

UNIVERSITAS INDONESIA

Analisis Sentimen and Pemodelan Topik terhadap Opini Masyarakat terkait Penyelenggaraan Pemilu 2024 di Media Sosial Twitter

KARYA AKHIR

AHMAD FAUZI

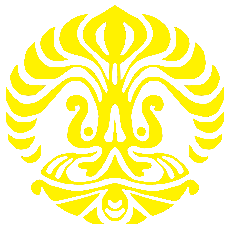
2106806630

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI

JAKARTA

JANUARI 2024



UNIVERSITAS INDONESIA

Analisis Sentimen dan Pemodelan Topik terhadap Opini Masyarakat terkait Penyelenggaraan Pemilu 2024 di Media Sosial Twitter

KARYA AKHIR

Diajukan sebagai salah satu syarat untuk memperoleh gelar

Magister Teknologi Informasi

AHMAD FAUZI

2106806630

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI

JAKARTA

JANUARI 2024

# HALAMAN PERNYATAAN ORISINALITAS

Karya Akhir ini adalah hasil karya saya sendiri,

Dan semua sumber baik yang dikutip maupun dirujuk

Telah saya nyatakan dengan benar.

Nama : …………………………………………

NPM : …………………………………………

Tanda Tangan : …………………………………………

Tanggal : …………………………………………

# HALAMAN PENGESAHAN

Karya Akhir ini ditujukan oleh:

Nama : …………………………………………

NPM : …………………………………………

Program Studi : Magister Teknologi Informasi

Judul Karya Akhir : …………………………………………

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Teknologi Informasi pada Program Studi Magister Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Indonesia.

DEWAN PENGUJI

Pembimbing I : …………………………………………… (………………)

Pembimbing II : ………………………………..………… (………………)

Penguji I : …………………………………………… (………………)

Penguji II : …………………………………………… (………………)

Ditetapkan di : Jakarta

Tanggal :

# KATA PENGANTAR

Segala puji bagi Allah Subhaanahu wata’ala karena berkat rahmat-Nya kepada penulis sehingga berhasil menyelesaikan Karya Akhir dengan judul “………………………………………………………………………………………..”. Adapun penulisan karya akhir ini disusun untuk melengkapi salah satu syarat untuk memperoleh gelar Magister Teknologi Informasi pada Universitas Indonesia.

Peneliti menyadari bahwa penelitian ini tidak akan berjalan dengan lancar tanpa bantuan dan bimbingan dari berbagai pihak, untuk itu peneliti mengucapkan terima kasih sebesar-besarnya kepada:

1. ……………………………………………………………………………………….. yang telah membimbing selama penelitian dengan penuh kesabaran.
2. Orang tua yang telah selalu tulus memberikan doa dan selalu sabar dalam memberikan dorongan untuk menyelesaiakan penyusunan Karya Akhir ini.
3. Istri yang selalu mendukung dan memahami kesibukan peneliti dan selalu mendoakan untuk kelancaran selama penelitian Karya Akhir ini.
4. Seluruh teman-teman selama perkuliahan di Program Studi Magister Teknologi Informasi di Universitas Indonesia yang memberi dukungan dalam penyelesaian Karya Akhir ini.

Akhir kata semoga Karya Akhir ini dapat bermanfaat bagi peneliti dan bagi pengembangan ilmu pengetahuan.

Jakarta, <tanggal>

<nama mahasiswa>

# HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan dibawah ini:

Nama : …………………………………………

NPM : …………………………………………

Program Studi : Magister Teknologi Informasi

Fakultas : Ilmu Komputer

Jenis Karya : Karya Akhir

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalti-Free Right*)** atas Karya Akhir saya yang berjudul:

…………………………………………………………………………………………….

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan Karya Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemiliki Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di: …………………………………………

Pada tanggal: …………………………………………

Yang menyatakan

ttd

(…………………)

# ABSTRAK

|  |  |
| --- | --- |
| Nama : | Ahmad Fauzi |
| Program Studi : | Magister Teknologi Informasi |
| Judul : | Analisis Sentimen dan Pemodelan Topik terhadap Opini Masyarakat terkait Penyelenggaraan Pemilihan Umum 2024 |
| Pembimbing : | …………………………………………  ………………………………………… |

<Abstract Indonesia>

Kata kunci: Analisis Sentimen, Pemilu, Pelanggaran Pemilu

# ABSTRACT

|  |  |
| --- | --- |
| Name : | Ahmad Fauzi |
| Study Program : | Magister Teknologi Informasi |
| Title : | Analisis Sentimen dan Pemodelan Topik terhadap Opini Masyarakat terkait Penyelenggaraan Pemilihan Umum 2024 |
| Counsellor : | …………………………………………  ………………………………………… |

<Abstract English>

Keywords: Sentiment Analysis, Electoral Integrity, Electoral Malpractice

# DAFTAR ISI

HALAMAN JUDUL.........................................................................................................i

[HALAMAN PERNYATAAN ORISINALITAS ii](#_Toc160717490)

[HALAMAN PENGESAHAN iii](#_Toc160717491)

[KATA PENGANTAR v](#_Toc160717492)

[HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS vi](#_Toc160717493)

[ABSTRAK vii](#_Toc160717494)

[ABSTRACT viii](#_Toc160717495)

[DAFTAR ISI ix](#_Toc160717496)

[DAFTAR GAMBAR xii](#_Toc160717497)

[DAFTAR TABEL xiii](#_Toc160717498)

[DAFTAR LAMPIRAN xiv](#_Toc160717499)

[BAB 1 PENDAHULUAN 15](#_Toc160717500)

[1.1 Latar Belakang 15](#_Toc160717501)

[1.2 Rumusan Masalah 23](#_Toc160717502)

[1.3 Pertanyaan Penelitian 24](#_Toc160717503)

[1.4 Tujuan Penelitian 25](#_Toc160717504)

[1.5 Ruang Lingkup Penelitian 25](#_Toc160717505)

[1.6 Manfaat Penelitian 25](#_Toc160717506)

[1.7 Sistematika Penulisan 26](#_Toc160717507)

[BAB 2 STUDI LITERATUR 28](#_Toc160717508)

[1.1 Media Sosial 28](#_Toc160717509)

[2.2 Analisa Media Sosial 29](#_Toc160717510)

[2.3 CRIPS-DM 31](#_Toc160717511)

[2.4 Analisis Sentimen 33](#_Toc160717512)

[2.5 Text Mining 34](#_Toc160717513)

[*2.5.1* *Data Collection* 35](#_Toc160717514)

[*2.5.2* *Text Preprocessing* 36](#_Toc160717515)

[*2.5.3* *Text Representation* 38](#_Toc160717516)

[*2.5.4* *Knowledge Discovery* 39](#_Toc160717517)

[2.6 Penelitian Terdahulu 51](#_Toc160717518)

[2.6.1 Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi 51](#_Toc160717519)

[2.6.2 Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19 52](#_Toc160717520)

[2.6.3 Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter. 53](#_Toc160717521)

[2.6.4 Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model *Machine Learning*. 55](#_Toc160717522)

[2.6.5 Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia. 57](#_Toc160717523)

[2.7Ringkasan Penelitian Terdahulu 59](#_Toc160717524)

[2.8Perbadingan Penelitian Terdahulu 61](#_Toc160717525)

[2.9Kerangka Penelitian 66](#_Toc160717526)

[BAB 3 METODOLOGI PENELITIAN 68](#_Toc160717527)

[3.1Rancangan Penelitian 68](#_Toc160717528)

[3.2Alur Penelitian 70](#_Toc160717529)

[3.2.1 Pengumpulan Data Awal 72](#_Toc160717531)

[3.2.2 Perumusan Masalah 72](#_Toc160717532)

[3.2.3 Studi Literatur 72](#_Toc160717533)

[3.2.4 Pengumpulan Data 73](#_Toc160717534)

[3.2.5 Membuat Korpus 74](#_Toc160717535)

[*3.2.6* *Data Preprocessing* 74](#_Toc160717536)

[*3.2.7* *Text Representation* 75](#_Toc160717537)

[3.2.8 Klasifikasi Sentimen dan Opini Tidak Relevan 75](#_Toc160717538)

[3.2.9 Evaluasi 76](#_Toc160717539)

[3.2.10 Pemodelan Topik 76](#_Toc160717540)

[3.2.11 Penarikan Kesimpulan 76](#_Toc160717541)

[BAB 4 HASIL DAN PEMBAHASAN 77](#_Toc160717542)

[BAB 5 KESIMPULAN DAN SARAN 78](#_Toc160717543)

[DAFTAR PUSTAKA 79](#_Toc160717544)

# DAFTAR GAMBAR

[Gambar 1. 1 Skor Pemilu berdasarkan kawasan dari waktu ke waktu 1990-2018, rentang nilai 0-4 (Van Ham, 2020). 17](#_Toc160717618)

[Gambar 1. 2 Distribusi kategori kerawanan berdasarkan jenis dimensi untuk tingkat provinsi (Sumber: IKP 2024, Bawaslu) 18](#_Toc160717619)

[Gambar 2. 1 Dua struktur *decision tree* yang berbeda yang terbentuk dari dataset yang sama 42](#_Toc160717637)

[Gambar 2. 2 Perbandingan kemungkinan hyperplane yang dapat digunakan dengan hyperplane yang optimal pada ruang 2 dimensi 43](#_Toc160717638)

[Gambar 2. 3 Contoh prediksi dengan algoritma KNN 45](#_Toc160717639)

[Gambar 2. 4 Struktur model *deep learning.* 47](#_Toc160717640)

[Gambar 2. 5 Theoretical framework 67](#_Toc160717641)

# DAFTAR TABEL

[Tabel 1. 1 Peringkat daerah provinsi dengan tingkat kerawanan tertinggi (rentang skor kerawanan Pemilu 0-100) 18](#_Toc160717673)

[Tabel 1. 2 Peringkat daerah kabupaten dan kota dengan tingkat kerawanan tertinggi (rentang skor kerawanan Pemilu 0-100) 19](#_Toc160717674)

[Tabel 1. 3 Harapan, Kenyataan, Masalah, dan Urgensi 23](#_Toc160717675)

[Tabel 2. 1 Jenis pendekatan machine learning berdasarkan karakteristik variable dependen dan kesediaan label variable dependen 40](#_Toc160717683)

[Tabel 2.2 Contoh data untuk pemodelan *decision tree* 41](#_Toc160717684)

[Tabel 2. 3 Struktur *confusion matrix* untuk tugas prediksi sederhana dengan 2 nilai nominal 48](#_Toc160717685)

[Tabel 2. 4 *Performance metrics* (*precision, recall, f1-measure*) yang digunakan pada tahapan evaluasi model. 48](#_Toc160717686)

[Tabel 2. 5 Ringkasan Penelitian Terdahulu 59](#_Toc160717687)

[Tabel 2. 6 Analisis 3C2S penelitian tentang Analisis Sentimen 61](#_Toc160717688)

[Tabel 3. 1 Ringkasan Rancangan Penelitian 68](#_Toc160717706)

[Tabel 3. 2 Alur Penelitian 70](#_Toc160717707)

[Tabel 3. 3 Contoh opini masyarakat tentang pemilu di media sosial 73](#_Toc160717708)

[Tabel 3. 4 Jumlah tweet pelabelan dari proses pelabelan. 74](#_Toc160717709)

# DAFTAR LAMPIRAN

# BAB 1 PENDAHULUAN

Pada bab ini dibahas mengenai latar belakang penelitian, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, sistematika penelitian.

## Latar Belakang

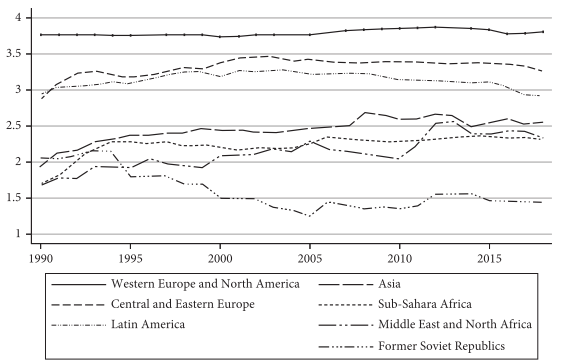
Tahun 2024 merupakan tahun politik bagi Negara Kesatuan Republik Indonesia, pasalnya, di tahun tersebut telah diselengarakannya Pemilihan Umum. Pemilihan Umum atau Pemilu adalah hak demokrasi dan perwujudan kedaulatan rakyat untuk memilih Presiden dan Wakil Presiden, anggota Dewan Perwakilan Rakyat, anggota Dewan Perwakilan Daerah dan anggota Dewan Perwakilan Rakyat Daerah (DPR RI, 2017). Berdasarkan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945, Pemilu yang diselengarakan haruslah berasaskan langsung, umum, rahasia, jujur dan adil. Maka untuk mewujudkan hal tersebut, Pemerintah telah menetapkan Undang-Undang Nomor 7 Tahun 2017 tentang Pemilihan Umum yang merupakan penyatuan dan penyederhanaan dari undang-undang sebelumnya yang mengatur tentang Pemilu.

Undang-Undang Nomor 7 Tahun 2017 diharapkan dapat menjadi wujud sistem ketatanegaraan yang menjunjung tinggi nilai-nilai demokrasi dan menjadi landasan hukum yang menjamin terselengaranya Pemilu yang ber (DPR RI, 2017). Menurut Pippa Noris, Pemilu yang ber adalah Pemilu yang mengadopsi norma dan kesepakatan universal dalam menjalankan siklusnya, mulai dari proses (a) perencanaan Pemilu, (b) pelaksanaan kampanye, (c) pemungutan dan perhitungan suara, dan (d) proses sesudahnya (Mauk, 2022). Dimana norma dan kesepakatan universal yang dimaksud merujuk pada kriteria yang dikemukakan Robert Dahl, dalam teori demokrasinya, yaitu partisipasi dan kontestasi diselenggarakan dengan jujur dan adil (Van Ham, 2020). Norma dan kesepakatan tersebut tertuang pada Pasal 25 Kovenan Internasional PBB untuk Hak Sipil dan Hak Politik pada tahun 1966 (Bagja et al., 2023).

Penelitian tentang *democratic satisfaction* dan *political trust*, menunjukkan betapa pentingnya Pemilu yang ber di mata masyarakat. Norris mengungkapkan bahwa Pemilu yang ber merupakan faktor terpenting dalam mendorong tingkat kepuasan masyarakat terhadap sistem demokrasi, bahkan jika dibandingkan dengan kebijakan yang dirasakan langsung oleh masyarakat (*policy performance*) seperti perpajakan, kesehatan, pendidikan, atau bahkan jaminan sosial (Norris, 2019). Sementara Marlene Mauk dalam penelitiannya mengungkapkan bahwa, persepsi publik akan Pemilu yang terselenggara dengan baik dan ber dapat meminimalisir terjadinya *political distrust* bagi para pendukung kontestan yang kalah, dimana *political distrust* dapat berdampak pada kurangnya dukungan publik terhadap pemerintahan yang baru terpilih, bahkan lebih dari itu, yaitu disintegrasi sosial di tengah masyarakat(Mauk, 2022).

Dalam upaya menyelenggarakan Pemilu yang ber sering kali tidak lepas dari adanya tantangan berupa temuan pelanggaran di dalamnya. Dalam konteks internasional, hal ini juga terjadi bahkan pada negara-negara yang menjadi kiblat demokrasi seperti negara-negara di kawasan Amerika Utara, Eropa Barat dan Eropa Utara, pelanggaran di dalam Pemilu masih saja terjadi (Van Ham, 2020). Pelanggaran seperti *garrymendering*[[1]](#footnote-2), manipulasi pendaftaraan pemilih, aturan *campaign finance* yang tidak adil, bias media, berita hoaks, *vote buying*, manipulasi proses pemunggutan suara, sampai dengan penggunaan teknologi yang tidak aman menjadi pelanggaran yang sering muncul di negara-negara tersebut. Walaupun skor persepsi Pemilu ber dari negara-negara tersebut masih lebih baik jika dibandingkan dengan negara di kawasan lain, sebagaimana ditunjukkan pada Gambar 1.1 (Van Ham, 2020). Hal ini telah menunjukkan bahwa pelanggaran di dalam Pemilu merupakan suatu hal yang umum terjadi (Tan, 2023).

Sedangkan dalam konteks penyelenggaraan Pemilu di Indonesia, berdasarkan data historis, terdapat beberapa laporan yang berhubungan dengan kontestasi politik di tahun-tahun sebelumnya yang dapat dikaitkan dengan adanya tantangan dalam penyelenggaraannya. Seperti yang dipublikasikan oleh BPS di tahun 2020 atau pasca Pemilu 2019, dimana ditemukan masih rendahnya tingkat kepercayaan masyarakat Indonesia terhadap proses penyelengaraan Pemilu (DataIndonesia.id, 2023). Nilai rata-rata tingkat kepercayaan untuk masing-masing provinsi adalah 69,72 poin. Provinsi Papua, Banten dan Aceh menjadi provinsi dengan tingkat kepercayaan terkecil, dengan nilai 63.7 poin, 63.88 poin dan 64.05 poin. Selain itu, berdasarkan data dari Bawaslu, ada pelanggaran netralitas ASN sebanyak 854 kasus yang terjadi selama Pilkada 2020 (Sirait et al., 2022). Pelanggaran dalam bentuk kampanye di media sosial menjadi yang terbanyak. Selain itu pelanggaran yang sama juga tercatat telah terjadi pada pemilihan presiden 2019. Hal ini tidak sejalan dengan apa yang tertuang dalam UU No 7 Tahun 2017, dimana Pemerintah, Pemerintah Provinsi, Pemerintah Kabupaten/Kota, Kecamatan, Kelurahan/Desa, Tentara Nasional Indonesia dan Kepolisian Negara republik Indonesia dilarang melakukan tindakan yang dapat menguntungkan atau merugikan Peserta Pemilu (DPR RI, 2017).



