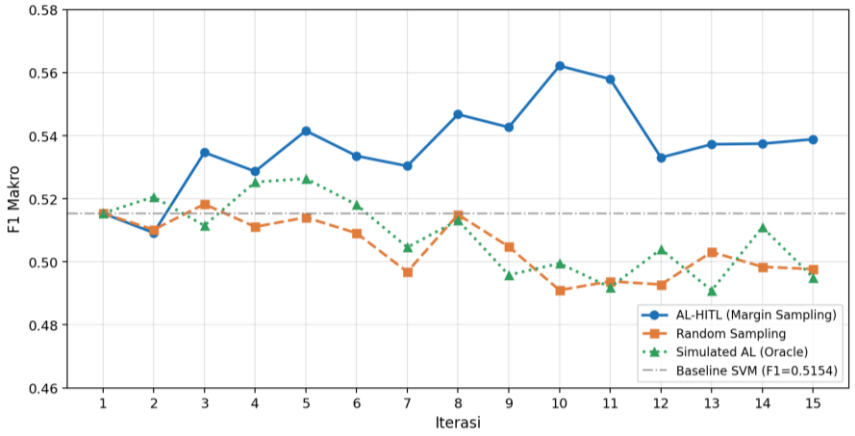
Sebelum dimulainya siklus *Active Learning*, model SVM Linear dengan *kernel* linear, parameter C sebesar 1.0, dan bobot kelas seimbang dilatih menggunakan 599 sampel awal sebagai baseline. Hasil evaluasinya disajikan pada Tabel 3.

Tabel 3. Performa Model SVM Baseline

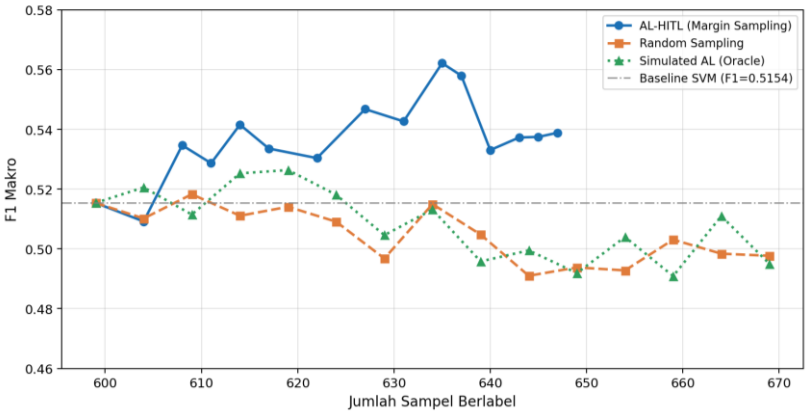
Model	Accuracy	Precision	Recall	F1 Macro
SVM Linear	0,5300	0,5307	0,5088	0,5154

Performa *baseline* yang berada di *F1 Macro* 0,5154 mencerminkan tantangan klasifikasi tiga kelas pada domain ini. Beberapa faktor yang diduga berkontribusi pada rendahnya performa awal adalah: (1) representasi TF-IDF yang bersifat statistik tidak mampu menangkap nuansa semantik bahasa informal Indonesia termasuk singkatan, kata gaul, dan *code-mixing*; (2) komentar YouTube MBG cenderung mengandung campuran ekspresi positif-bersyarat dan kritik halus dalam satu komentar; serta (3) kelas Positif yang merupakan kelas minoritas (20%) memiliki representasi fitur yang tumpang tindih dengan dua kelas lainnya, menyebabkan *decision boundary* yang tidak tegas pada ruang fitur TF-IDF.

Eksperimen *Active Learning* dijalankan selama 15 iterasi dengan penambahan maksimum 5 sampel per iterasi menggunakan strategi *Margin Sampling*. Proses pelabelan HITL dilakukan langsung oleh peneliti sebagai *Oracle* dengan panduan anotasi tiga kelas. Mekanisme *skip* tersedia untuk komentar yang dinilai terlalu ambigu. Perbandingan perkembangan *F1 Macro* antara tiga kondisi eksperimen (AL-HITL, *Simulated* AL, dan *Random Sampling*) ditampilkan pada Gambar 4 dan Gambar 5.



Gambar 4. Kurva Pembelajaran F1 Macro vs Iterasi



Gambar 5. Label Efficiency Curve: F1 Macro vs Jumlah Sampel Berlabel

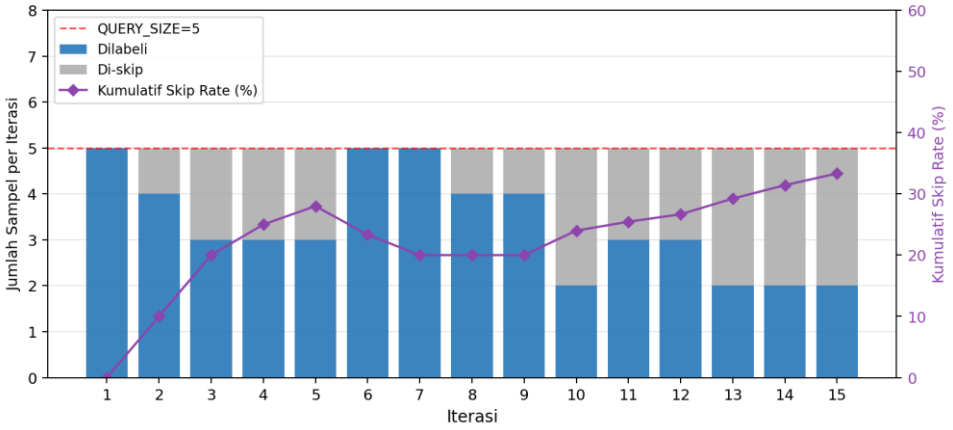
Nilai *F1 Macro* pada iterasi-iterasi kunci untuk ketiga kondisi eksperimen disajikan secara lengkap pada Tabel 4.

Tabel 4. Perkembangan F1 Macro per Iterasi Active Learning

Iterasi	Labeled Size	AL-HITL (Margin)	Simulated AL	Random
0 (Baseline)	599	0,5154	0,5154	0,5154
3	608	0,5347	0,5115	0,5183
5	614	0,5416	0,5264	0,5141
8	627	0,5468	0,5132	0,5150
10	635	0,5622	0,4995	0,4910
12	640	0,5331	0,5040	0,4928
15 (Akhir)	647	0,5389	0,4950	0,4977

Berdasarkan Tabel 4 dan Gambar 4, terdapat empat temuan utama. Pertama, AL-HITL berbasis *Margin Sampling* mencapai *F1 Macro* tertinggi pada iterasi ke-15 sebesar 0,5389, melampaui Random Sampling (0,4977) dan Simulated AL (0,4950). Peningkatan dari *baseline* sebesar Δ+0,0235 dicapai hanya dengan 50 sampel berlabel tambahan yang dipilih secara informatif oleh strategi *Margin Sampling*. Kedua, temuan yang paling signifikan secara konseptual adalah bahwa AL-HITL justru mengungguli *Simulated* AL yang menggunakan label *ground-truth* tanpa *noise*. Analisis mendalam menunjukkan bahwa mekanisme *skip* dalam proses HITL dengan *skip rate* keseluruhan sebesar 33,3% (25 dari 75 sampel diajukan) secara implisit bertindak sebagai *quality filter*: sampel yang paling ambigu secara semantik dihindari dari proses pelabelan, mencegah masuknya *noisy label* ke dalam set latih. Dalam kondisi *Simulated* AL, seluruh 75 sampel dipaksakan masuk termasuk yang sangat ambigu, sehingga beberapa sampel justru merusak *decision boundary* yang telah terbentuk. Ketiga, *Random Sampling* dan *Simulated* AL mengalami *negative learning*: nilai *F1 Macro* akhir keduanya berada di bawah nilai *baseline* (masing-masing 0,4977 dan 0,4950 vs. baseline 0,5154). Pada *Random Sampling*, penambahan sampel acak memperkenalkan data yang tidak relevan secara geometris bagi *hyperplane* SVM. Keempat, fluktuasi *F1 Macro* yang cukup tinggi antar-iterasi pada ketiga kondisi mengindikasikan bahwa dengan *QUERY_SIZE*=5, perubahan *decision boundary* SVM sangat sensitif terhadap karakteristik individu sampel yang terpilih, sehingga interpretasi berfokus pada tren kumulatif 15 iterasi dan bukan perubahan per-iterasi.