Gambar 1. 1 Skor Pemilu berdasarkan kawasan dari waktu ke waktu 1990-2018, rentang nilai 0-4.

Sumber: (Van Ham, 2020)

Kemudian berdasarkan data *post-factum* Pemilu dan Pilkada sebelumnya, pada periode 2017-2020, yang disusun Bawaslu ke dalam Indeks Kerawanan Pemilu (IKP), menunjukkan selama tahun 2017-2020 masih banyak masalah yang harus dihadapi yang berkaitan dengan Dimensi Penyelenggaraan Pemilu (Bagja et al., 2023). Hal ini dikhawatirkan masih akan berlangsung ketika Pemilu serentak 2024 diselenggarakan. IKP mengelompokan indikator pelanggaran Pemilu ke dalam 4 dimensi, yaitu: (a) Dimensi Sosial Politik yang berkaitan dengan penggunaan otoritas penyelenggara negara dan penyelenggara Pemilu, (b) Dimensi Penyelenggaraan Pemilu yang berkaitan dengan kegiatan atau tahapan di dalam Pemilu mulai dari pendaftaran pemilih, kampanye, pemungutan suara, ajudikasi, sampai dengan pengawasan, (c) Dimensi Partisipasi terkait dengan hak warga negara untuk memberikan suara dan (d) Dimensi Kontestasi terkait dengan hak warga negara untuk berkontestasi menjadi peserta Pemilu.

IKP disajikan dalam bentuk pemeringkatan daerah tingkat provinsi dan kabupaten/kota, yang dikelompokkan ke dalam 3 tingkat kategori kerawanan (a) tinggi, (b) sedang, dan (c) rendah. Dimana DKI Jakarta (dengan skor IKP 88,95), Sulawesi Utara (87,48), Maluku Utara (84,86), Jawa Barat (77,04), Kalimantan Timur (77,04) masuk sebagai 5 provinsi dengan tingkat kerawanan tertinggi. Sedangkan pada tingkat kabupaten/kota, 5 kabupaten/kota yang masuk dalam provinsi Papua dan Papua Barat teridentifikasi masuk dalam jajaran daerah dengan tingkat kerawanan tinggi. Gambar 1.2 menunjukkan sebaran tingkat kerawanan berdasarkan jenis dimensi untuk tingkat provinsi. Sementara, Tabel 1.1 dan 1.2 menunjukkan daerah yang masuk ke dalam kategori kerawanan tinggi yang dipetakan berdasarkan dimensi kerawanannya.

Gambar 1. 2 Distribusi kategori kerawanan berdasarkan jenis dimensi untuk tingkat provinsi (Sumber: IKP 2024, Bawaslu)

Sumber: (Bagja et al., 2023)

Tabel 1. 1 Peringkat daerah provinsi dengan tingkat kerawanan tertinggi (rentang skor kerawanan Pemilu 0-100)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Dimensi Sosial Politik** | **Dimensi Penyelenggaraan Pemilu** | **Dimensi Kontestasi** | **Dimensi Partisipasi** |
| 1 | Sulawesi Utara (55,67) | Banten (70,28) | DKI Jakarta (69,92) | Papua (24,08) |
| 2 | Maluku Utara (48,56) | Papua (56,09) | DI Yogyakarta (63,67) | Kalimantan Selatan (12,95) |
| 3 | Papua (46,60) | Sulawesi Tengah (54,96) | Maluku Utara (42,74) | Kalimantan Barat (11,27) |
| 4 | Jambi (43,98) | Jawa Tengah (54,58) | Jawa Barat (38,70) | Bangka Belitung (8,03) |
| 5 | Sulawesi Tengah (43,45) | Jawa Barat (51,56) | Kalimantan Barat (37,91) | Riau (7,69) |
| 6 | Nusa Tenggara Barat (43,17) | Aceh (50,04) | Banten (36,50) | Jawa Barat (6,22) |
| 7 | Sulawesi Barat (42,16) | DI Yogyakarta (49,94) | Sulawesi Tengah (35,82) | Bali (5,67) |
| 8 | Bali (38,95) | Sulawesi Utara (49,17) | Papua (35,16) | Sulawesi Utara (4,88) |
| 9 | Sulawesi Tenggara (38,02) | Maluku Utara (48,69) | Nusa Tenggara Barat (34,27) | Nusa Tenggara Timur (4,56) |
| 10 | Jawa Barat (36,79) | Nusa Tenggara Barat (47,86) | Kepulauan Riau (33,95) | Kalimantan Utara (4,18) |

Sumber: (Bagja et al., 2023)

Tabel 1. 2 Peringkat daerah kabupaten dan kota dengan tingkat kerawanan tertinggi (rentang skor kerawanan Pemilu 0-100)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Dimensi Sosial Politik** | **Dimensi Penyelenggaraan Pemilu** | **Dimensi Kontestasi** | **Dimensi Partisipasi** |
| 1 | Kabupaten Bandung, Jabar (100) | Kota Banjarbaru, Kalsel (100) | Kabupaten Intan Jaya, Papua (100) | Kabupaten Tolikara, Papua (100) |
| 2 | Kota Banjarbaru, Kalsel (100) | Kabupaten Jayawijaya, Papua (100) | Kota Semarang, Jateng (100) | Kabupaten Intan Jaya, Papua (100) |
| 3 | Kabupaten Puncak, Papua (100) | Kabupaten Timor Tengah Selatan, NTT (100) | Kabupaten Yalimo, Papua (100) | Kabupaten Majalengka, Jabar (100) |
| 4 | Kabupaten Kerinci, Jambi (100) | Kabupaten Bojonegoro, Jatim (100) | Kabupaten Fakfak, Papua Barat (100) | Kabupaten Jayawijaya, Papua (100) |
| 5 | Kabupaten Jayawijaya, Papua (100) | Kabupaten Nabire, Papua (100) | Kabupaten Sleman, DIY (100) | Kabupaten Malaka, NTT (100) |
| 6 | Kabupaten Intan Jaya, Papua (100) | Kabupaten Intan Jaya, Papua (100) | Kabupaten Jayawijaya, Papua (100) | Kota Banjarbaru, Kalsel (100) |
| 7 | Kabupaten Sukoharjo, Jateng (99,69) | Kabupaten Purworejo, Jateng (98,14) | Kota Salatiga, Jateng (99,30) | Kota Pekanbaru, Riau (91,72) |
| 8 | Kabupaten Malaka, NTT (98,57) | Kabupaten Sarmi, Papua (97,33) | Kabupaten Pandeglang, Banten (98,52) | Kabupaten Nduga, Papua (82,37) |
| 9 | Kabupaten Batang Hari, Jambi (97,50) | Kabupaten Yalimo, Papua (97,22) | Kabupaten Malaka, NTT (97,44) | Kabupaten Mimika, Papua (68,59) |
| 10 | Kabupaten Mappi, Papua (96,92) | Kabupaten Pidie, Aceh (94,92) | Kabupaten Tuban, Jatim (96,95) | Kabupaten Bolaang Mongondow Utara, Sulut (67,71) |

Sumber: (Bagja et al., 2023)

Walaupun pada tingkat provinsi, mayoritas wilayah berada pada kategori sedang. Namun untuk sebaran kategori tinggi, terjadi di seluruh pulau besar di Indonesia. Sedangkan untuk tingkat kabupaten/kota, mayoritas berasal dari provinsi Papua (Bagja et al., 2023). Lebih lanjut lagi penilaian *Electoral Integtiry Project* menyebutkan Indonesia masih masuk ke dalam kategori moderat sebagai negara yang telah mengadopsi nilai-nilai pemilu yang ber (Garnett et al., 2023). Hal ini mengindikasikan tidak menutup kemungkinan adanya potensi kerawanan pada penyelenggaraan Pemilu 2024.

Kemudian ditambah dengan adanya kemunculan pernyataan, putusan dan kabar di sepanjang tahapan penyelenggaraan Pemilu. Seperti contoh: (a) Pada tahapan awal, (i) Pernyataan Presiden Jokowi yang akan cawe-cawe dalam Pemilu 2024 dan (ii) Putusan Majelis Kehormatan Makamah Konstitusi (MKMK) yang menyatakan Makamah Konstitusi (MK) telah melakukan pelanggaran kode etik dalam menghasilkan produk putusan 90/PUU-XXI/2023, yang mengubah batas usia calon Presiden atau calon Wakil Presiden[[2]](#footnote-3). (b) Pada masa kampanye, isu politisasi bantuan sosial dan netralitas ASN[[3]](#footnote-4). (c) Pada masa tenang, perilisan film *Dirty Vote* yang menjadi *trending topic* di media sosial[[4]](#footnote-5). (d) Pada masa penghitungan suara, kendala yang dialami Sistem Rekapitulasi Informasi (Sirekap)[[5]](#footnote-6),[[6]](#footnote-7). Menurut Titi Anggreini, pengajar hukum pemilu di Universitas Indonesia, kemunculan kabar tersebut dapat berpotensi menciptakan stigma, opini atau persepsi buruk bahwa hasil Pemilu 2024 tidaklah mencerminkan kehendak suara rakyat yang murni (Kompas, 2023).

Menurut Van Ham (2020), ketika terdapat tanda-tanda dugaan pelanggaran dalam penyelenggaraan pemilihan umum, maka dapat berpotensi memimbulkan persepsi telah rusaknya proses translasi dari preferensi menjadi dukungan, dukungan menjadi suara, suara menjadi representasi, representasi menjadi kekuatan politik (Van Ham, 2020). Menurut Norris (2019), stigma dan persepsi buruk dalam penyelenggaraan pemilihan umum pada akhirnya dapat mengikis nilai-nilai demokrasi yang telah lama dibangun dan telah memperoleh legitimasi publik, yang selama ini dapat menjadi penyanggah ketika terjadi turbulensi politik seperti: kasus korupsi, krisis ekonomi, atau bahkan skandal kepemimpinan (Norris, 2019). Maka kehadirannya tidak hanya memicu terjadinya *political distrust,* lebih jauh dari itu, bisa menyebabkan disintegrasi sosial di masyarakat (Mauk, 2022; Norris, 2019). Maka untuk mengantisipasi hal tersebut, dibutuhkan pengukuran untuk dapat mengetahui sejauh mana sentimen positif dan negatif dari opini masyarakat terhadap penyelenggaraan Pemilu 2024.

Dewasa ini media sosial telah menjadi tempat bagi masyarakat untuk menunjukkan opini yang mereka miliki dan mendiskusikannya (Liu, 2015). Termasuk juga opini terkait dengan perkembangan situasi politik, dimana media sosial memfasilitasi warga negara untuk menyalurkan hak demokrasinya seperti mengeskpresikan pandangan politik, berkomunikasi permasalahan politik, dan mengapresiasi atau mengkritik kinerja penyelenggara negara (Saud et al., 2023; Saud & Margono, 2021). Sebagaimana umumnya terjadi di tahun Pemilu, opini dan topik pembicaraan yang paling hangat dibicarakan saat ini adalah opini dan topik pembicaraan terkait dengan penyelenggaraan Pemilu 2024. Opini dan topik pembicaraan yang berkembang tentunya tidak lepas dari pernyataan, putusan dan kabar, yang menurut Titi Anggreini, dapat berpotensi menciptakan sentimen negatif terhadap penyelenggaraan pemilu 2024 (Kompas, 2023).

Salah satu faktor yang menyebabkan banyaknya opini tentang penyelenggaraan pemilu yang disampaikan masyarakat melalui internet dan media sosial tidak terlepas dari normalnya penggunaan internet dan media sosial di Indonesia. Survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) selama periode 10 Januari -27 Januari 2023 menyatakan bahwa penetrasi penggunaan internet di Indonesia pada tahun 2023 meningkat menjadi 215.626.156 pengguna atau mencapai 78.19 persen dari total populasi Indonesia sebanyak 275.773.903 penduduk[[7]](#footnote-8). Sementara pemerataan penggunaan internet di Indonesia juga cukup baik dimana tidak hanya masyarakat yang berasal dari kota besar yang menggunakannya. Tercatat sebanyak 79,79 persen dari masyarakat di daerah pedesaan telah menggunakan internet, bahkan angka tersebut melebihi besaran masyarakat Indonesia di perkotaan yaitu 77,36 persen.

Sementara media sosial mencatatkan angka pengguna yang juga tidak kalah tinggi di Indonesia. Berdasarkan laporan dari *We Are Social*, pengguna media sosial di Indonesia pada tahun 2023 telah mencapai 60,4 persen dari total populasi atau sebanyak 167 juta pengguna[[8]](#footnote-9). Sedangkan waktu rata-rata yang dihabiskan dalam menggunakan media sosial, Indonesia masuk ke dalam 10 besar negara di dunia, dimana sekitar 3 jam 18 menit masyarakat Indonesia menghabiskan waktu mereka untuk menggunakan media sosial. Sementara proyeksi yang dilakukan oleh Statista meyebutkan pada tahun 2026 sebanyak 81,82 persen dari total populasi penduduk Indonesia akan aktif menjadi pengguna media sosial, angka ini meningkat 2 kali lipat dari jumlah pengguna media sosial di Indonesia pada tahun 2017[[9]](#footnote-10).

Sementara itu salah satu media sosial yang paling banyak digunakan dengan tujuan untuk mengkomunikasikan peristiwa yang sedang hangat dibicarakan adalah media sosial Twitter. Di Indonesia sendiri media sosial Twitter memiliki pengguna aktif yang cukup banyak. Berdasarkan laporan dari *We Are Social* pada Januari 2023, Twitter menempatkan posisi ke 6 sebagai media sosial yang paling sering digunakan masyarakat Indonesia, dengan presentase pengguna 8.7 persen dari keseluruhan populasi masyarakat Indonesia atau sebanyak 24 juta pengguna. Jumlahnya meningkat 30.1 persen dari tahun sebelumnya atau sekitar 5.6 juta pengguna baru[[10]](#footnote-11). Oleh karena itu dapat disimpulkan bahwa platform media sosial Twitter dapat dijadikan sumber data dalam mengukur sentimen masyarakat terkait dengan penyelenggaraan Pemilu 2024. Tabel 1.3 menunjukkan beberapa contoh cuitan di media sosial Twitter yang merupakan opini masyarakat terkait dengan penyelenggaraan Pemilu 2024, baik yang bersentimen negatif maupun bersentimen positif.

Tabel 1. 3 Contoh cuitan opini masyarakat terkait dengan penyelenggaraan Pemilu 2024

|  |  |
| --- | --- |
| **Cuitan bersentimen positif** | **Cuitan bersentimen negatif** |
|  |  |
|  |  |

Salah satu teknik yang sering digunakan untuk mengukur opini, sentiment dan persepsi masyarakat yang mereka tuangkan pada media sosial adalah analisis sentimen (Liu, 2015). Beberapa penelitian sebelumnya telah menunjukkan bahwa analisis sentimen telah banyak digunakan untuk mengetahui opini masyarakat terhadap situasi politik dan kinerja penyelenggara negara. Penelitian yang dilakukan oleh Pandu Maulana (2022) menggunakan analisis sentimen untuk mengukur opini masyarakat terhadap upaya pemerintah mengatasi masalah pengangguran selama masa pandemi COVID-19 (Maulana et al., 2022). Sementara Dian Isnaeni (2022) menggunakan analisis sentimen untuk mengukur penilaian masyarakat terhadap implementasi sistem Laporan Harta Kekayaan Penyelenggara Negara (LHKPN) yang diselenggarakan oleh Komisi Pemberantasan Korupsi (KPK) (Afra, 2022).

Berdasarkan penelitian terkini terkait analisis sentimen menunjukkan bahwa *state-of-art* teknik yang digunakan adalah dengan mengadopsi teknik *machine learning* (Drus & Khalid, 2019). *Machine learning* merupakan analisis prediktif, dimana berdasarkan model yang terbentuk akan mengautomasi proses klasifikasi sentimen. Kemampuan automasi dengan menggunakan teknik *machine learning* dalam menganalisis data yang berasal dari media sosial telah menjadi solusi, dimana data yang berasal dari media sosial memiliki 3 karakteristik yang menjadi tantangan dalam menganalisisnya. Pertama, *velocity* dimana data yang berasal dari media sosial memiliki pertumbuhan yang sangat cepat*,* kedua, *volume* dimana data yang berasal dari media sosial memiliki jumlah yang besar*,* ketiga, *variety* dimana data yang berasal dari media sosial memiliki bentuk berupa teks yang tidak terstruktur (Khan, 2018). Sehingga dibutuhkan cara yang dapat memberikan waktu analisis yang cepat untuk mengimbangi waktu pertumbuhan data dan efisiensi untuk dapat memproses data dengan jumlah yang besar dalam menggali pola tersembunyi dari data tidak terstruktur. Sedangkan *state-of-art* algoritma pembentukan model *machine learning* seperti *artificial neural network* biasanya dikomparasikan dengan algoritma *machine learning* yang lebih klasik seperti *support vector machine* (SVM), *random forest,* dan *logistic regression* untuk dapat mengidentifikasi algoritma yang dapat memberikan hasil prediksi paling akurat (Jain et al., 2021).

Berdasarkan uraian diatas maka dapat disimpulkan bahwa analisis sentimen dengan menggunakan pendekatan *machine learning* dapat digunakan untuk mengukur seberapa besar pola sentimen positif dan negatif terhadap penyelenggaraan Pemilu 2024 berdasarkan opini masyarakat yang mereka sampaikan melalui media sosial Twitter. Kemudian dengan mengadopsi teknik pemodelan topik diharapkan dapat menyaring tema-tema aspirasi yang dikemukakan masyarakat baik yang berasal dari opini bersentimen positif dan maupun negatif. Hasil dari penelitian ini, yang berupa analisis deskriptif, diharapkan dapat memberikan manfaat bagi para pemangku kepentingan seperti Penyelenggara Pemilu, Pemerintah secara umum, Konsultan Politik dan juga Peserta Pemilu.

## Rumusan Masalah

Analisis kesenjangan dilakukan berdasarkan hasil studi literatur dan analisis data skunder. Analisis kesenjangan yang dilakukan berupaya untuk menemukan permasalah yang akan dijadikan fokus dalam penelitian ini. Berdasarkan adanya *gap* antara harapan dan kenyataan, dimana diharapkan Pemilu 2024 dapat berjalan dengan luber jurdil, namun kenyataannya terdapat dugaan terciptanya stigma buruk di tengah masyarakat terkait dengan penyelenggaraannya. Maka dapat disimpulkan dibutuhkan pengukuran untuk dapat mengetahui sejauh mana opini masyarakat terkait dengan penyelenggaraan Pemilu 2024. Hasil dari pengukuran yang dilakukan dapat dijadikan bahan pertimbangan bagi pemangku kepentingan untuk dapat menjaga stabilitas sosial pasca Pemilu 2024. Hal ini mendorong peneliti untuk melakukan penelitian terhadap opini atau sentimen masyarakat terkait penyelenggaraan Pemilu 2024.

Tabel 1. 4 Harapan, Kenyataan, Masalah, dan Urgensi

|  |  |  |
| --- | --- | --- |
| **Harapan** | **Kenyataan** | **Permasalahan** |
| Terselenggaranya pemilihan umum yang luber jurdil sesuai dengan apa yang termaktub dalam UUD 1945 dan Undang-Undang No 7 Tahun 2017 tentang Pemilu (DPR RI, 2017). | Adanya dugaan terciptanya stigma atau persepsi buruk bahwa penyelenggaraan Pemilu 2024 tidaklah berjalan dengan luber jurdil. Berikut adalah indikasinya:   1. Berdasarkan survey BPS sebelum penyelenggaraan Pemilu 2024 dilakukan, masih rendahnya tingkat kepercayaan masyarakat Indonesia terhadap proses penyelengaraan Pemilu (DataIndonesia.id, 2023). 2. Berdasarkan data historis adanya pelanggaran di dalam pemilu: 3. Ada pelanggaran netralitas ASN sebanyak 854 kasus yang terjadi selama Pilkada 2020 (Sirait et al., 2022). 4. Indeks Kerawanan Pemilu (IKP) menunjukkan selama tahun 2017-2020 masih banyak temuan kerawanan pada Dimensi Penyelenggaraan Pemilu (Bagja et al., 2023). 5. Adanya pernyataan, putusan dan kabar di sepanjang tahapan Pemilu yang menurut Titi Anggreini, dapat berpotensi menciptakan stigma dan persepsi buruk bahwa hasil Pemilu 2024 tidaklah mencerminkan kehendak suara rakyat yang murni (Kompas, 2023). | Dibutuhkan pengukuran untuk dapat mengetahui sejauh mana sentimen positif dan negatif dari opini masyarakat terhadap penyelenggaraan Pemilu 2024. |
| **Urgensi:**  Pengukuran yang dilakukan dapat menjadi langkah antisipasi dan bahan pertimbangan bagi pemangku kepentingan dalam menjaga stabilitas sosial pasca Pemilu 2024. Dimana dikhawaritkan dampak negatif dari stigma atau persepsi buruk dalam Pemilu dapat menciptakan *political distrust* terhadap pemerintahan yang baru terpilih atau bahkan disintegrasi sosial di tengah masyarakat (Mauk, 2022; Norris, 2019). | | |

## Pertanyaan Penelitian

Berdasarkan pemaparan pada bagian latar belakang dan rumusan masalah, maka penelitian ini mengajukan pertanyaan penelitian sebagai berikut:

RQ1: Bagaimana model klasifikasi sentimen yang dapat mengklasifikasikan secara otomatis opini masyarakat berupa cuitan di media sosial Twitter terkait dengan penyelenggaraan Pemilu 2024?

RQ2: Sejauh mana opini bersentimen positif dan negatif yang masyarakat sampaikan di media sosial Twitter terkait dengan penyelenggaraan Pemilu 2024?

RQ3: Seberapa efektif pemodelan topik dapat menyaring tema-tema aspirasi masyarakat terkait penyelenggaraan Pemilu 2024, baik yang berasal dari opini bersentimen positif maupun negatif?

## Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Menghasilkan model analisis sentimen yang dapat mengukur opini atau persepsi masyarakat terkait dengan penyelenggaraan Pemilu 2024 di Indonesia.
2. Mengetahui pola sentimen masyarakat terkait dengan penyelenggaraan Pemilu 2024 di Indonesia
3. Mengidentifikasi topik pembicaraan atau aspirasi masyarakat terkait dengan penyelenggaraan pemilihan umum 2024.

## Ruang Lingkup Penelitian

Batasan pada penelitian ini adalah sebagai berikut:

1. Pada Penelitian ini akan hanya menggunakan data pada media sosial Twitter.
2. Data yang diambil merupakan data *tweet* yang berupa komentar masyarakat terhadap Penyelenggaraan Pemilu 2024.
3. Data *tweet* yang dikumpulkan akan dikempokkan ke dalam 2 jenis klasifinasi. Jenis klasifikasi pertama adalah jenis klasifikasi untuk data relevan dan tidak relevan. Kemudian untuk jenis klasifikasi kedua adalah klasifikasi sentimen apakah opini bersentimen positif atau negatif. Lalu dilakukan identifikasi topik atau tema-tema aspirasi masyarakat yang berasal dari dataset bersentimen positif dan negatif.
4. Rentang waktu pengambilan data dilakukan ketika masa kampanye, masa tenang, dan masa pemungutan dan rekapitulasi suara yang berlangsung pada periode 21-Januari-2024 sampai dengan 20-Maret-2024.
5. Data Tweet yang digunakan pada penelitian ini merupakan *tweet* berbahasa Indonesia.
6. Penelitian ini menggunakan instrumen seperti *google colab* dan *jupyter notebook,* dimana bahasa pemograman *python* akan digunakan untuk mengeksekusi proses *text mining* yang akan dilakukan.

## Manfaat Penelitian

Berikut adalah manfaat yang dapat diberikan oleh penelitian ini kepada para pihak seperti akademisi dan praktisi yang berkepentingan:

1. Manfaat Akademis

Penelitian terkait mengembangan model untuk menggali sentimen, opini dan aspirasi masyarakat terhadap penyelenggaraan Pemilu 2024 di Indonesia yang: (1) dapat dijadikan referensi untuk penelitian selanjutnya dan (2) data yang dikumpulkan dari penelitian ini dapat digunakan untuk penelitian selanjutnya.

1. Manfaat Praktis

Penelitian ini diharapkan bisa bermanfaat bagi para praktisi seperti:

1. Penyelenggara Pemilu.

Wawasan yang dihasilkan dalam penelitian ini diharapkan dapat memberikan manfaat bagi 3 lembaga Penyelenggara Pemilu, yaitu (a) Komisi Pemilihan Umum (KPU) yang bertanggung jawab menjalankan tahapan Pemilu, (b) Badan Pengawas Pemilu (Bawaslu) yang bertanggung jawab untuk melakukan pengawasan terhadap jalannya penyelenggaraan Pemilu, dan (d) Dewan Kehormatan Penyelenggara Pemilu (DKPP) yang bertugas mengawasi fungsi KPU dan Bawaslu dan potensi pelanggaraan kode etik yang dilakukan oleh kedua lembaga tersebut. Hasil penelitian diharapkan dapat memberikan saran perbaikan terhadap ketiga lembaga dalam menyongsong kontestasi yang akan datang, seperti Pemilihan Gubernur DKI Jakarta 2024. (tergantung hasil penelitian, disertakan jika topik banyak membahas penyelenggara pemilu)

1. Pemerintah secara umum.

Hasil penelitian ini dihadapkan bisa memberikan informasi tambahan bagi Pemerintah dalam upaya menjaga stabilitas sosial pasca Pemilu 2024, seperti mengantisipasi dari adanya potensi polarisasi atau disintegrasi sosial di tengah masyarakat.

1. Peserta Pemilu dan Konsultan Politik.

Hasil penelitian ini diharapkan bisa memberikan wawasan tambahan bagi peserta Pemilu baik partai politik atau kontestan calon independen dan Konsultan Politik. Hasil penelitian diharapkan dapat menjadi bahan evaluasi apakah ada strategi, pendekatan, atau aksi kegiatan kampanye yang telah dilakukan telah dikaitkan dengan adanya opini pelanggaran dalam Pemilu yang pada akhirnya dapat menggerus elektabilitas Peserta Pemilu.

## Sistematika Penulisan

Sistematika penulisan pada penelitian ini disusun ke dalam enam bab dengan sistematika penulisan sebagai berikut:

* BAB 1 PENDAHULUAN

Pada bagian ini penulis menguraikan latar belakang penelitian. Bab ini membahas tentang bagaimana penulis (1) menguraikan harapan dan kenyataan terkait topik penelitian, (2) mengidentifikasi masalah, (3) merumuskan pertanyaan, tujuan, ruang lingkup dan manfaat penelitian.

* BAB 2 TINJAUAN PUSTAKA

Pada bagian ini penulis membahas tentang beberapa teori, metode yang menjadi rujukan pada penelitian ini. Kemudian dilanjutkan dengan memaparkan 5 penelitian terdahulu dengan topik Analisis sentimen dalam bingkai analisis 3C2S. Bagian ini diakhiri dengan menjelaskan kerangka teoritis yang digunakan untuk penelitian ini.

* BAB 3 METODOLOGI PENELITIAN

Pada bagian ini penulis memjelaskan rancangan penelitian, seperti menjelaskan metode pengumpulan dan pengelolahan data sampai dengan pendekatan dalam penarikan kesimpulan. Kemudian dilanjutkan dengan menjelaskan alur penelitian yang terdiri 11 tahapan yang akan dilalui pada penelitian ini.

* BAB 4 IMPLEMENTASI

Pada bagian ini penulis menunjukan langkah-langkah implementasi dari apa yang telah dijelaskan pada bagian metodologi penelitian dan alur penelitian.

* BAB 5 HASIL DAN PEMBAHASAN

Pada bagian ini penulis memaparkan hasil penelitian yang berupa analisis deskriptif dari analisis sentimen dan pemodelan topik yang disajikan berdasarkan tahapan penyelenggaraan Pemilu 2024, yaitu: (a) masa kampanye, (b) masa tenang, (c) masa perhitungan dan pemungutan suara, (d) masa ajudikasi sampai dengan penetapan pemenang Pemilu.

* BAB 6 PENUTUP

Pada bagian ini membahas tentang kesimpulan berdasarkan temuan-temuan pada penelitian ini dan saran yang dapat menjadi rujukan untuk penelitian selanjutnya

# 

# BAB 2 STUDI LITERATUR

Pada bab ini akan dibahas mengenai teori,teknik, literatur yang akan digunakan pada penelitian ini. Selain itu juga akan dibahas beberapa penelitian Analisis Sentimen terdahulu, yang disajikan dalam bentuk analisis *Compare, Contrast, Critize, Synthesize, Summarize* (3CS2). Pada akhir bagian akan ditampilkan kerangka teoritis yang digunakan pada penelitian ini.

## Media Sosial

Menurut Kaplan dan Haenlein (2010) media sosial adalah suatu istilah yang terlahir dari perpaduan 2 konsep yaitu *web 2.0* dan *user-generated content* (Haenlein & Kaplan, 2010). Untuk konsep yang pertama, *web 2.0* merupakan istilah dari cara baru dalam pemanfaatan internet sebagai sarana media dalam bentuk platform. Konsep ini menjadi tempat dimana konten atau informasi tidak diciptakan oleh satu entitas yang memiliki kendali akan media tersebut (*web 1.0*). *Web 2.0* memungkinkan semua pengguna berpartisipasi dan berkolaborasi dalam membuat, mengubah, dan menjaga isi informasi dan berita yang disiarkan melalui *platform* tersebut. Sedangkan *user-generated content* merupakan konten media yang dibuat dan dibagikan oleh pengguna akhir dalam bentuk yang bervariasi seperti teks, gambar, audio dan video yang dapat diakses secara publik.

Lebih lanjut lagi *Organization for Economic Cooperation and Development* (OECD) menjelaskan bahwa ada 3 komponen penting sehingga suatu informasi atau konten bisa dikatakan sebagai *user-generated-content*[[11]](#footnote-12)*.* (1) Konten atau Informasi haruslah dapat diakses secara publik pada tempat seperti website atau situs *social networking* yang ditujukan kepada suatu kelompok. (2) Adanya upaya kreatif dalam lahirnya konten dan informasi (3) Konten dan informasi harus dibuat di luar konteks rutinitas profesional. Dari ketiga kriteria tersebut kita dapat memisahkan beberapa jenis konten dan informasi yang tidak masuk ke dalam kategori *user-generated content* seperti: (1) email dan *instant messager* yang tidak lolos dalam kriteria pertama. (2) Memposting berita tanpa memodifikasinya dan menyebarkannya ke internet juga bukan masuk ke dalam *user generated content* berdasarkan kriteria kedua dan (3) Semua konten yang mengandung tujuan komersial tidak masuk kriteria ketiga. Dengan demikian dapat disimpulkan bahwa media sosial merupakan aplikasi berbasis internet yang dibangun berlandaskan fondasi ideologi dan teknologi dari *web 2.0* dan memungkinkan penggunanya untuk membuat dan membagikan *user-generated content.*

## 2.2 Analisa Media Sosial

Menurut Khan (2015), Analisa media sosial adalah perpaduan dari seni dan ilmu pengetahuan dalam menggali wawasan yang tersimpan di dalam data, baik yang terstruktur maupun yang tidak terstruktur, yang terdapat pada media sosial dan dipergunakan untuk membantu proses pengambilan keputusan (Khan, 2018). Terdapat 8 *layers* dalam pendekatan menganalisa media sosial. Setiap *layer* memiliki kegunaan dan karakteristik khusus, berikut adalah penjelasannya:

1. *Text Layer*

Analisa media sosial yang dilakukan terhadap konten berbasis teks seperti komentar, *tweets,* status di Facebook. Terdapat berbagai macam jenis pengaplikasian dari analisa teks media sosial, seperti: penggalian *sentiment mining, intention mining,* dan *topic & idea mining*.

1. *Network Layer*

Analisa media sosial yang dilakukan terhadap jejaring relasi antar entitas di media sosial, seperti jejaring relasi pertemanan, jejaring relasi komunitas yang diikuti pengguna. Analisa *network layer* memungkinkan peneliti dan praktisi untuk dapat mengidentifikasi entitas yang berpengaruh, mengidentifikasi struktur jaringan pada media sosial. Kemampuan tersebut dapat bermanfaat dalam mengidentifikasi efisiensi dalam transmisi informasi pada media sosial.

1. *Actions Layer*

Analisa media sosial yang dilakukan terhadap aktifitas pengguna pada media sosial, seperti aktifitas *likes*, *dislikes*, *mentions* yang dilakukan oleh atau tertuju kepada pengguna. Analisa *actions layers* memungkinkan peneliti dan praktisi untuk menganalisa performa kampanye, mengukur popularitas *brand*, atau *engagement tracking* di media sosial*.*

1. *Application Layer*

Analisa pada lapisan ini dilakukan dengan tujuan untuk mengukur dan mengoptimalkan kepuasan pengguna terhadap penggunaan aplikasi media sosial.

1. *Hyperlinks Layer*

Analisa *hyperlinks layer* dilakukan dengan menganalisa dan menginterpretasikan keberadaan *hyperlink* pada media sosial, baik *in-links* maupun *out-links*[[12]](#footnote-13).

1. *Locations Layer*

Analisa *locations layers* juga dikenal sebagai *spatial analysis* atau *geospatial analytics,* digunakan untuk menggali dan memetakan informasi pada media sosial berdasarkan area geografis. Contoh penerapannya adalah analisa dukungan terhadap calon presiden berdasarkan demografi lokasi.

1. *Search Engines Layer*

Analisa *search engine layer* dilakukan terhadap histori kata kunci yang digunakan untuk proses pencarian pada mesin pencari. Analisa *search engine layer* digunakan untuk menganalisa *trend*, memantau kata kunci yang sering digunakan dan hasil yang ditemukan berdasarkan kata kunci yang digunakan.

1. *Multimedia Layer*

Analisa *multimedia layer* dilakukan dengan tujuan untuk dapat menggali *business value* pada media sosial dari data berformat *video,* gambar, *audio* dan bahkan animasi. Oleh karena itu dibutuhkan teknik khusus yang berbeda dengan analisa media sosial untuk *text layer*.