Distribusi pelabelan dan *skip* per iterasi disajikan pada Gambar 6 dan Tabel 5. Tren *skip rate* kumulatif yang meningkat seiring bertambahnya iterasi mengindikasikan bahwa sampel yang dipilih *Margin Sampling* di iterasi akhir semakin ambigu konsisten dengan ekspektasi teoritis bahwa setelah iterasi awal, mayoritas sampel yang mudah dilabeli sudah terpilih dan yang tersisa adalah kasus-kasus batas.

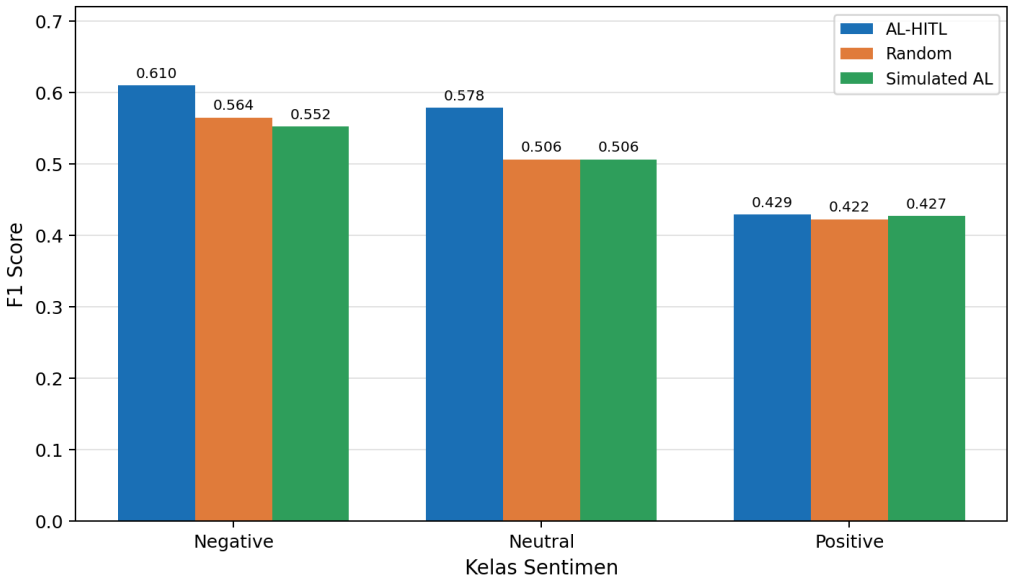


Gambar 6. Distribusi Pelabelan dan Skip per Iterasi AL-HITL

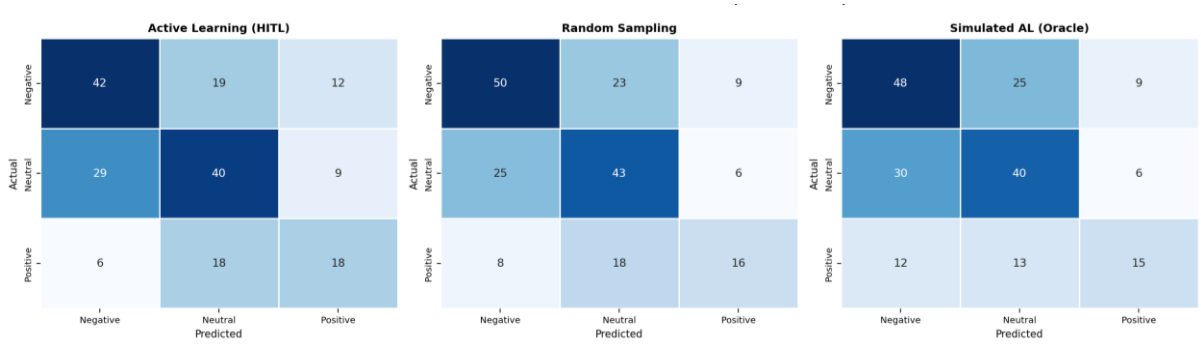
Tabel 5. Detail Pelabelan dan Skip per Iterasi AL-HITL

Iterasi	Labeled Size	Dilabeli	Di-skip	Pool Sisa	F1 Macro
1	599	5	0	195	0,5154
2	599	4	1	190	0,5092
3	604	3	2	185	0,5347
4	608	3	2	180	0,5287
5	611	3	2	175	0,5416
6	614	5	0	170	0,5336
7	617	5	0	165	0,5304
8	622	4	1	160	0,5468
9	627	4	1	155	0,5427
10	631	2	3	150	0,5622
11	635	3	2	145	0,5580
12	637	3	2	140	0,5331
13	640	2	3	135	0,5373
14	643	2	3	130	0,5375
15	645	2	3	125	0,5389
Total	—	50	25	130	—

Analisis *F1-score* per kelas dan *confusion matrix* pada model iterasi ke-15 disajikan pada Gambar 7, Gambar 8, dan Tabel 6.