Sedangkan pada perspektif tujuan dari analisa media sosial, Khan (2018) membaginya ke dalam 4 kategori, yaitu:

1. *Descriptive Analytics*

Tujuan dari *descriptive analytics* adalah untuk menjabarkan secara mendetail data dari suatu fenomena yang sedang atau telah terjadi. Contoh dari *descriptive analytics* adalah aplikasi *business intellegance* dimana penggunaan fungsi seperti operasi aritmatika, *mean, median, max, percentage* dan sebagainya digunakan terhadap data eksisting.

1. *Diagnostic Analytics*

Tujuan dari *diagnostic analytics* adalah untuk mengatahui alasan atau faktor pendorong penyebab terjadinya suatu fenomena. Teknik yang biasa digunakan adalah *drill-down, data discovery, data mining* dan *correlation.*

1. *Predictive Analytics*

Tujuan dari *predictive analytics* adalah untuk memprediksi kemungkinan suatu fenomena akan terjadi. Teknik yang biasa digunakan adalah dengan menciptakan model *machine learning*. Model yang terbentuk akan divalidasi dengan mengukur tingkat akurasi dari prediksi yang dihasilkan. Model dikatakan layak digunakan ketika menunjukkan hasil di atas ambang batas tingkat akurasi yang diharapkan.

1. *Prescriptive Analytics*

Tujuan dari *prescriptive analytics* adalah untuk dapat memberikan solusi terbaik berdasarkan 3 analisa sebelumnya *descriptive analytics, diagnostic analytics,* dan *predictive analytics*. Contoh dari penerapannya adalah *recommender system, decision modeling* dan *expert system.*

Analisis media sosial pada penelitian ini dilakukan pada bagian *text layer* dimana data yang akan dianalisa berupa *tweet* yang bertipe teks. Sedangkan dalam perspektif tujuan analisis media sosial pada penelitian ini dapat dikategorikan sebagai *descriptive analitycs* dan *predictive analytics.* Dikategorikan *descriptive analytics* dikarenakan hasil penelitian diharapkan dapat menjelaskan secara mendetail opini dan aspirasi masyarakat terhadap peristiwa yang telah terjadi yaitu penyelenggaraan Pemilu 2024. Dikategorikan sebagai *predictive analycis* dikarenakan penelitian ini mengadopsi teknik analisis prediktif yaitu *machine learning*.

## Analisis Sentimen

Analisis sentimen merupakan studi terkait analisa opini, sentimen, sikap, dan emosi yang diekspresikan dalam bentuk teks terhadap suatu entitas atau atributnya (Liu, 2015). Proses analisa menekankan pada identifikasi sentimen positif atau negatif dari suatu opini. Tidak hanya menjadi topik penelitian pada domain komputer, analisis sentimen juga menjadi topik penelitian dari domain yang berhubungan dengan busnis atau opini publik, seperti manajemen, politik, sosial dan ekonomi. Namun berbeda dengan penelitian pada domain bisnis atau opini publik, dimana yang fokus penelitiannya adalah untuk mengukur dampak *business value* dari opini yang berkembang, pada domain *computer science,* yang menjadi fokus adalah teknik yang digunakan adalah dengan melibatkan proses komputasi dengan memfaatkan teknologi komputer.

Contoh penerapan analisis sentimen sangat beragam, bagi organisasi analisis sentimen bisa mengantikan cara lama dalam mengukur opini publik seperti sentimen masyarakat terhadap calon presiden, kebijakan publik yang diambil pemerintah, analisa dalam memprediksi pasar saham, sampai dengan analisa sentimen masyarakat terhadap produk dan jasa yang dipasarkan. Pada awalnya perhatiaan terhadap topik penelitian analisis sentimen kurang begitu dilirik, namun munculnya media sosial pada awal tahun 2000an menjadi katalisator yang membawa antusias terhadap penelitian analisis sentimen. Hal ini dikarenakan media sosial dapat menciptakan data opini dengan pertumbuhan yang cepat dan dapat diakses oleh siapapun. Dilihat dari perspektif tingkat analisis, analisis sentimen dapat dibagi menjadi 3 tingkat, yaitu: *document level analysis, sentiment level analysis* dan *entity / aspect level analysis* (Liu, 2015)*.*

1. *Document level Analysis.*

Tujuannya adalah untuk menidentifikasi sentimen positif atau negatif dari keseluruhan teks pada dokumen. Analisis pada tingkat ini dilakukan dengan asumsi bahwa isi dokumen hanya tertuju ke satu entitas sebagai target. Untuk dapat mengevaluasi sentimen terhadap banyak entitas dibutuhkan tingkat analisis yang lebih spesifik yaitu *sentence level analysis* atau *aspect level analysis.*

1. *Sentence level Analysis.*

Tujuannya adalah untuk mengidentifikasi sentimen positif atau negatif dari setiap kalimat dari suatu paragraf atau dokumen. Pada analisis tingkat ini setiap kalimat hanya memiliki satu entitas yang menjadi target sentimen.

1. *Aspect level Analysis*

Tujuannya adalah untuk mengidentifikasi sentimen positif atau negatif dari setiap aspek atau entitas yang terdapat pada suatu kalimat. Perbedaan paling mendasar dari *sentence-level-analysis* dan *aspect-level analysis* adalah *sentence-level-analysis* hanya dapat mengidentifikasi satu entitas sebagai target sentimen dalam satu kalimat, sedangkan *aspect-level analysis* memungkinkan lebih dari satu entitas sebagai target dalam suatu kalimat. Sebagai contoh “IPhone 15 memiliki spesifikasi yang sangat tinggi, namun harganya sangat mahal”. Dimana terdapat 2 aspek yaitu “spesifikasi” dan “harga” yang menjadi target sentimen.

Penelitian ini menggunakan teori analisis sentimen untuk mengetahui opini masyarakat terkait penyelenggaraan Pemilu 2024 di media sosial. Sedangkan berdasarkan jenis tingkat analisis penelitian ini dapat dikategorikan menggunakan jenis *document-level analysis* dimana satu *tweet* dapat dianalogikan sebagai satu dokumen yang memiliki satu jenis sentimen.

## Text Mining

*Text mining* atau *text analytics* membahas sekumpulan teknik linguistik, statistik dan *machine learning* yang digunakan untuk membuat model dan struktur yang berguna untuk menganalisa, mengambil informasi inti dan menemukan setiap pola yang menarik seperti tren dan *outlier* yang berasal dari data yang berlimpah dan berdimensi tinggi, yaitu teks (Aggarwal & Zhai, 2013). Hasil dari *text analytics* pada akhirnya digunakan untuk membantu dalam proses pengambilan keputusan. Setelah melakukan Pembentukan Korpus*,* menurut Zhai (2013) terdapat 3 fase dalam proses *text analytics*, yaitu *text preprocessing*, *text representation*, dan *knowledge discovery.* Gambar 2.1 menunjukkan urutan kerangka kerja *text mining.* Berikut penjelasan lebih detail terkait dengan ke 4 fase tersebut:

A diagram of a diagram

Description automatically generated with medium confidence

Gambar 2. 1 Alurkerangka kerja *text mining.*

Sumber: (Aggarwal & Zhai, 2013)

### *Text Corpus Building*

Pada tahapan ini meliputi pengumpulan data*,* pelabelan data, dan evaluasi hasil pelabelan. Hasil dari tahapan ini adalah dataset dengan nilai *dependent variable* atau atribut kelas yang telah terdefinisikan. Berikut adalah penjelasan dari ketiga subproses dari tahapan ini:

#### *2.5.1.1 Data Collection*

*Data Collection* atau akuisisi data adalah tahap awal dalam setiap studi dan pengembangan *data science* (Khder, 2021). Pada tahapan ini data diperoleh dari berbagai sumber, baik sumber pribadi seperti catatan penjualan perusahaan dan laporan keuangan, atau dari sumber publik seperti jurnal, situs web, atau bahkan dengan cara membelinya. Pada tahapan ini data yang terkumpul belum memiliki nilai *dependent variable* atau atribut kelas. Sementara itu ada 2 cara yang sering digunakan untuk mengumpulkan data mentah dari media sosial. (1) menggunakan *Application Programming Interface* (API) yang disediakan oleh situs media sosial atau (2) menggambil informasi menggunakan proses *web scraping* (Zafarani et al., 2014).

*Web scraping* yang juga dikenal sebagai ekstraksi *web*, adalah teknik untuk mengekstrak data dari halaman web dan menyimpannya ke dalam sistem file atau database untuk analisis lebih lanjut di kemudian hari (Khder, 2021). Sementara menurut Sirisuriya (2015) *web scraping* adalah teknik untuk mengubah data web yang tidak terstruktur menjadi data yang terstruktur yang dapat disimpan dan dianalisis dalam basis data pusat atau lembar kerja *spreadsheet*. Berikut cara kerja *web scrapper* (Laurie A. Schintler; Connie L. McNeely, 2022):

1. Mengkonstruk HTTP *request* untuk mengakuisisi data dari website yang dituju. HTTP *request* yang digunakan dapat berisikan sebuah URL ditambah dengan *query* metode GET atau POST yang dibutuhkan.
2. Mengirimkan HTTP *request* ke website yang dituju.
3. Website yang dituju akan mengirimkan jawaban balik dalam bentuk HTTP *response*. Pada HTTP *response* tersebut juga membawa data yang dibungkus dengan format seperti HTML, JSON atau XML.
4. *Web Scrapper* akan mengekstrak data inti yang dibutuhkan dari format HTML, JSON, atau XML dan menyimpannya ke dalam format yang lebih standar seperti file *spreadsheet*.

Namun terdapat kontroversi dalam penggunaan *web scrapping* seperti: i) Pengambilan data dapat terjadi pada data berhak cipta dan dapat mengakibatkan pelanggaran hak cipta. ii) *Web scrapping* yang dilakukan secara agresif dengan jumlah permintaan besar dapat membebankan sistem yang dituju, yang tentunya tidak diinginkan oleh pemilik situs. Hal tersebut dapat dianggap sebagai bagian dari *denial-of-service* (DOS) *attack* (Khder, 2021)*.* Terlepas kontroversi yang ada, penggunakan *web scraping* sangat disarankan karena lebih ekonomis dibandingkan dengan metode pengumpulan data lainnya, asalkan dilakukan dengan cara yang beretika dan tidak merugikan pihak lain (Khder, 2021).

#### *2.5.1.2* Evaluasi Hasil Pelabelan

Dalam tahapan pembentukan korpus, setelahproses pengumpulan data, terdapat upaya untuk mengdefinisikan nilai *dependent variable.* Nilai *dependent variable* dapat ditentukan dengan memberikan penilaian atau pelabelan secara manual. Dalam proses penilaian atau pelabelan secara manual tidak terlepas dari tantangan adanya penilaian yang bersifat tidak objektif. Oleh karena itu dibutuhkan uji reliabilitas terhadap proses penilaian dan pelabelan *dependent variable* yang dilakukan. Salah satu indeks yang sering digunakan untuk tujuan objektifitas ini adalah *Cohen Kappa Coefficient* (Tang et al., 2015). Berikut formula dalam menentukan nilai indeks cohen kappa:



Dimana *K* adalah nilai *cohen kappa. Po* adalah persentase penilaian atau pelabelan dengan pilihan yang sama. *Pe* adalah persentase penilaian atau pelabelan dengan pilihan yang berbeda. Cohen Kappa dapat digunakan untuk mengukur seberapa tinggi tingkat kesepakatan penilaian kategorikal antar 2 orang annotator (Tang et al., 2015). Tabel 2.5 menunjukkan interpretasi tingkat kesepakatan berdasarkan nilai indeks cohen kappa.

Tabel 2. 1 Tingkat Nilai Cohen Kappa

|  |  |
| --- | --- |
| **Nilai Cohen Kappa** | **Keterangan** |
| < 0 | Sangat Kurang |
| 0.01 – 0.2 | Kurang |
| 0.21 – 0.4 | Cukup |
| 0.41 – 0.6 | Sedang |
| 0.61 – 0.8 | Baik |
| 0.81 – 0.99 | Sempurna |

### *Text Preprocessing*

Merupakan proses membersihkan dan mempersiapkan teks yang akan dianalisa sehingga lebih mudah diolah dan diinterpretasikan. Subproses dari *text preprocessing* meliputi tokenisasi, penghapusan *stopword*, lematisasi, *stemming* dan penyaringan teks. Hal ini dibutuhkan karena pada proses selanjutnya, *text representation*, data yang diberikan sebagai input haruslah konsisten sehingga meminimalisir kesalahan interpretasi (Aggarwal & Zhai, 2013). Seperti contoh normalisasi teks "senangggg" menjadi "senang" membuat representasi teks menjadi konsisten. Berikut adalah penjelasan tentang sebagian subproses pada tahapan *text preprocessing*.

#### 2.5.2.1 *Tokenisasi*

Tokenisasi adalah bagian dari text preprocessing yang mengubah teks atau *string* menjadi unit yang lebih kecil yang disebut token. Terdapat 2 unit token, yaitu “kalimat” yang merupakan token dari paragraf dan “kata” yang merupakan token dari kalimat (Perkins, 2014). Unit kalimat dihasilkan dari proses *sentence tokenization*, dimana dari proses tersebut mengubah teks menjadi daftar atau kumpulan kalimat. Sedangkan unit kata dihasilkan dari proses *word tokenization*, dimana proses tersebut mengubah teks menjadi daftar atau kumpulan kata. Kata adalah *unit* terkecil dari data teks yang dapat diproses oleh mesin (Hardeniya et al., 2016).

#### 2.5.2.2 Penghapusan *Stopword*

*Stopword* merupakan daftar kata yang tidak memberikan arti secara signifikan pada teks yang dianalisis dengan tujuan penggalian wawasan dan klasifikasi dokumen (Perkins, 2014). Sehingga "Penghapusan *stopword*" merupakan proses menghapus kata dalam teks yang tidak memiliki signifikansi dalam proses penggalian wawasan dan klasifikasi dokumen. Hal ini dapat menghemat proses komputasi dan kebutuhan kapasitas penyimpanan data. Dalam bahasa inggris jenis kata seperti *pronouns* (*i, she, he, they*, dsb) dan *article* (*a, the*, dsb) dapat dikategorikan sebagai *stopword* (Hardeniya et al., 2016). Dalam praktiknya daftar dari *stopword* biasanya dikumpulkan dan dirawat dalam suatu *repository* yang bisa diakses melalui internet[[13]](#footnote-14),[[14]](#footnote-15).

#### 2.5.2.3 *Stemming*

*Stemming*, secara harfiah, adalah proses memotong cabang-cabang pohon hingga hanya tersisa batangnya saja (Hardeniya et al., 2016). Sedangkan dalam konteks *text mining*, *stemming* adalah teknik untuk menghilangkan imbuhan dari sebuah kata, sehingga hanya tersisa kata dasarnya. Sebagai contoh, kata dasar dari "bermain" adalah "main", *stemmer* menghilangkan imbuhan "ber-" pada kata "bermain". Dengan *stemming* secara signifikan mengurangi kebutuhan *space*, *memory* dan meningkatkan tingkat akurasi proses klasifikasi (Perkins, 2014). Penggunaan *stemming* biasanya dihindari atau dilakukan setelah proses penyaringan jenis kata (*part-of-speech*) dan *Named Entity Recognition* (NER) (Hardeniya et al., 2016). Hal ini dikarenakan *stemming* akan mengubah hasil identifikasi NER dan jenis kata dari token yang berubah.

#### 2.5.2.4 Lematisasi

Keluaran atau hasil dari lematisasi mirip dengan stemming, namun secara fungsi jika *stemming* menghilangkan imbuhan pada kata, lematisasi mengganti kata dengan sinonimnya (Perkins, 2014). Lematisasi adalah cara yang lebih sistematis untuk mengubah variasi kata ke akar katanya. Cara yang lebih sistematis ini menggunakan analisis morfologi (asal usul kata), konteks dan aturan normalisasi yang berbeda untuk setiap jenis kata (*part-of-speech*) (Hardeniya et al., 2016). Contoh hasil fungsi lematisasi dengan input berupa variasi kata dalam bahasa inggris "ate" akan ditransformasi menjadi kata "eat". Hal tersebut tidak bisa dilakukan dengan menggunakan *stemming* yang hanya menghapus imbuhan pada kata. Karena lebih sederhana, penggunaan *stemming* sebagai pengganti lematisasi lebih sering dilakukan, namun beberapa kasus NLP mengharuskan penggunaan lematisasi ketimbang penggunaan *stemming.*

#### 2.5.2.5 Normalisasi

Normalisasi adalah upaya yang dilakukan untuk mentransformasi kata informal, yang menimbulkan kebisingan, ke dalam bentuk versi kata yang lebih formal atau resmi (Mosquera et al., 2017). Hal ini dibutuhkan dikarenakan opini masyarakat yang berasal dari media sosial bisa saja mengandung berbagai variasi leksikal, seperti: kata informal atau *slang*. Selain itu penggunaan karakter yang berulang, dengan tujuan untuk memberikan penekanan terhadap pesan yang disampaikan, sering kali dilakukan dalam mengekspresikan suatu opini (Perkins, 2014). Penggunaan kamus normalisasi leksikal sering kali diadopsi dalam menjalankan proses ini (Mosquera et al., 2017). Tabel x menunjukkan perbedaan masukan dan keluaran dari proses transformasi normalisasi teks.

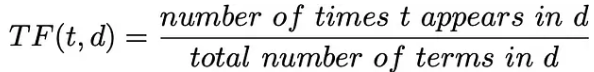
Tabel 2. 2Contoh transformasi normalisasi teks

|  |  |
| --- | --- |
| **Masukan** | **Keluaran** |
| Curaaaaaang | Curang |
| currrrrrrrrraaaaang |
| Tdk | Tidak |
| jngn | Jangan |

### *Text Representation*

*Text representation* merupakan proses untuk mengubah daftar token atau kata yang dihasilkan pada fase *text preprocessing* ke dalam bentuk vektor numerik yang nantinya akan diolah baik mengunakan metode probabilistik, *deep learning*, atau liniear aljabar pada fase *Knowledge Discovery* (Aggarwal & Zhai, 2013). Vektor numerik yang dihasilkan juga dapat disebut sebagai model *Bag of Word* (BOW) atau *Vector Space Model* (VSM), dimana dalam bentuk tersebut struktur linguistik pada teks telah diabaikan. Model direpresentasikan dalam bentuk matrik dengan dimensi n \* m, dimana n adalah jumlah kata dalam korpus (semua dokumen), dikalikan dengan m yang mana adalah jumlah dokumen. Setiap entri dari matrik memiliki bobot yang nilainya akan digunakan dalam penentuan pada tahapan *knowledge discovery* (Aggarwal & Zhai, 2013).

Terdapat beberapa teknik yang dapat digunakan dalam pembobotan nilai, diantaranya adalah *Term-frequency Inverse document frequency* (TF-IDF), *count vectorizer, boolean vectorizer,* dan *word embedding* (Aggarwal & Zhai, 2013)*.* Teknik vektorisasi numerik yang paling sering digunakan adalah TF-IDF, teknik ini memberikan bobot berdasarkan hasil perkalian dari 2 jenis frekuensi, yaitu *Inverse Document Frequency* (IDF) dan *Term Frequency* (TF). Nilai IDF akan lebih tinggi ketika suatu token memiliki frekuensi kemunculan yang jarang di seluruh dokumen atau korpus. Sementara nilai TF akan lebih tinggi ketika token memiliki frekuensi yang tinggi di dalam dokumen atau kalimat dimana token tersebut digunakan (Perkins, 2014). Berikut adalah formula dari fungsi TF-IDF.



A black and white image of a symbol

Description automatically generated



Dimana:

* d adalah *instance* atau salah satu dokumen yang dianalisa
* t adalah suatu token
* N adalah jumlah dokumen atau *instance* pada korpus
* df adalah jumlah dokumen yang di dalamnya terdapat “t” (token)

Sehingga nilai kata atau entri yang memiliki nilai TF-IDF yang lebih tinggi dapat digunakan sebagai atribute atau fitur yang secara signifikan dapat representasi kelas atau jenis dokumen (Zafarani et al., 2014).

### *Knowledge Discovery*

Pada tahapan ini dilakukan analisa terhadap vektor numerik yang terbentuk dengan menggunakan metode *data mining* atau *machine learning*. Hal ini dilakukan untuk dapat menemukan pola yang menarik dan memiliki nilai bisnis (Aggarwal & Zhai, 2013). *Machine learning* juga disebut sebagai analisis prediktif berdasarkan pembelajaran statistik (Laurie A. Schintler; Connie L. McNeely, 2022). Alur kerja umum dari *machine learning* adalah menerima input (disebut juga sebagai dataset), melatih model prediksi, melakukan tugas prediksi, dan akhirnya menghasilkan output.

#### 2.5.3.1 Penentuan Pendekatan Teknik *Machine Learning*

Jordan dan Mitchell (2015) mengemukakan bahwa paradigma *machine learning* dapat dibagi menjadi 3, yaitu (1) *supervised learning*, (2) *unsupervised*, dan (3) *reinforcement learning*. Sedangkan pendekatan pemilihan paradigma *machine learning* dikategorikan berdasarkan dua kriteria: (1) jenis data dari variabel dependen, apakah kontinue atau disktit (2) ketersediaan label dari variabel dependen. Variable dependen merupakan nilai yang dihasilkan dari *predictive analytics.* (Laurie A. Schintler; Connie L. McNeely, 2022). Berikut kriteria dalam menentukan pendekatan teknik *machine learning* dan tabel 2.1 mengambarkan pemetaan kriteria dan pendekatan yang akan diambil:

1. Jika dataset telah memiliki label, maka pendekatan yang dipilih adalah *supervised learning*. Sebaliknya, jika dataset tidak memiliki label, maka pendekatan yang diambil adalah *unsupervised learning*.
2. Sedangkan tipe kontinue pada variable dependen berarti nilai yang diprediksikan berkelanjutan, seperti contoh *predictive analytics* untuk harga rumah berdasarkan fitur-fitur yang dimiliki, seperti ukuran rumah, dekat dari rumah sakit atau sekolah dan sebagainya. Sedangkan tipe *diskrit* berarti *predictive analytics* menghasilkan nilai yang tetap, seperti contoh analisa sentimen terhadap *tweet* hanya memiliki kemungkinan positif atau negative.

Tabel 2. 3 Jenis pendekatan machine learning berdasarkan karakteristik variable dependen dan ketersediaan label variable dependen

|  |  |  |
| --- | --- | --- |
|  | Dataset memiliki label / *Supervised Machine Learning* | Dataset tidak memilki label / *Unsupervised machine learning* |
| Tipe nilai kontinue | Regresi | *Dimentionality Reduction* |
| Tipe nilai diskrit | Klasifikasi | *Clustering* |

#### 2.5.3.2 Teknik Pemodelan *Supervised Machine Learning*

*Supervised machine learning* digunakan untuk membangun model matematis berdasarkan input berupa dataset yang mencakup fitur-fitur variable independen dan variable dependen yang nilainya sudah diketahui (Laurie A. Schintler; Connie L. McNeely, 2022). Variable dependen sering disebut sebagai kelas variable yang nilainya dipengaruhi oleh variable independen. Dengan kata lain input berbentuk *tuple (x, y)* dari satu item pada dataset, dimana *x* adalah fitur-fitur dan *y* adalah variable dependen. Model matematis yang dihasilkan dapat dinotasikan dengan *f(x),* dimana digunakan untuk memprediksi nilai *y* (variable dependen)dari item baru yang nilai *y*-nyabelum diketahui.

Proses dalam *supervised machine learning* bisa dibagi menjadi 2, yaitu proses pembentukan model yang disebut sebagai proses induksi dan proses pemanfaatan model untuk melakukan prediksi yang disebut sebagai proses deduksi (Zafarani et al., 2014). Berdasarkan tipe data dari variable dependen jenis tugas *supervised machine learning* dapat dibagi menjadi 2, yaitu *classification* dan *regression* (Laurie A. Schintler; Connie L. McNeely, 2022). Pada penelitian ini mengadopsi *supervised machine learning* dengan tujuan hanya digunakan untuk tugas *classification,* yaitu untuk memprediksi sentiment masyarakat terhadap penyelenggaraan Pemilu 2024. Berikut macam-macam algoritma *supervised machine learning* untuk tugas *classification* yang dapat digunakan:

##### 2.5.3.2.1 *Decision Tree*

*Decision tree* adalah salah satu teknik pemodelan *machine learning* tertua (Hardeniya et al., 2016). Berdasarkan fitur yang tersedia, algoritma ini akan mengolah fitur tersebut dan membuat pohon logika. Lebih dari satu *decision tree* dapat dibentuk berdasarkan dataset yang sama walaupun hasilnya memiliki struktur logika yang berbeda. *Decision tree* dengan struktur yang berbeda dapat memberikan hasil prediksi yang sama jika dihasilkan dari data latih yang sama (Zafarani et al., 2014). Gambar 2.2 memberikan contoh *decision tree* yang terbentuk dari data latih yang ditunjukkan pada tabel 2.2. *Decision tree* memiliki beberapa komponen diantaranya adalah *non-leaf node, leaf node* dan *branch.* (a) *Non-leaf* *node* merepresentasikan sebuah fitur, (b) *branch* merepresentasikan kondisi berdasarkan nilai dari fitur dan (c) *leaf node* merepresentasikan nilai dari atribut kelas yang sedang coba diprediksi. Prediksi dihasilkan dengan mengikuti jalur logika yang dimulai dari *root node* dan berakhir sampai dengan *leaf node*. Jalur logika yang dilewati berdasarkan nilai yang dimiliki masing masing fitur pada *instance* (Zafarani et al., 2014).

Tabel 2.4 Contoh data untuk pemodelan *decision tree*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Celebrity** | **Verified Account** | **# Follower** | **Influential?** |
| 1 | Yes | No | 1.25 M | No |
| 2 | No | Yes | 1 M | No |
| 3 | No | Yes | 600 K | No |
| 4 | Yes | Unknown | 2.2 M | No |
| 5 | No | No | 850 K | Yes |
| 6 | No | Yes | 750 K | NO |
| 7 | No | No | 900 K | Yes |
| 8 | No | No | 700 K | No |
| 9 | Yes | Yes | 1.2 M | No |
| 10 | No | Unknown | 950 K | Yes |

Sumber : (Zafarani et al., 2014)

|  |  |
| --- | --- |
| Decision Tree 1 | Decision Tree 2 |

Gambar 2. 2 Dua struktur *decision tree* yang berbeda yang terbentuk dari dataset yang sama

Sumber: (Zafarani et al., 2014)

*Decision tree* dikonstrukksi secara rekursif berdasarkan data latih dengan menggunakan pendekatan *top-down greedy* (Zafarani et al., 2014). Berikut urutan proses konstruksi *decision tree:*

1. Proses konstruksi diawali dengan pemilihan fitur yang dijadikan sebagai *non-leaf node* yang dilakukan berdasarkan *purity*-nya. Fitur dengan *purity* yang tinggi merupakan fitur yang ketika digunakan untuk mempartisi data latih, partisi data latih yang dihasilkan berisikan *instance* yang semua nilai kelas atributnya sama, atau yang disebut *pure subset.*
2. Kemudian *branch* dibentuk berdasarkan nilai diskrit atau diskritisasi nilai kontinue dari fitur yang terpilih. Data latih kemudian dipecah berdasarkan *branch* yang dibentuk.
3. *Branch* dengan *pure subset* akan membentuk *leaf node,* dan jalur pada *node* tersebut berakhir.
4. Ketika ada hasil partisi yang tidak *pure* maka secara rekursif pemilihan fitur sebagai *leaf node* kembali dilakukan, namun kali ini dilakukan berdasarkan data latih yang telah terpartisi untuk masing-masing *branch*.

Terdapat beberapa algoritma konstruksi *decision tree*, diantaranya adalah CART, ID3, C4.5, dan C5 (Hardeniya et al., 2016).

##### 2.5.3.2.2 *Support Vector Machine* (SVM)

Dikembangkan oleh Vapnik di pertengahan tahun 1960-an, SVM telah berevolusi menjadi salah satu teknik *machine learning* yang efektif (Kuhn & Johnson, 2013). Sebagai teknik *mechine learning* yang popular digunakan untuktujuan klasifikasi, SVM dapat memberikan waktu cepat dalam proses pelatihan dan hasil prediksi yang akurat terhadap data dengan dimensi yang tinggi (Kelleher et al., 2020). SVM bekerja dengan berupaya menemukan *linear hyperplanes* sebagai batas pemisah kelompok data, dengan memaksimalkan jarak posisi antara kelompok data yang memiliki kelas atribut yang berbeda pada ruang dimensi (Laurie A. Schintler; Connie L. McNeely, 2022).

Komponen penting dalam SVM adalah *instance* yang terletak pada garis *margin* yang disebut *support vector*, sedangkan *margin* (yang ditunjukkan sebagai garis putus-putus pada Gambar 2.2 bagian kanan) adalah jarak antara *hyperplane* dengan *support vector* dari masing-masing kelas (Kelleher et al., 2020). *Support vector* akan menjadi penentu dalam upaya menemukan *hyperplane* paling optimal, dimana *hyperplane* yang paling optimal adalah *hyperplane* yang memiliki *margin* terjauh ke *support vector* (Kuhn & Johnson, 2013). Gambar 2.2 sebelah kiri menunjukan terdapat banyak atau bahkan tidak terhingga kemungkinan *hyperplane* yang dapat digunakan untuk mengklasifikasikan 2 kelompok data pada ruang 2 dimensi. SVM mengoptimalisasi *parameter* sehingga menghasilkan *hyperplane* yang ditunjukkan pada Gambar 2.2 sebelah kanan, proses optimalisasi ini disebut sebagai *quadratic optimization* (Kelleher et al., 2020).

A comparison of colored lines and a graph

Description automatically generated with medium confidence

Gambar 2. 3 Perbandingan kemungkinan hyperplane yang dapat digunakan dengan hyperplane yang optimal pada ruang 2 dimensi

Sumber: (Kelleher et al., 2020)

##### 2.5.3.2.3 *Naïve Bayes*

Klasifikasi Bayesian adalah kerangka kerja probabilitas untuk mengatasi kebutuhan klasifikasi dan regresi. Algoritma ini didasarkan pada penerapan Teorema Bayes dan *conditional probability* (Laurie A. Schintler; Connie L. McNeely, 2022)*. Conditional probability* merupakan kemungkinan terjadinya suatu peristiwa yang dipengaruhi oleh peristiwa lainnya[[15]](#footnote-16). Asumsi utama dalam menerapkan Teorema Bayes adalah bahwa masing-masing fitur harus bersifat independen antara satu dengan yang lain (*naive*) untuk mempermudah operasi perhitungan (Zafarani et al., 2014). Naive Bayes merupakan algoritma yang sangat sederhana dan mudah untuk diimplementasikan, dimana yang dibutuhkan hanya menyimpan frekuensi dan menghitung probabilitas (Hardeniya et al., 2016). Selain itu waktu pengeksekusian algoritma ini terbilang cepat, baik pada proses *training* maupun proses *testing.* Berikut adalah formula perhitungan *naïve bayes.*

A mathematical equation with black text

Description automatically generatedA mathematical equation with black text

Description automatically generated

Dimana *Y* adalah nilai yang mungkin dimiliki oleh atribut kelas atau label, dimana masing-masing dapat dinotasikan dengan *yi*. variable *X* adalah fitur dari *instance,* sedangkan fungsi *P()* adalah probabilitas (Zafarani et al., 2014). Sehingga (a) *P(yi|X)* adalah probabilitas terjadinya label *Y* dengan fitur-fitur *X,* (b) *P(yi)* adalah probabilitas munculnya label *yi* pada proses pelatihan yang sebelumnya, dan (c) *P(X|Y)* merupakan probabilitas munculnya fitur-fitur *X* ketika label *Y* muncul. Formula dibawah menjabarkan fungsi *P(X|yi)* ke dalam bentuk yang lebih detail.



##### 2.5.3.2.4 *K-Nearest Neighbors* (K-NN)

KNN memprediksi nilai atribut kelas berdasarkan mayoritas nilai atribut kelas dari *instance* dengan posisi terdekat sejumlah *k* yang juga disebut *neighbors* (Zafarani et al., 2014). Namun ketika tidak ada nilai atribut yang menjadi mayoritas maka nilai atribut akan ditentukan secara acak. Dalam menentukan *neighbors* penghitungan similaritas dan jarak dengan menggunakan teknik seperti *pearson correlation, cosine similarities, euclidian distance* atau teknik pengukuran lainnya dilakukan (Laurie A. Schintler; Connie L. McNeely, 2022). Kemudian dalam menentukan nilai *k* juga menjadi salah satu hal penting dengan algoritma KNN, hal ini dikarenakan nilai *k* yang berbeda dapat memberikan label atau hasil prediksi yang berbeda. Sehingga proses tuning untuk menemukan parameter *k* yang dapat memberikan hasil akurasi tertinggi menjadi suatu hal yang tidak bisa diabaikan (Laurie A. Schintler; Connie L. McNeely, 2022). Pada Gambar 2.3 memberikan contoh, nilai *k* yang berbeda dapat memberikan hasil klasifikasi yang berbeda pula, dimana nilai *k=*5 menghasilkan label segi tiga, sedangkan nilai *k=*10 menghasilkan persegi.

A circular pattern with a question mark

Description automatically generated

Gambar 2. 4 Contoh prediksi dengan algoritma KNN

Sumber: (Laurie A. Schintler; Connie L. McNeely, 2022)

KNN banyak digunakan dalam implementasi tugas klasifikasi data besar, hal tersebut dikarenakan selain mudah untuk terapkan dan cukup sederhana, KNN juga dapat memberikan hasil klasifikasi dengan tingkat akurasi yang tinggi (Deng et al., 2016). KNN juga dikenal sebagai algoritma *lazy learner* atau *instance-based learning* dikarenakan proses pembelajaran belum dilakukan atau model belum dibentuk sebelum proses klasifikasi terhadap suatu *instance* dilakukan (Laurie A. Schintler; Connie L. McNeely, 2022)*.* Hal tersebut berbeda dengan algoritma lain yang dikategorikan sebagai *eager learner* seperti *naïve bayes, decision tree* dan *artificial neural network,* dimana proses pembelajaran telah dilakukan atau model telah terbentuk sebelum proses klasifikasi terhadap suatu *instance* dilakukan[[16]](#footnote-17)

##### 2.5.3.2.5 *Random Forest*

Algoritma ini melakukan klasifikasi berdasarkan nilai modus (nilai paling serirng muncul) dari hasil klasifikasi terhadap sekumpulan *decision tree,* dimana sekumpulan *decision tree* yang disebut *forest* dibentuk berdasarkan subset data latih dan fitur yang dipilih secara acak (Rodríguez et al., 2018). Salah satu kelebihan yang dimiliki *random forest* adalah untuk mengantisipasi permasalahan *over-fitting* (klasifikasi akan akurat dengan data latih namun tidak dengan data baru)dan mereduksi kemunculan *variance* antar *decision tree.* Proses pembentukan kombinasi dari sekumpulan *decision tree* disebut *Bootstrap.* Sedangkan proses agregasi berdasarkan nilai modus dari setiap *decision tree* pada *forest* disebut *Bagging*. Menurut Hardeniya di tahun 2016, *random forest* merupakan salah satu alrogitma yang sering kali memberikan hasil akurasi terbaik (Hardeniya et al., 2016).