Gambar 7. F1-Score per Kelas — AL-HITL vs Random Sampling vs Simulated AL (iterasi ke-15)



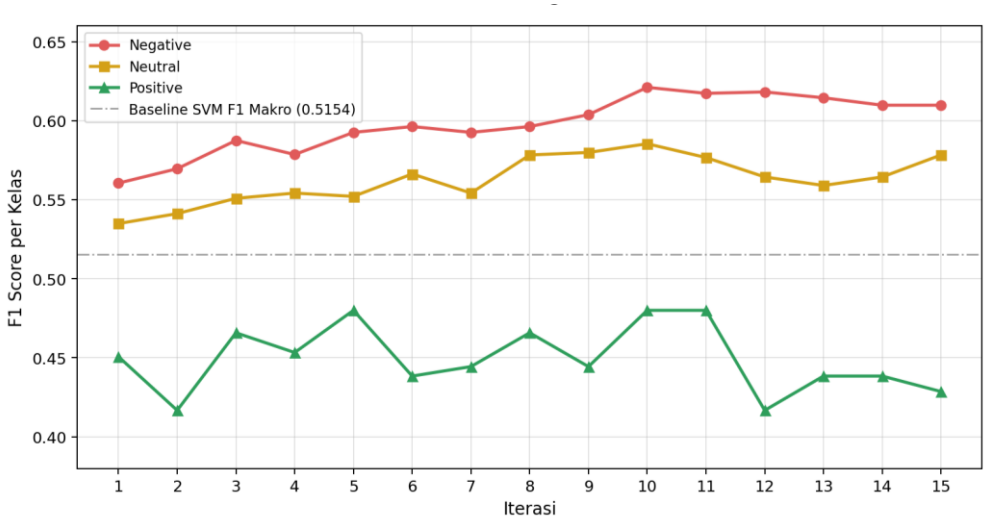
Gambar 8. Confusion Matrix (iterasi ke-15): (a) AL-HITL, (b) Random Sampling, (c) Simulated AL

Tabel 6. F1-Score per Kelas pada Iterasi Terakhir

Kelas	AL-HITL	Random	Simulated AL
Negative	0,6098	0,5644	0,5521
Neutral	0,5783	0,5060	0,5062
Positive	0,4286	0,4225	0,4267

Dari Tabel 6 dan Gambar 7 tampak pola yang konsisten di seluruh kondisi eksperimen: **F1 Negatif > F1 Netral > F1 Positif**. Pola ini sejalan dengan kondisi ketidakseimbangan dataset, di mana kelas Negatif sebagai kelas mayoritas (40,7%) memiliki jumlah contoh latih terbanyak sehingga *decision boundary*-nya lebih mudah dibentuk oleh SVM. Meskipun *class_weight='balanced'* diterapkan untuk mengkompensasi ketidakseimbangan tersebut, kelas Positif sebagai kelas minoritas (20%) tetap memperoleh *F1-score* terendah di seluruh kondisi, mengindikasikan bahwa permasalahan utama bukan semata-mata pada ketidakseimbangan jumlah sampel. Kelas Negatif memperoleh *F1-score* tertinggi (0,6098) karena komentar negatif terhadap MBG cenderung menggunakan kosakata yang lebih spesifik dan khas seperti kata-kata kritis dan penolakan eksplisit, sehingga representasi TF-IDF mampu menangkap fitur pembeda yang lebih tegas bagi SVM. Kelas Netral memperoleh F1 sebesar 0,5783 pada AL-HITL secara signifikan lebih tinggi dari Random (0,5060) dan *Simulated* AL (0,5062) mengindikasikan bahwa mekanisme *skip* HITL secara tidak langsung membantu akurasi kelas Netral dengan menghindari pelabelan sampel ambigu yang berpotensi *mislabeled*. Kelas Positif secara konsisten memperoleh *F1-score* terendah (0,4286). Rendahnya performa kelas ini disebabkan oleh heterogenitas semantik yang tinggi: komentar positif terhadap MBG mencakup pujian langsung, dukungan bersyarat, ekspresi aspiratif, hingga apresiasi bersyarat, sehingga representasi TF-IDF yang dihasilkan tersebar luas dan tumpang tindih dengan kedua kelas lainnya di ruang fitur. Temuan ini konsisten dengan analisis *confusion matrix* pada Gambar 8: proporsi kesalahan klasifikasi tertinggi terdapat pada kelas Positif (60% sampel aktual Positif salah diprediksi), sedangkan misklasifikasi silang antara kelas Negatif dan Positif yang memiliki polaritas paling berjauhan terjadi paling jarang.