##### 2.5.3.2.6 *Logistic Regression*

##### 2.5.3.2.7 *Deep Learning*

*Deep Learning* adalah teknik *machine learning* yang terispirasi dari struktur dan cara kerja otak manusia (Kelleher et al., 2020). Teknik ini tersusun dari banyak unit pemprosesan terkecil yang disebut *neuron*. *Neuron* disusun ke dalam lapisan-lapisan yang saling terkoneksi dan berinteraksi. Teknik ini dianggap teknik *machine learning* yang paling ampuh dalam mempelajari data dengan pola yang tidak linear dan sangat cocok digunakan pada bidang dengan data input yang besar seperti pemprosesan gambar, suara atau teks.

Istilah *deep learning*, yang merupakan pengembangan dari *artificial neural network*, menekankan pada jumlah lapisan yang semakin banyak atau semakin dalam jika dibandingkan dengan versi terdahulunya (Kelleher et al., 2020). Lapisan yang semakin dalam memungkinkan model ini untuk dapat mengidentifikasi pola yang semakin kompleks. McCulloch and Pitts (1943), sebagai penggagas awal cara kerja *deep learning*, meniru aktifitas otak manusia berdasarkan logika proporsional, dimana mereka mendesain model yang terdiri dari sekumpulan *neuron* yang menerima input dan menghasilkan output dengan sinyal tinggi atau rendah (Kelleher et al., 2020).

Secara struktur model *deep learning* memiliki 3 jenis lapisan atau layer, yaitu *input layer*, *hidden layer*, dan *output layer*. Masing masing layer memiliki sekumpulan *neuron* didalamnya. Berikut penjelasan untuk ketiga lapisan dalam model *deep learning*.

1. *Input layer* merupakan lapisan pertama pada model *deep learning*. Pada *layer input* terdapat *sensing neuron*, *sensing neuron* pada layer ini merepresentasikan nilai fitur pada *instance*.
2. *Hidden layer* merupakan lapisan yang posisinya berada diantara *input layer* dan *output layer*. Jumlah lapisannya dinamis dan ditentukan berdasarkan *rule of thumb*, dimana ketika model dapat memberikan hasil akurasi yang tinggi.
3. *Output layer* merupakan lapisan yang di dalamnya terdapat *neuron* yang merepresentasikan nilai atribut kelas.

Gambar 2.4 menunjukan struktur model *deep learning*. Setiap koneksi memiliki nilai bobot atau *weight,* yang dinotasikan dengan wi,j. Dimana i adalah indeks dari *target neuron* dan j adalah indeks dari *source neuron*. Nilai tersebut akan terus disesuaikan ketika proses pembelajaran dilakukan. Secara alur eksekusi pembelajaran *deep learning* dikenal dengan nama *backpropagarion,* yang merupakan upaya untuk mengidentifikasi kontribusi *weight* dengan menggunakan perhitungan kalkulus *partial derivative* yang disesuaikan melalui proses *gradient descend* terhadap hasil dari *loss function* (Kelleher et al., 2020).

A diagram of a network

Description automatically generated

Gambar 2. 5 Struktur model *deep learning.*

Sumber: (Kelleher et al., 2020)

#### Evaluasi Pemodelan *Supervised Machine Learning*

Evaluasi pemodelan bertujuan untuk memastikan model yang terbentuk dapat digunakan sesuai dengan kebutuhan, mengestimasi kinerja model ketika diimplentasikan, dan menunjukkan kepada pengguna bahwa model yang terbentuk dapat memenuhi harapan (Kelleher et al., 2020). Ketiga hal tersebut dilakukan dengan mengkomparasikan hasil pelabelan manual dengan prediksi dari model terlatih menggunakan *confusion matrix*. Tabel 2.3 mengambarkan pemetaan *confusion matrix* yang memiliki 4 kemungkinan prediksi, yaitu:

* *True Positive* (TP): merupakan *instance* pada *testing set* yang baik hasil pelabelan dan hasil prediksi bernilai positif.
* *True Negative* (TN): merupakan *instance* pada *testing set* yang baik hasil pelabelan dan hasil prediksi bernilai negatif.
* *False Positive* (FP): merupakan *instance* pada *testing set* dengan hasil pelabelan bernilai negatif, sedangkan hasil prediksi bernilai positif.
* *False Negative* (FN): merupakan *instance* pada *testing set* dengan hasil pelabelan bernilai positif, sedangkan hasil prediksi bernilai negatif.

Tabel 2. 5 Struktur *confusion matrix* untuk tugas prediksi sederhana dengan 2 nilai nominal

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | *Prediction* | | | |
| Positif | | Negatif | |
| *Target / Actual* | positif | TP | | FN | |
| negatif | FP | | TN | |

Sekilas jika dilihat secara diagonal masing masing kemungkinan prediksi (TP dan TN atau FN dan FP) dapat menunjukan kinerja model, seperti tingginya nilai TP dan TN menunjukkan kinerja model yang baik (Kelleher et al., 2020). Selain itu *confusion matrix* juga dapat digunakan sebagai dasar dalam menghitung *performance metrics* sebagai acuan pemeringkatan model terbaik, seperti: *precision, recall* dan *f1-measure.* Tabel 2.4 menjelaskan lebih detail tentang ketiga metrik tersebut.

Tabel 2. 6 *Performance metrics* (*precision, recall, f1-measure*) yang digunakan pada tahapan evaluasi model.

|  |  |
| --- | --- |
| Penjelasan | Formula |
| *Precision* menunjukkan seberapa persenkan dari keseluruhan *instance* yang diprediksi sebagai kelas positif, yang berhasil diprediksi dengan tepat. |  |
| *Recall* menunjukkan seberapa persenkan dari keseluruhan *instance* yang secara aktual berlabel positif, yang berhasil diprediksi dengan tepat. |  |
| *F1-measure* merupakan harmonisasi dari *precision* dan *recall.* |  |

Ketiga *performance metrics* tersebut memiliki rentang nilai 0 sampai dengan 1 dimana nilai yang lebih tinggi mengindikasikan performa model yang lebih baik. Selain itu dalam evaluasi model juga dibutuhkan desain eksperimen yang baik. Hal ini untuk memastikan dilakukanya proses estimasi yang ketat sebagimana model yang terbentuk nantinya akan diimplementasikan (Kelleher et al., 2020). Terdapat 5 pendekatan desain eksperimen evaluasi pemodelan diantaranya adalah:

* *Hold-out sampling* memecah dataset secara acak ke dalam 2 bagian, bagian pertama dipergunakan sebagai data latih pembentukan model (*training set*) dan bagian kedua sebagai data pengujian (*testing set*). Tidak ada rekomendasi pasti dari persentase *training set* dan *testing set*, namun konfigurasi *training set:testing set* yang biasa digunakan adalah 70:30 dan 60:40 (Kelleher et al., 2020).
* K-*fold cross validation* memecah data ke dalam beberapa partisi sebanyak *k* (p1, p2, p3, …, p*k*) dan eksperimen evaluasi sejumlah *k* akan dilakukan secara terpisah. Setiap partisi data akan secara bergantian dijadikan sebagai *testing set* (Kelleher et al., 2020)*.* Seperti contoh pada eksperimen evaluasi pertama, data partisi pertama (p1) akan dijadikan *testing set* sedangkan partisi lainnya dijadikan sebagai *training set.* Dilanjutkan dengan eksperimen evaluasi kedua, dimana data partisi kedua (p2) dijadikan *testing set* sedangkan data partisi pertama dan partisi lainnya dijadikan sebagai *training set.* Eksperimen evaluasi dilanjutkan sampai eksperimen ke-*k* telah selesai dilakukan*.* Sebagai hasil akhir, setiap *performance metrics* yang dihasilkan dari masing-masing eksperimen evaluasi akan dihitung nilai rata-ratanya. Walaupun tidak ada rekomendasi terkait nilai *k* yang ideal, namun banyak penelitian menggunakan nilai 10 atau 10-*fold cross validation.*
* *Leave-one-out cross validation* mirip seperti *k-fold cross validation* dimana nilai *k* sejumlah *instance* pada dataset. Desain eksperimen evaluasi ini cocok digunakan ketika jumlah dataset sedikit, dimana memungkinkan *training set* mendapatkan porsi data yang cukup banyak (Kelleher et al., 2020).
* *Bootstraping* mirip seperti *k-fold cross validation* namun partisi *testing set* dibentuk secara acak (Kelleher et al., 2020). Desain eksperimen evaluasi ini cocok digunakan dalam konteks jumlah data pada dataset sedikit (di bawah 300 *instance*)*.*
* *Out of time sampling* mirip seperti *hold-out sampling* namun pemilihan *testing set* tidak dilakukan secara acak (Kelleher et al., 2020). Desain eksperimen evaluasi ini dapat digunakan ketika *instance* pada dataset memiliki atribut waktu, yang memungkinkan untuk membangun model berdasarkan dataset (*training set*) dengan atribut periode waktu tertentu dan membentuk *testing set* dengan atribut periode waktu yang lain.

Penelitian ini akan menggunakan 3 *performance metrics,* yaitu: *precision, recall* dan *f1-measure* dan pendekatan *k-fold cross validation* sebagai desain eksperimen evaluasi model.

#### 2.5.3.4 Pemodelan Topik

Pemodelan Topik merupakan pendekatan yang digunakan untuk menemukan struktur (semantik) atau pola tersembunyi dari korpus yang jumlahnya sangat besar (Jelodar et al., 2019). Dalam bidang NLP, pemodelan topik sangat banyak digunakan dengan tujuan mendeteksi topik dan penggalian pola semantik (*semantic mining*) baik pada dokumen yang terstruktur maupun yang tidak terstruktur. Dalam bidang ilmu komputer, daftar topik yang dihasilkan, bukan berasal dari analisa makna ataupun konsep linguistik dari teks pada korpus, melainkan berdasarkan statistik kemunculannya dalam korpus (Jelodar et al., 2019).

Salah satu teknik pemodelan topik yang banyak digunakan saat ini adalah *Latent Dirichlet Allocation* (LDA). LDA, teknik yang dikenalkan oleh Blei, NG dan Jordan pada 2003, merepresentasikan topik berdasarkan probabilitas kemunculan kata (Jelodar et al., 2019). Kata dengan probabilitas kemunculan tertinggi dapat memberikan makna bagi topik yang sedang dideteksi. Teknik ini mengasumsikan dokumen dalam korpus merupakan distribusi probabilistik dari topik dan topik merupakan distribusi probabilistik dari kemunculan kata pada korpus. Berikut algoritma pembentukan model dengan teknik LDA:

1. Berdasarkan korpus *D* yang berisi dokumen berjumlah *M*, dimana di dalam satu dokumen *d* memiliki *Nd* kata.
2. Menentukan distribusi multinomial *φ*t untuk semua topik dari distribusi *derichlet* dengan parameter *β.*
3. Menentukan distribusi multinomial *ϴd* untuk semua dokumen dari distribusi *derichlet* dengan parameter *α.*
4. Distribusi multinomial yang ditentukan pada langkah sebelumnya digunakan untuk mementukan probabilitas *wn* pada dokumen *d*. Berikut adalah formula dari 4 langkah ini:

A black and white math symbol

Description automatically generated

## Tinjauan Penelitian Terdahulu

*Systematic literature review* (SLR) adalah upaya untuk menggali wawasan terkini dan mengidentifikasi arah penelitian ke depan dengan melakukan meninjauan terhadap penelitian terdahulu yang relevan (Jain et al., 2021). Terdapat 6 tahapan dalam SLR, yaitu: (1) menentukan *research question,* (2) menentukan kriteria pencarian, (3) mengumpulkan literatur berdasarkan kriteria, (4) memilah literatur yang relevan, (5) menyintesis informasi yang relevan, (6) melaporkan hasil peninjauan (Drus & Khalid, 2019).

### Menentukan *Research Question*

Penelitian ini telah menentukan research question sebagaimana telah disebutkan pada bagian 1.3 Pertanyaan Penelitian.

### Menentukan Kriteria Pencarian

Berdasarkan *research question* maka disusun kriteria dalam melakukan pencarian, Penelitian ini menggunakan pendekatan *machine learning* dalam melakukan analisis sentimen, maka dibutuhkan kata kunci sebagai berikut: “machine learning”, “analisis sentimen” OR “sentiment analysis”. Berdasarkan domain bahasan yaitu penelitian terkait dengan penyelenggaraan Pemilu 2024, maka dibutuhkan kriteria tambahan yaitu: “pemilu” OR “election”, “kebijakan umum” OR “public policy”, “government” OR “pemerintah”. Berikut susunan *query* pencarian: “machine learning” AND (“analisis sentimen” OR “sentiment analysis”) AND ((“pemilu” OR “election”) OR (“kebijakan umum” OR “public policy”) OR (“government” OR “pemerintah”)).

### Mengumpulkan Literatur

Pencarian dilakukan pada 2 database literatur yaitu Scopus dan UI lib[[17]](#footnote-18). Khusus pencarian pada UI lib tidak disertakan kata kunci berdasarkan domain pembahasan. Selain Scopus dan UI Lib, peneliti juga menggunakan database literatur seperti ACM, Google Scholar dan SpringerLink untuk mencari rujukan yang berasal dari literatur pada 2 database sebelumnya. Tabel 2.6 menunjukkan jumlah literatur hasil pencarian dari 2 database literatur tersebut:

Tabel 2. 7 *Jumlah literatur hasil pencarian*

|  |  |
| --- | --- |
| Database | Jumlah Literatur |
| Scopus | 214 |
| Lib UI | 50 |
| Total | 264 |

### Memilih Literatur yang Relevan

Pada tahapan ini, dari 264 literatur yang ditemukan maka dilakukan seleksi berdasarkan isi atau kandungan pada bagian judul dan abstrak. Berikut kriteria yang digunakan dalam melakukan seleksi: (1) Apakah penelitian membahas tentang analisis sentimen menggunakan pendekatan *machine learning*? dan (2) Apakah bahasan dalam penelitian terkait dengan kebijakan publik atau Pemilihan Umum?. Berdasarkan seleksi yang dilakukan, maka didapat 5 penelitian yang akan digunakan pada 2 tahapan selanjutnya, yaitu: menyintesis informasi yang relevan dan melaporkan hasil peninjauan. Kedua tahapan tersebut akan dibahas pada bagian 2.7 Penelitian Terdahulu, 2.8 Ringkasan Penelitian Terhadulu dan 2.9 Perbandingan Penelitian Terdahulu.

## Penelitian Terdahulu

Bagian ini menjelaskan 5 penelitian terdahulu hasil dari proses sebelumnya, dimana akan menjadi landasan bagi peneliti untuk melakukan penelitian ini. Berikut adalah daftarnya:

### Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi

Penelitian yang dilakukan oleh Dian Isnaeni (2022) ini bertujuan untuk memberikan masukan kepada Komisi Pemberantasan Korupsi (KPK) terkait penerapaan Laporan Harta Kekayaan Penyelengara Negara (LHKPN). Masukan yang diberikan bertujuan agar pemberantasan korupsi menjadi lebih optimal. Dengan menganlisa opini masyarakat terkait pelaksanaan LHKPN di media sosial dengan menggunakan pendekatan analisis sentimen dan pemodelan topik. Berikut adalah ringkasan dari penelitian tersebut:

1. *Dataset*

Pengumpulan data dilakukan periode 1 Agustus – 5 November 2021 pada media sosial Twitter, dengan menggunakan kata kunci “harta kekayaan pejabat” dan “LHKPN”. Pengumpulan data dilakukan dengan pendekatan *web scrapping* dalam bahasa pemrograman *python*. Sebanyak 894 *tweet* diperoleh pada tahapan pengumpulan data ini.

1. Anotasi

Proses ini dilakukan dengan melibatkan 3 orang *annotator*. *Tweet* dilabelkan ke dalam 3 kategori yaitu positif, negatif dan netral. Dalam menentukan label dilakukan *scoring* menggunakan koefisien *Kappa*.

1. *Preprocessing* dan *Text Representation*

Penelitian ini menggunakan beberapa teknik teks *preprocessing,* yaitu: (1) *case folding*, (2) menghapus *URL*, *number*, *puncuation*, *special character*, (3) *tokenization* (4) *normalization* (5) *stemming* dan (6) *stopwords removal*. Selanjutnya dilakukan representasi teks dengan pendekatan TF-IDF (unigram, bigram, dan trigram) dan *CountVectorizer*.

1. Metode Klasifikasi

Penelitian ini menggunakan teknik *supervised machine learning* seperti *Artificial Neural Network*, *Support Vector Machine*, *Logistic Regression*, *Decision Tree*, *Gradient Boosting*, dan *Random Forest*.

1. Evaluasi

Dalam proses evaluasi penelitian ini memanfaatkan teknik *confusion matrix* dan 4 metrik pengukuran yaitu: *precision, accuracy, recall* dan *F1* *score.*

1. Hasil

Pada penelitian ini *multilayer perceptron* atau *artificial neural network* dengan representasi teks unigram *CountVectorizer* menjadi metode klasifikasi terbaik dengan nilai *accuracy* 78.60%, *precision* 78,19%, *recall* 76,60%, dan *F1 score* 76,95%. Sedangkan untuk pemodelan topik, penelitian ini menggunakan metode *Latent Dirichlet Allocation* (LDA). Pomodelan topik diterapkan pada data dengan hasil prediksi positif dan negatif. Untuk data dengan hasil label negatif menunjukkan kekecewaan masyarakat terhadap laporan semakin tingginya angka rata-rata kekayaan Penyelenggara Negara (PN) ketika pandemi Covid-19. Sedangkan label positif memiliki topik yang bervariasi, salah satunya seperti kewajaran peningkatan harta kekayaan PN yang disebabkan peningkatan harta *fixed asset.*

### Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19

Penelitian yang dilakukan oleh Pandu Maulana (2022) ini bertujuan untuk dapat memberikan masukan bagi pemerintah dalam membuat kebijakan terkait penangganan masalah pengangguran selama masa pandemi COVID-19 berdasarkan opini masyarakat yang mereka utarakan di media sosial Twitter. Berikut adalah detail dari penelitian ini:

* 1. *Dataset*

Penelitian ini berhasil mengumpulkan 6589 *tweet* dalam rentang waktu 2 bulan pada *platform* media sosial Twitter. *Web crawling* Twitter APIdengan *library tweepy* menjadi pendekatan yang dipilih untuk proses ini.

* 1. Anotasi

Penelitian ini melibatkan 3 orang annotator, yang bertugas untuk mengkategorikan dataset ke dalam 4 label, yaitu: positif, negatif, netral, dan tidak relevan. Tahapan ini menghasilkan 1710 data berlabel positif, 1553 data berlabel negatif, 2198 data berlabel netral, dan 1128 data berlabel tidak relevan.

* 1. *Preprocessing* dan *Text Representation*

Penelitian ini menggunakan teknik pra-pemprosesan teks seperti *tokenizing*, *stop word removal,* identifikasi jenis kata (*POS Tagging*). Penelitian ini menggunakan representasi teks *Bag of Word,* namun tidak dijelaskan secara terperinci pendekatan yang diambil seperti TF-IDF, *CountVectorizer* atau lainnya*.*

* 1. Metode Klasifikasi

Penelitian ini mengkomparasikan 3 metode klasifikasi yaitu *Naïve Bayes, Decision Tree* dan *Random Forest.*

* 1. Evaluasi

Penelitian ini memanfaatkan *confusion matrix* dengan metrik *accuracy, precision, recall* untuk mengevaluasi kinerja model. Sedangkan teknik distribusi pengujian yang dilakukan adalah dengan menggunakan pendekatan *K-fold cross validation* dengan nilai k=5.

* 1. Hasil

Penelitian ini menunjukkan bahwa *Random Forest* sebagai metode klasifikasi terbaik dengan *score accuracy* 79%, *precision* 70% dan *recall* 78%. Penelitian ini juga memaparkan bahwa sentimen positif mendominasi dengan 1710 *tweet* dibandingkan dengan sentimen negatif sebanyak 1553 *tweet.* Selain itu penelitian ini juga memaparkan fitur yang dominan muncul, fitur tersebut dapat dimanfaatkan sebagai indikator topik atau *concern* dari opini yang disampaikan.

### Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter.

Penelitian yang dilakukan oleh Ghanim Kanugrahan (2021) ini bertujuan untuk membantu pemerintah dalam menampung aspirasi masyarakat yang mereka utarakan melalui media sosial Twitter terkait dengan kembalinya sistem pendidikan tatap muka dengan menggunakan pendekatan analisis sentimen. Berikut rangkuman dari penelitian ini:

* + 1. Dataset

Dalam proses pengumpulan data, penelitian ini memanfaatkan bahasa pemrograman *python* dengan *library tweepy* dalam melakukan *web crawling* Twitter API*.* Dengan menggunakan kata kunci seperti “sekolah tatap muka”, “sekolah offline”, dan “kuliah offline”dan dilakukan pada periode 11 April 2021 sampai dengan 26 April 2021, proses ini berhasil mengumpulkan sebanyak 15.000 *tweet*.

* + 1. Anotasi

Penelitian ini melibatkan 3 orang *annotator* untuk melabeli *tweet* ke dalam empat kategori yaitu: positif, negatif, netral dan tidak relevan. Proses ini menghasilkan 2697 *tweet* positif, 1565 *tweet* negatif, 799 *tweet* netral dan 10063 *tweet* yang tidak relevan dengan topik.

* + 1. *Preprocessing* dan *Text Representation*

Teknik *preprocessing* teks seperti (1) *case folding,* (2) menghilangkan *mentions*, angka, tanda baca (3) *tokenization* (4) *normalization* (5) *stopword removal* dan (6) *stemming* dilakukan pada penelitian ini*.* Disamping itu peneliti juga menambahkan 2 fitur baru yaitu “jumlah kata” dan “waktu *posting* *tweet*”. Alasan peneliti dalam menambahkan 2 fitur tersebut adalah yang pertama peneliti menyebutkan ada asumsi bahwa fitur “jumlah kata” berkorelasi dengan jenis sentimen. Sedangkan untuk fitur yang kedua, “waktu *posting tweet”,* berdasarkan penelitian yang dilakukan oleh Arakawa et. al (2018)depat meningkatkan hasil akurasi. Untuk representasi teks penelitian ini menggunakan *bag of word* dengan teknik penghitungan TF-IDF unigram dan bigram.

* + 1. Metode Klasifikasi

Penelitian ini mengkomparasikan metode SVMdan *multilayer perceptron* (*artificial neural network* dengan 10 *hidden layer* dan 8 *neuron* pada masing-masing *hidden layer*)*.*

* + 1. Evaluasi

Dengan menggunakan *confusion matrix* penelitian ini menghitung nilai *accuracy, precision, recall* dan *F1-score* dari model yang dihasilkan.

* + 1. Hasil

Pada penelitian ini SVM keluar sebagai algoritma klasifikasi terbaik dengan *score* 85,78% untuk TF-IDF unigram dan 81,0% untuk TF-IDF bigram. Wawasan yang dihasilkkan oleh penelitian ini adalah masyarakat lebih mendukung kembalinya sistem pendidikan tatap muka walaupun ada kekhawatiran bagi mereka kembali merembaknya virus COVID-19 ke tengah-tengah mereka. Hal ini diindikasikan dengan *tweet* yang diprediksi bersentimen positif lebih mendominasi dengan jumlah 2.697 *tweet* jika dibandingkan dengan yang bersentimen negatif, dimana hanya menyentuh angka 1.565 *tweet*.

### Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model *Machine Learning*.

Penelitian yang dilakukan oleh Dennis et.al. (2021) ini bertujuan untuk dapat memberikan rekomendasi kepada Pemerintah terkait dengan kebijakan Pemberlakukan Pembatasan Kegiatan Masyarakat (PPKM) yang berlaku selama masa pandemi COVID-19 di Jakarta. Rekomendasi yang diberikan diperoleh berdasarkan penelitian analisis sentimen terhadap komentar masyarakat di media sosial Facebook dan Youtube. Berikut adalah ringkasan dari penelitian ini:

* + - 1. Dataset

Penelitian ini menggunakan pendekatan *web scrapping* berbahasa pemprograman R (*library Rselenium* and *Rvest*)untuk mengumpulkan data yang bersumber dari media sosial Facebook dan Youtube. Dengan menggunakan kata kunci “PPKM Jakarta” diperoleh 3.583 komentar masyarakat terkait PPKM. Lebih spesifik lagi komentar yang diperoleh berasal dari tautan yang dikirim oleh Pemprov DKI di media sosial Facebook dan berita tentang PPKM di Jakarta dari beberapa *channel* TV nasional di media sosial Youtube.

* + - 1. Anotasi

Penelitian ini melibarkan 2 orang *annotator* yang ditugaskan untuk memberikan label terhadap dataset. 3 label yang menjadi nilai nominal dari *dependent variable* adalah positif, negatif dan netral.

* + - 1. *Preprocessing* dan *Text Representation*

Beberapa teknik *text preprocessing* dilakukan pada penelitian ini diantaranya adalah (1) *case folding,* (2) menghapus *links* (3) *stopword removal* (4) mengubah *hashtag* menjadi kata (5) normalisasi kata bahasa gaul menjadi bahasa bakunya. (6) menghilangkan komentar yang duplikat, langkah ini mereduksi dataset hingga 1.214 komentar, dengan 659 komentar positif dan 343 komentar negatif. Penelitian ini menggunakan teknik undersampling untuk menyeimbangkan data, hasilnya masing masing sentimen memiliki 340 *instance.* Untuk representasi teks penelitian ini menggunakan pendekatan *bag of word,* dimana ada 6 teknik yang digunakan dan dikomparasikan, yaitu: *word unigram*, *word bigram, word n-gram, character trigram, character quadrigram* dan *character n-gram.*

* + - 1. Metode Klasifikasi

Penelitian ini menggunakan 3 metode klasifikasi, yaitu: *Naïve Bayes*, *Logistic Regression* dan *Random Forest.*

* + - 1. Evaluasi

Penelitian ini menggunakan metrik *F-measure* dengan teknik distribusi pengujian 10-*fold cross validation.*

* + - 1. Hasil

Dalam penelitian ini menunjukkan bahwa teknik *Logistic Regression* dengan representasi teks *quandrigram character* menjadi pendekatan dengan *F-measure* tertinggi dengan *score* 79.6%. Sentimen netral mendominasi hasil prediksi dengan 48%, diikuti dengan sentimen negatif 31.7% dan sentimen positif 18.4% . Peneliti juga mengemukakan bahwa pengujian dengan data tidak seimbang menyebabkan hasil prediksi yang buruk dan menghapus fitur berdasarkan *document frequency* (DF) meningkatkan *score* rata-rata *F-measure* sebanyak 4.4 persen. Penelitian ini juga menghasilkan wawasan berupa opini masyarakat hasil dari analisa menggunakan *word cloud* dari masing-masing kategori sentimen. Untuk komentar bersentimen negatif PPKM berdampak pada sulitnya masyarakat dalam memenuhi kebutuhan dasar sehari-hari dan perlunya bantuan sosial dari pemerintah untuk mengatasi masalah tersebut. Sedangkan untuk komentar bersentimen netral penyegeraan distribusi vaksin COVID-19 menjadi *concern* masyarakat yang mendominasi untuk kategori tersebut. Sementara pada komentar bersentimen positif ‘kerja nyata’, ‘terima kasih’ menjadi fitur yang paling banyak muncul, mengindikasikan dukungan atas kebijakan yang diambil pemerintah.

### Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia.

Penelitian yang dilakukan oleh Taufan Ardi (2019) ini bertujuan untuk menidentifikasi pola demografi dukungan masyarakat pada kontestasi pemilihan Presiden Indonesia 2019. Analisis demografi dan sentimen yang dilakukan memanfaatkan pendekatan *machine learning* terhadap komentar masyarakat di media sosial Twitter. Berikut ringkasan dari penelitian ini:

* + - * 1. *Dataset*

Pada penelitian ini terdapat 6 tahap pengumpulan data yang dibagi berdasarkan fungsi data yang dikumpulkan. Diantaranya adalah (1) Pengumpulan data *tweet,* (2) Pengumpulan data *akun twitter*, (3) Pengumpulan data aktifitas akun *twitter,* (4) Pengumpulan kamus nama dan jenis kelamin (bersumber dari KPU), (5) Pengumpulan data training dan (6) Pengumpulan data untuk evaluasi. Pengumpulan data pada Twitter menghasilkan 199,727 *tweet* dari total 16,170 akun.

* + - * 1. Anotasi

Penelitian ini tidak menerangkan secara detail proses anotasi yang dilakukan.

* + - * 1. *Preprocessing* dan *Text Representation*

Penelitian ini tidak menerangkan secara detail proses *text cleansing* yang dilakukan. Namun pada tahapan representasi teks, penelitian ini menggunakan pendekatan TF-IDF unigram. Selain memanfatkan deskripsi akun dan *tweet* untuk membentuk fitur teks, beberapa fitur juga dibentuk untuk nantinya digunakan dalam klasifikasi jenis kelamin, seperti: First\_male, Middle\_male, Last\_male, First\_female, Middle\_female, Last\_female. Fitur-fitur tersebut dihasilkan berdasarkan perbandingan yang dilakukan antara nama pada akun twitter dengan nama pada kamus data nama dan jenis kelamin yang bersumber dari KPU.

* + - * 1. Metode Klasifikasi

Penelitian ini melakukan 4 macam pemodelan yang dibagi berdasarkan tujuan prediksi, yaitu: (1) Klasifikasi jenis kelamin, (2) Klasifikasi umur, (3) Klasifikasi lokasi, (4) Klasifikasi sentiment. Untuk Klasifikasi 1-3, penelitian ini menggunakan teknik *machine learning* seperti: *Multinomial Naïve Bayes, Logistic Regression, Random Forest* dan SVM. Sedangkan teknik *SentiStrength,* dengan *scoring* berdasarkan *hashtag* kampanye pada *tweet,* digunakan untuk klasifikasi sentimen.

* + - * 1. Evaluasi

Penelitian ini menggunakan metrik *F1-score*.

* + - * 1. Hasil

Untuk klasifikasi jenis kelamin, metode klasifikasi *multinomial naïve bayes* dengan fitur First\_male, Middle\_male, Last\_male, First\_female, Middle\_female, Last\_female dan nama lengkap menunjukkan hasil akurasi tertinggi dengan nilai 74,5%. Sedangkan untuk klasifikasi umur, metode klasifikasi *Logistic Regression* unigram mampu menghasilkan akurasi tertinggi dengan nilai 70%, dengan selisih tidak terlalu jauh untuk perbandingan antar fitur yang digunakan. Untuk klasifikasi lokasi, metode *linear* SVM dengan fitur nama dan username menunjukkan hasil tertinggi. Terakhir untuk klasifikasi sentimen, pendukung berjenis kelamin laki-laki, 61% netral, 18% mendukung paslon 01 dan 21% pendukung paslon 02. Sedangkan untuk pendukung berjenis kelamin perempuan 59% netral, 20% mendukung paslon 01 dan 21% mendukung paslon 02. Wawasan yang dihasilkan pada penelitian ini adalah: (1) generasi milenial berjenis kelamin perempuan adalah segmen yang paling aktif, (2) pendukung paslon 01 walaupun lebih sedikit jumlahnya namun lebih aktif dibandingkan dengan pendukung paslon 02.

## Ringkasan Penelitian Terdahulu

Tabel 2.6 meringkas daftar penelitian terdahulu yang disebutkan pada bagian sebelumnya. Ringkasan disajikan dengan mengkomparasikan *dataset*, *text representation*, metode klasifikasi, metode evaluasi model dan metode pembentukan topik.

Tabel 2. 8 Ringkasan Penelitian Terdahulu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Judul | Dataset | *Text Representation* dan Metode Klasifikasi | Evaluasi | Metode Pembentukan Topik. |
| 1 | Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi (Afra, 2022). | 894 *tweet* berbahasa Indonesia.  *dependent variable* datasetmemiliki 3 nilai nominal: positif, negatif dan netral  Perbandingan 70:30 untuk pembagian data latih dan data uji. | *Text Representation*:   * TF-IDF (unigram, bigram, dan trigram) * *CountVectorizer*   Metode Klasifikasi:   * *Artificial Neural Network* * SVM * *Logistic Regression* * *Decision Tree* * *Gradient Boosting* * *Random Forest* | Metriks: *precision, accuracy, recall* dan *F1* *score.*  Pendekatan eksperimentasi: *hold-out sampling.* | *Latent Dirichlet Allocation* (LDA). |
| 2 | Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19 (Maulana et al., 2022). | 6859 *tweet* berbahasa Indonesia  *dependent variable* datasetmemiliki 4 nilai nominal: positif, negatif, netral, dan tidak relevant.  Perbandingan 70:30 untuk pembagian data latih dan data uji. | *Text Representation*:   * TF-IDF (unigram dan bigram).   Metode Klasifikasi:   * *Naive Bayes* * *Decision Tree* * *Random Forest* | Metriks: *accuracy, precision, recall.*  Pendekatan eksperimentasi: 5-*fold cross validation.* | Penelitian ini menggunakan analisis frekuensi kemunculan kata untuk dapat mengidentifikasi topik atau tema-tema aspirasi masyarakat. |
| 3 | Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter (Kanugrahan & Wicaksono, 2021). | 15.000 *tweet* berbahasa Indonesia.  *dependent variable* datasetmemiliki 4 nilai nominal: positif, negatif, netral, dan tidak relevant.  Perbandingan 80:20 untuk pembagian data latih dan data uji. | *Text Representation*:   * TF-IDF (unigram dan bigram).   Metode Klasifikasi:   * *Artificial Neural Network* * SVM | Metriks: *accuracy, precision, recall* dan *F1-score.*  Pendekatan eksperimentasi: 10-*fold cross validation.* | Penelitian ini menggunakan analisis frekuensi kemunculan kata untuk dapat mengidentifikasi topik atau tema-tema aspirasi masyarakat. |
| 4 | Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model *Machine Learning* (Dennis et al., 2021). | 3.583 komentar pada *platform* Facebook dan YouTube berbahasa Indonesia.  *dependent variable* datasetmemiliki 3 nilai nominal: positif, negatif, dan netral.  Perbandingan 70:30 untuk pembagian data latih dan data uji. | *Text Representation*:   * *CountVectorizer* (*word unigram*, *word bigram, word n-gram, character trigram, character quadrigram* dan *character n-gram*)   Metode Klasifikasi:   * *Naive Bayes* * *Logistic Regression* * *Random Forest* | Metriks: *f-measure*  Pendekatan eksperimentasi: 10-*fold cross validation* | Penelitian ini menggunakan analisis frekuensi kemunculan kata untuk dapat mengidentifikasi topik atau tema-tema aspirasi masyarakat. |
| 5 | Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia (Arafat et al., 2020). | 199,727 *tweet* berbahasa Indonesia dari total 16,170 akun.  Penulis ini tidak menjelaskan pembagian data latih dan data uji. | *Text Representation*:   * TF-IDF (unigram)   Metode klasifikasi usia, lokasi dan jenis kelamin:   * *Naïve Bayes* * *Logistic Regression* * SVM * *Random Forest*   Metode klasifikasi sentimen: *SentiStrength* | Metrik: *F1-score*  Penulis tidak menjelaskan pendekatan eksperimentasi evaluasi yang dilakukan. | Penelitian ini tidak mengadopsi metode pemodelan topik. |

## Perbadingan Penelitian Terdahulu

Tabel 2.7 menunjukkan hasil analisis *compare, contrast, critisize, summarize, syntesize* (3C2S) dari daftar penelitian terhadulu yang disebutkan pada bagian sebelumnya. *Compare* mengidentifikasi kesamaan, *contrast* mengidentifikasi perbedaan, *criticize* mengidentifikasi kekurangan, *syntesize* mengidentifikasi hal yang dapat diambil untuk memperkaya penelitian yang akan dilakukan, *summarize* meringkas hasil analisa yang telah dilakukan.