Dinamika perkembangan *F1-score* per kelas selama siklus *Active Learning* AL-HITL ditampilkan pada Gambar 9. F1 Negatif menunjukkan tren naik yang paling konsisten dari 0,5605 pada *baseline* ke puncak 0,6211 (iterasi 10), sebelum stabil di 0,6098 pada iterasi akhir. F1 Netral menunjukkan tren positif dengan *volatilitas* sedang, naik dari 0,5349 ke 0,5783 di iterasi akhir. F1 Positif menunjukkan fluktuasi tinggi tanpa tren naik yang jelas dari 0,4507 menjadi 0,4286, mengkonfirmasi kesulitan model dalam mempelajari *decision boundary* kelas ini secara inkremental.



Gambar 9. Class-wise Learning Curve AL-HITL Perkembangan F1 per Kelas

Ringkasan efisiensi pelabelan ketiga kondisi disajikan pada Tabel 7. AL-HITL adalah satu-satunya kondisi yang menghasilkan peningkatan *F1 Macro* dari baseline, sementara *Random Sampling* dan *Simulated* AL mengalami penurunan.

Tabel 7. Ringkasan Efisiensi Pelabelan

Kondisi	F1 Baseline	F1 Akhir	Δ F1	Sampel Ditambahkan
AL-HITL (Margin Sampling)	0,5154	0,5389	+0,0235	50 (dari 75 diajukan)
Simulated AL (Oracle)	0,5154	0,4950	−0,0204	75
Random Sampling	0,5154	0,4977	−0,0177	75

Hasil ini menunjukkan bahwa dengan menginvestasikan upaya selektif dalam proses pelabelan termasuk kemampuan untuk *skip* sampel yang tidak dapat dilabeli dengan keyakinan tinggi peneliti dapat memperoleh model yang lebih baik dibandingkan melabeli secara acak maupun mensimulasikan *Oracle* sempurna. Namun demikian, penting untuk memposisikan temuan ini secara hati-hati: nilai *F1 Macro* akhir 0,5389 masih berada di bawah ambang 0,6, sehingga performa model belum dapat dikategorikan optimal untuk keperluan praktis. Keunggulan AL-HITL lebih tepat dipahami sebagai peningkatan relatif dalam kondisi keterbatasan sumber daya anotasi, bukan sebagai bukti final efisiensi metode. Fenomena *negative learning* yang diamati pada *Random Sampling* dan *Simulated* AL serta *F1 Positif* yang konsisten terendah menggarisbawahi bahwa pengembangan lebih lanjut khususnya integrasi representasi semantik berbasis konteks seperti *IndoBERT* dan eksplorasi strategi kueri yang lebih beragam masih diperlukan untuk meningkatkan kualitas klasifikasi analisis sentimen sumber daya terbatas pada domain isu publik yang kompleks [16]. Penelitian ini dengan demikian berkontribusi sebagai studi pendahuluan yang memberikan bukti empiris awal mengenai potensi dan tantangan implementasi *Active Learning* berbasis *Margin Sampling* pada analisis sentimen komentar YouTube MBG.

4. KESIMPULAN

Penelitian ini telah mengimplementasikan dan mengevaluasi kerangka *Active Learning* (AL) berbasis *Support Vector Machine* (SVM) dengan strategi *Margin Sampling* dan pendekatan *Human-in-the-Loop* (HITL) untuk analisis sentimen komentar YouTube terkait program Makan Bergizi Gratis (MBG). Eksperimen dijalankan pada 999 komentar YouTube berlabel yang dibagi dengan rasio 60/20/20 menjadi 599 sampel latih awal, 200 sampel *Oracle pool*, dan 200 sampel uji tetap, dengan representasi teks TF-IDF berkonfigurasi *unigram* dan *bigram* (*max_features*=5.000). Tiga kondisi dibandingkan selama 15 iterasi: AL-HITL berbasis *Margin Sampling*, *Simulated Active Learning*, dan *Random Sampling* sebagai *baseline*.

Pertama, AL-HITL berbasis *Margin Sampling* adalah satu-satunya kondisi yang menghasilkan peningkatan *F1 Macro* dari *baseline*. Model AL-HITL mencapai *F1 Macro* sebesar 0,5389 pada iterasi ke-15, meningkat Δ+0,0235 dari *baseline* SVM (0,5154) hanya dengan menambahkan 50 sampel berlabel dalam 15 iterasi. Peningkatan ini lebih tinggi dibandingkan *Simulated* AL (0,4950) maupun *Random Sampling* (0,4977) yang keduanya justru mengalami penurunan performa di bawah *baseline*. *Margin Sampling* memanfaatkan karakteristik *decision function* SVM untuk mengidentifikasi sampel yang paling ambigu secara geometris, sehingga setiap sampel yang dilabeli memberikan kontribusi yang lebih terarah terhadap pembaruan *hyperplane*.

Kedua, temuan yang paling signifikan secara konseptual adalah mekanisme *skip* dalam proses HITL yang secara implisit berperan sebagai *quality filter*. Dari 75 sampel yang diajukan sistem (15 iterasi × 5 sampel), sebanyak 25 sampel di-*skip* oleh peneliti sehingga hanya 50 sampel yang benar-benar masuk ke dalam set latih (*skip rate* 33,3%). Sampel yang di-*skip* cenderung merupakan komentar yang terlalu ambigu untuk dilabeli dengan keyakinan tinggi. Dengan tidak memaksakan pelabelan pada sampel tersebut, AL-HITL justru berhasil menghindari masuknya

noisy label ke dalam set latih sesuatu yang tidak terjadi pada kondisi *Simulated AL* yang memaksakan semua 75 sampel dilabeli. Hal inilah yang menjelaskan mengapa AL-HITL mengungguli *Simulated* AL, sebuah fenomena yang secara konseptual bertentangan dengan asumsi teoritis bahwa *oracle* sempurna seharusnya menghasilkan performa lebih baik.

Ketiga, fenomena *negative learning* yang diamati pada *Random Sampling* (F1 akhir 0,4977 < *baseline* 0,5154) dan *Simulated* AL (0,4950 < 0,5154) perlu mendapat perhatian serius. Penambahan sampel yang tidak selektif baik secara acak maupun melalui seleksi berbasis ketidakpastian tanpa mekanisme penyaringan berpotensi memperkenalkan data yang mengganggu *decision boundary* SVM yang sudah terbentuk. Pada domain analisis sentimen isu publik dengan tingkat ambiguitas semantik tinggi seperti komentar YouTube MBG, kondisi ini justru lebih rentan terjadi dibandingkan pada domain teks yang lebih formal dan terstruktur. Temuan ini menggarisbawahi bahwa keberhasilan AL tidak hanya ditentukan oleh strategi pemilihan sampel, tetapi juga oleh kualitas proses pelabelan yang menyertainya.

Keempat, pola F1-score per kelas yang konsisten di seluruh kondisi Negatif (0,6098) > Netral (0,5783) > Positif (0,4286) mengungkap tantangan struktural yang lebih dalam. Kelas Negatif sebagai kelas mayoritas (40,7%) memperoleh F1 tertinggi karena kosakatanya yang khas dan spesifik memudahkan pembentukan *decision boundary* yang tegas dalam ruang fitur TF-IDF. Sebaliknya, kelas Positif sebagai kelas minoritas (20,0%) secara konsisten memperoleh F1 terendah meskipun *class_weight='balanced'* diterapkan, mengindikasikan bahwa permasalahannya bukan semata-mata pada ketidakseimbangan jumlah sampel, melainkan pada heterogenitas semantik yang tinggi: komentar positif terhadap MBG mencakup berbagai ekspresi mulai dari pujian langsung hingga dukungan bersyarat yang memiliki representasi TF-IDF tumpang tindih dengan kelas lain. Representasi berbasis frekuensi statistik seperti TF-IDF terbukti belum mampu menangkap nuansa semantik tersebut secara memadai.

Kelima, penelitian ini perlu diposisikan secara jujur sebagai studi pendahuluan yang mengeksplorasi tantangan implementasi AL-HITL pada domain analisis sentimen isu publik berbahasa Indonesia. Terdapat keterbatasan metodologis yang perlu diakui: *Oracle pool* dalam eksperimen ini berasal dari partisi 200 sampel yang telah memiliki label tersembunyi dari 999 data berlabel, bukan dari pool data yang benar-benar tidak berlabel. Kondisi ini membatasi validitas klaim efisiensi pelabelan dalam konteks *real-world* di mana peneliti menghadapi pool data yang murni tidak berlabel. Selain itu, penggunaan satu *annotator* tunggal sebagai *Oracle* memberikan konsistensi internal pelabelan, namun membatasi objektivitas dan generalisabilitas panduan anotasi. Dengan demikian, peningkatan performa AL-HITL yang diamati (Δ+0,0235) lebih tepat diinterpretasikan sebagai *bukti empiris awal* yang mendukung eksplorasi lebih lanjut, bukan sebagai bukti final efektivitas metode.

Berdasarkan temuan dan keterbatasan tersebut, beberapa rekomendasi untuk penelitian lanjutan dapat dikemukakan. Pertama, replikasi eksperimen menggunakan pool data yang benar-benar tidak berlabel (seperti 7.967 komentar mentah yang tersedia dalam penelitian ini) akan memberikan validasi yang lebih kuat terhadap efisiensi AL dalam kondisi *real-world*. Kedua, integrasi representasi teks berbasis konteks seperti IndoBERT atau model *Transformer* lainnya berpotensi meningkatkan performa klasifikasi secara signifikan, khususnya pada kelas Positif yang memiliki heterogenitas semantik tinggi. Ketiga, eksplorasi strategi kueri yang lebih beragam seperti *Core-Set*, *BADGE*, atau *Expected Model Change* dapat dibandingkan dengan *Margin Sampling* untuk mendapatkan gambaran yang lebih komprehensif. Keempat, pelibatan lebih dari satu anotator dengan mekanisme *inter-annotator agreement* yang terstruktur akan meningkatkan objektivitas proses pelabelan HITL dan mengurangi dampak *annotation noise*. Secara keseluruhan, penelitian ini memberikan kontribusi berupa bukti empiris awal dan pelajaran metodologis yang relevan bagi pengembangan sistem analisis sentimen berbasis AL pada domain data publik berbahasa Indonesia yang kompleks.

DAFTAR PUSTAKA

[1] A. Sitanggang, Y. Umaidah, R. I. Adam, U. S. Karawang, and T. Timur, "ANALISIS SENTIMEN MASYARAKAT TERHADAP PROGRAM MAKAN SIANG GRATIS PADA MEDIA," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4902. https://doi.org/10.23960/jitet.v12i3.4902

[2] T. A. Aziz, I. Ismayadi, and B. Budiman, "Analisis Sentimen Terhadap Program Makan Siang Gratis pada Media Sosial X Menggunakan Logistic Regression dan SVM," In Search, vol. 24, no. 1, pp. 18-28, 2025, doi: 10.37278/insearch.v24i1.1238. https://doi.org/10.37278/insearch.v24i1.1238

[3] Saurabh Chakraborty, Rutika Zambre, Hrishikesh Sharma, Tapash Gaikwad, and Dr. Nitin Janwe, "Sentiment Analysis of YouTube Comments," International Journal of Advanced Research in Science, Communication and Technology, pp. 194-198, 2024, doi: 10.48175/ijarsct-18131. https://doi.org/10.48175/IJARSCT-18131

[4] M. Liebenlito, N. Inayah, E. Choerunnisa, T. E. Sutanto, and S. Inna, "Active Learning on Indonesian Twitter Sentiment Analysis Using Uncertainty Sampling," Journal of Applied Data Sciences, vol. 5, no. 1, pp. 114-121, 2024, doi: 10.47738/jads.v5i1.144. https://doi.org/10.47738/jads.v5i1.144

[5] S. M. Daudpota, S. Hassan, Y. Alkhurayyif, A. S. Alqahtani, and M. H. Aziz, "Active Learning Strategies for Textual Dataset-Automatic Labelling," Computers, Materials & Continua, 2023, doi: 10.32604/cmc.2023.034157. https://doi.org/10.32604/cmc.2023.034157

[6] N. W. S. Saraswati, I. K. G. D. Putra, M. Sudarma, and I. M. Sukarsa, "Enhance Sentiment Analysis in Big Data Tourism using Hybrid Lexicon and Active Learning Support Vector Machine," Bulletin of Electrical Engineering and Informatics, vol. 13, no. 5, pp. 3663-3674, 2024, doi: 10.11591/eei.v13i5.7807. https://doi.org/10.11591/eei.v13i5.7807

[7] Settles, B. (2010). Computer sciences active learning literature survey. Laporan Teknis, University of Wisconsin-Madison. https://burrsettles.com/pub/settles.activelearning.pdf

[8] Sekar Cinta Amaria and Nurtriana Hidayati, "ANALISIS PROGRAM MAKAN BERGIZI GRATIS DENGAN SUPPORT VECTOR MACHINE (SVM) PADA APLIKASI X," JSiI (Jurnal Sistem Informasi), vol. 12, no. 2, pp. 16-24, 2025, doi: 10.30656/jsii.v12i2.10708. https://doi.org/10.30656/jsii.v12i2.10708

[9] M. Venugopalan and D. Gupta, "A Reinforced Active Learning Approach for Optimal Sampling in Aspect Term Extraction for Sentiment Analysis," Expert Syst. Appl., vol. 209, no. May, p. 118228, 2022, doi: 10.1016/j.eswa.2022.118228. https://doi.org/10.1016/j.eswa.2022.118228

[10] J. Z. Bengar, "Reducing Label Effort : Self-Supervised meets Active Learning," IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1631-1639, 2021, doi: 10.1109/iccvw54120.2021.00188. https://doi.org/10.1109/ICCVW54120.2021.00188

[11] D. Kartchner, D. N. An, W. Ren, C. Zhang, and C. S. Mitchell, "Rule-Enhanced Active Learning for Semi-Automated Weak Supervision," AI, vol. 3, no. 1, pp. 211-228, 2022, doi: 10.3390/ai3010013. https://doi.org/10.3390/ai3010013

[12] M. Learning, "Machine Learning as an Experimental Science," Mach. Learn., vol. 3, no. 1, pp. 5-8, 1988, doi: 10.1023/A:1022623814640. https://doi.org/10.1023/A:1022623814640

[13] J. Kamiri and G. Mariga, "Research Methods in Machine Learning: A Content Analysis," International Journal of Computer and Information Technology, vol. 10, no. 2, 2021, doi: 10.24203/ijcit.v10i2.79. https://doi.org/10.24203/ijcit.v10i2.79

[14] A. Andhini, F. N. Handayani, I. Diasih, and Nurmalitasari, "Analisis Sentimen Opini Publik pada Channel Youtube Mata Najwa Menggunakan Metode SVM Universitas Duta Bangsa Surakarta , Indonesia Perkembangan pesat media sosial dan platform video seperti YouTube telah menciptakan ruang baru bagi masyarakat untuk menge," Jurnal Teknik Informatika dan Teknologi Informasi, vol. 5, no. 2, 2025. https://doi.org/10.55606/jutiti.v5i2.5426

[15] M. Afnan Ul Haque, A. Rahman, and M. M. A. Hashem, "Sentiment Analysis in Low-Resource Bangla Text Using Active Learning," 2021 5th International Conference on

Electrical Information and Communication Technology, EICT 2021, no. December, pp. 17-19, 2021, doi: 10.1109/EICT54103.2021.9733711. https://doi.org/10.1109/EICT54103.2021.9733711

[16] G. Yu et al., "Learning by Querying Subexamples CMAL : Cost-effective Multi-label Active Learning by Querying Subexamples," IEEE Trans. Knowl. Data Eng., vol. 34, no. 5, pp. 2091-2104, 2022, doi: 10.1109/tkde.2020.3003899. https://doi.org/10.1109/TKDE.2020.3003899