Tabel 2. 9 Analisis 3C2S penelitian tentang Analisis Sentimen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Judul Penelitian | *Compare* | *Contrast* | *Criticize* | *Syntesize* |
| 1 | Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi (Afra, 2022). | Melakukan analisis sentimenterhadap aspirasi masyarakat di media sosial Twitter dengan teknik *machine learning* untuk memberikan rekomendasi bagi Pemerintah. | (1) Menggunakan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA) yang bertujuan untuk menggali wawasan akan *concern* masyarakat.  (2) Mengkomparasikan teknik *Artficial Neural Network* dengan 5 classical machine learning yaitu: *Support Vector Machine, Logistic Regression, Decision Tree, Gradient Boosting, dan Random Forest* | (1) Penelitian tidak menerangkan secara terperinci arsitektur *Artificial Neural Network* yang digunakan.  (2) Penelitian tidak melibatkan proses klasifikasi untuk data yang ‘tidak relevan’. Beberapa penelitian telah menunjukkan klasifikasi data ‘tidak relevan’ dapat membantu meningkatkan akurasiprediksi *machine learning.* | Penelitian ini menunjukkan *Artificial Neural Network* dengan unigram *CountVectorizer* menjadi teknik dengan nilai *F1 score* tertinggi 76,95% dan identifikasi topik dapat dilakukan dengan metode *Latent Dirichlet Allocation* (LDA). |
| 2 | Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19 (Maulana et al., 2022). | Melakukan analisis sentimenterhadap aspirasi masyarakat di media sosial Twitter dengan teknik *machine learning* untuk memberikan rekomendasi bagi Pemerintah. | Menggunakan teknik klasifikasi Naïve Bayes, Decision Tree, dan Random Forest. | (1) Penelitian tidak menyebutkan periode pengumpulan data, peneliti hanya menyebutkan berapa lama waktu yang dibutuhkan untuk pengumpulan data.  (2) Dari sisi teknik yang digunakan penelitian ini tidak menunjukan sesuatu yang bersifat khusus, seperti melakukan eksperimental atau menggunakan teknik *machine learning* yang tergolongbaru, penelitian ini mengkomparasikan 3 *classical machine learning.* | Penelitian ini menggunakan frekuensi kemunculan kata untuk mengidentifikasi topik atau tema-tema aspirasi masyarakat. |
| 3 | Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter (Kanugrahan & Wicaksono, 2021). | Melakukan analisis sentimenterhadap aspirasi masyarakat di media sosial Twitter dengan teknik *machine learning* untuk memberikan rekomendasi bagi Pemerintah. | (1) Mengkomparasikan teknik machine learning *Artificial Neural Network* (*Multilayer Perceptron*) dengan *classifical machine learning* *Support Vector Machine* (SVM)  (2) Memanfaatkan fitur “jumlah kata” dan “waktu *posting tweet”* untuk menunjang hasil klasifikasi. | (1) Dataset yang digunakan untuk melatih machine learning tidak seimbang. Hal ini bisa mengurangi tingkat prediksi akurasi *machine learning.*  (2) Penulis tidak menjelaskan kriteria yang digunakan pada tahapan anotasi atau labeling data.  (3) Peneliti tidak menjelaskan teknis pembagian training dan testing dataset. | Hal yang menonjol pada penelitian ini adalah pemanfaatan fitur “jumlah kata” dan “waktu *posting tweet”* untuk menunjang hasil klasifikasi. Disamping itu komparasi antara teknik *multilayer perceptron* yang dianggap lebih moderenketimbang *classical machine learning* seperti SVM, ternyata memberikan hasil yang berbanding terbalik. SVM menjadi teknik dengan nilai *f-measure* lebih tinggi, dengan *score* 85,78%. |
| 4 | Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model Machine Learning (Dennis et al., 2021). | Melakukan analisis sentimenterhadap aspirasi masyarakat di media sosial Twitter dengan teknik *machine learning* untuk memberikan rekomendasi bagi Pemerintah. | (1) Menggunakan teknik *undersampling* untuk membuat dataset menjadi seimbang.  (2) Mengkomparasikan 6 teknik representasi teks. yaitu: word unigram, word bigram, word n-gram, character trigram, character quadrigram dan character n-gram.  (3) Menggunakan 3 teknik machine learning. yaitu: Naïve Bayes, Logistic Regression dan Random Forest. | Penelitian tidak melibatkan proses klasifikasi untuk data yang ‘tidak relevan’. Beberapa penelitian telah menunjukkan klasifikasi data ‘tidak relevan’ dapat membantu meningkatkan akurasiprediksi *machine learning.* | Penelitian ini menunjukkan bahwa representasi teks berbasis jumlah karakter (*character n-gram*) bisa memberikan hasil yang lebih baik ketimbang representasi teks berbasis kata (*word n-gram*). |
| 5 | Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia (Arafat et al., 2020). | Melakukan analisis sentimenterhadap aspirasi masyarakat di media sosial Twitter dengan teknik *machine learning* untuk memberikan informasi dan menambah wawasan bagi akademisi dan praktisi. | (1) Melakukan 4 pemodelan yang berbeda untuk menggali pola demografi pendukung, yaitu: klasifikasi jenis kelamin, klasifikasi umur, klasifikasi lokasi, dan klasifikasi sentiment.  (2) Memanfaatkan atribut akun Twitter seperti nama lengkap, deskripsi akun yang digunakan untuk mengklasifikasikan jenis kelamin dan umur.  (3) Menciptakan beberapa fitur untuk mendukung proses klasifikasi yang informasinya tidak secara langsung disediakan oleh metadata Twitter, seperti: First\_male, Middle\_male, First\_female, dan Middle\_male. | (1) Tidak menjelaskan secara mendetail proses *text cleansing* pada tahapan *data preprocessing.*  (2) Penelitian ini tidak menjelaskan secara mendetail proses anotasi yang dilakukan.  (3) Penelitian tidak menyebutkan seberapa tinggi tingkat akurasi yang dicapai pada kasus klasifikasi sentimen. | Penelitian ini melakukan banyak eksperimentasi dengan membuat fitur-fitur yang digunakan untuk klasifikasi jenis kelamin, usia dan lokasi. Atribut pada akun yang disediakan oleh Twitter dapat digunakan untuk keperluan 3 klasifikasi tersebut. Penelitian ini juga menggunakan pendekatan analisis sentimen berbasis lexicon, yaitu *SentiStrength.* |
| ***Summarize:***   1. Kelima penelitian analisis sentimen pada analisis 3C2S ini telah mengadopsi kerangka kerja *Text Mining* mulai dari *data collection, text preprocessing, text representation, model building* sampai dengan *evaluation.* 2. Mayoritas penelitian telah menggunakan data teks opini yang berasal dari media sosial Twitter. 3. Untuk tahapan *text representation* 4 dari 5 penelitian sebelumnya telah mengadopsi teknik TF-IDF, baik dengan bentuk unigram, bigram maupun trigram. Dimana pendekatan *n-gram* memberikan hasil lebih baik dibandingkan unigram. 4. Untuk metode klasifikasi, algoritma pemodelan *classical machine learning* seperti *naïve bayes,* SVM, dan *decision tree* bersama dengan teknik pemodelan yang lebih modern seperti *deep learning* atau *multilayer perception* menjadi teknik yang sering digunakan. Studi literatur ini memberikan banyak masukan bagi penulis, seperti tidak selalu *machine learning* yang dianggap lebih modern seperti *multilayer perception* memberikan hasil lebih baik daripada *classical machine learning*, namun hal tersebut tampaknya juga terpengaruh dari arsitektur *multilayer perception* dan teknik representasi teks yang digunakan. Eksperimentasi seperti membentuk fitur baru berdasarkan teori atau rekomendasi penelitian sebelumnya bisa diadopsi untuk meningkatkan hasil akurasi prediksi. 5. Untuk tahapan *evaluation,* mayoritas penelitian menggunakan metrik *accuracy, precision, recall,* dan *f1-score*. Sedangkan untuk pendekatan eksperimentasi evaluasi 3 penelitian menggunakan *k-fold-cross-validation.* Selain itu mayoritas penelitian (3 penelitian) menggunakan perbandingan data latih dan data uji sebesar 70:30 persen. 6. Sedangkan untuk tujuan pemodelan topik, teknik seperti LDA dan frekuensi penggunaan token (*wordcloud* dan *scatterplot*) menjadi yang paling sering digunakan. | | | | | |

## Kerangka Penelitian

Berdasarkan studi lituratur yang telah dilakukan, peneliti menyusun kerangka penelitian yang dipresentasikan pada Gambar 2.5. Komponen pada kerangka penelitian disusun berdasarkan tahapan pada kerangka kerja *Text Mining*. Komponen (1) *data collection* menjelaskan proses pengumpulan data yang dilakukan akan menggunakan pendekatan *web scrapping* atau ekstraksi data yang berasal dari *platform* media sosial Twitter. Komponen (2) *text preprocessing* menjelaskan data opini masyarakat yang berupa *text* akan melalui tahapan penghapusan *links* dan kerakter *non-alphanumeric, tokenization, stemming, lemmatization, stopword removal, post tagging* dan *normalization*. Tujuannya adalah untuk membuat data *text* yang lebih konsisten agar dapat memberikan hasil prediksi yang lebih optimal dan menghemat sumber daya yang dibutuhkan dalam proses komputasi data.

Komponen (3) *text representation* menjelaskan bentuk terstruktur dari dataset yang awalnya berupa *text* agar dapat diteliti secara kuantitatif dengan formula matematika dan statistik. *Bag of Words* TF-IDF dalam bentuk unigram, bigram dan trigram menjadi pendekatan *text representation* yang dipilih. (4) Komponen *classification model* menjelaskan teknik *machine learning* yang dianggap lebih modern yaitu *artificial neural network* akan dikomparasikan dengan *classical machine learning* seperti SVM, *naïve bayes, decisicion tree* dan KNN. Model dengan tingkat akurasi paling tinggi dianggap mampu menjadi alat klasifikasi otomatis terhadap *unlabeled dataset*.

Komponen (5) evaluasi menjelaskan 3 metrik, yaitu: *precision, recall* dan *f1-score* yang digunakan untuk mengukur performa model *machine learning*, selain itu komponen ini juga menjelaskan strategi yang dapat dipilih dalam proses pemilahan data latih dan data uji. (6) Komponen *topic modelling* menjelaskan penelitian ini akan berupaya mengidentifikasi tema-tema aspirasi masyarakat terkait dengan penyelenggaraan Pemilihan Umum 2024 dengan menggunakan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA). (7) Komponen Tujuan Penelitian “Analisis sentimen dan pemodelan topik terhadap opini masyarakat terkait dengan penyelenggaraan Pemilu 2024” ini terhubung dengan semua komponen yang telah disebutkan sebelumnya. Komponen tujuan penelitian akan mempegaruhi kriteria dan parameter yang diperlukan pada komponen yang lain.

[relasi hubungan antar bagan]

A group of white rectangular boxes with black text

Description automatically generated

Gambar 2. 6 *Theoretical framework*

# BAB 3 METODOLOGI PENELITIAN

Pada bab ini penulis menjelaskan metodologi penelitian yang digunakan. Bab disusun ke dalam 2 sub bab, yaitu: Rancangan Penelitian dan Alur Penelitian.

## Rancangan Penelitian

Penelitian ini menggali opini masyarakat terkait dengan penyelenggaraan Pemilihan Umum 2024 di Indonesia dengan mengukur sentimen masyarakat dan mengidentifikasi aspirasi atau topik yang menjadi perhatian masyarakat terkait dengan kasus tersebut. Penelitian ini menggunakan pendekatan kuantitatif, dimana teknik perhitungan matematis dan statistik akan banyak dimanfaatkan (Recker, 2013). Perhitungan matematis dan statistik diimplementasikan dalam bentuk teknik *text preprocessing, text representation,* dan klasifikasi menggunakan *machine learning,* evaluasi model dan pemodelan topik.

Lebih spesifik lagi penelitian ini dapat diklasifikasikan sebagai penelitian eksprimental. Penelitian eksprimental berupaya mengidentifikasi hubungan sebab akibat dari variable-variable yang menjadi objek penelitian (Recker, 2013). Selain itu penelitian eksperimental menjawab tantangan yang muncul ketika menggunakan pendekatan metode penelitian observasi seperti survey, dimana dilakukan penyesuaian yang diperlukan ketika terdapat faktor pembeda yang muncul. Metode penelitian berdasarkan observasi dianggap tidak dapat mengindetifikasi adanya faktor pembeda yang muncul (Recker, 2013). Dengan menggunakan teknik *machine learning* memungkinkan penelitian ini untuk dapat melakukan penyesuaian terhadap bobot suatu fitur atau variable independen yang menjadi faktor pembeda. Table 3.1 menjelaskan secara singkat atribut-atribut rancangan penelitian.

Tabel 3. 1 Ringkasan Rancangan Penelitian

|  |  |  |
| --- | --- | --- |
| **No** | **Atribut Penelitian** | **Deskripsi** |
| 1 | Jenis Penelitian | Kuantitatif |
| 2 | Klasifikasi | Eksperimental |
| 3 | Tujuan Penelitian | Membentuk model untuk melakukan analisis sentimen dan menggali tema-tema aspirasi masyarakat terkait dengan penyelenggaraan Pemilihan Umum 2024. |
| 4 | Hasil Penelitian | Analisis deskriptif opini masyarakat terkait penyelenggaraan Pemilu 2024 yang dapat bermanfaat bagi pihak yang berkepentingan, seperti: (1) Penyelenggara Pemilu, (2) Pemerintah, (3) Konsultan Politik dan (4) Peserta Pemilu. |
| 5 | Analisis Data | Kuantitatif |
| 6 | Jenis Data | Data Primer |
| 7 | Metode Pengumpulan Data | *Web Scrapping* |
| 8 | Metode Pengolahan Data | *Text Preprocessing*  *Text Representation*  *Sentiment Classification Modelling*  *Evaluation*  *Topic Modelling* |
| 9 | Pendekatan Pengambilan Keputusan | Deduktif |
| 10 | Alat bantu Penelitian | *Google Collab, Jupyter Notebook* |
| 11 | Bahasa pemprograman | *Python* |

Penelitian ini menggunakan data primer yang dikumpulkan langsung oleh peneliti. Data yang dikumpulkan merupakan opini masyarakat terkait penyelenggaraan Pemilihan Umum 2024 dari media sosial Twitter*.* Dengan menggunakan teknik *web scrapping* sebagai cara untuk pengumpulan datanya. Metode pengolahan data seperti *tokenization, text cleansing, stopword removal, lemmatization, normalization* dilakukan pada tahapan *text preprocessing.* Teknik *bag of word* dengan TF-IDF unigram, bigram, dan trigram digunakan untuk *text representation.* Komparasi model *artificial neural network* dan *classical machine learning model* seperti: *naïve bayes*, SVM, *decision tree*, dan KNN dilakukan untuk dapat menentukan model yang paling layak digunakan untuk analisis sentimen. Proses komparasi dilakukan dengan mengevaluasi metrik *precision, recall*, *f1-score* dari masing-masing model *machine learning*. Model dengan tingkat akurasi terbaik akan digunakan untuk mengklasifikasikan jenis sentimen dari dataset secara keseluruhan. Selain analisis sentimen, dengan tujuan untuk dapat mengidentifikasi tema-tema aspirasi masyarakat, baik yang berasal dari sentimen positif maupun negatif, maka pemodelan topik juga dilakukan berdasarkan data yang dihasilkan dari proses klasifikasi secara keseluruhan. Pemodelan topik dilakukan dengan menggunakan metode *Latent Dirichlet Allocation* (LDA).

## Alur Penelitian

Pada penelitian ini, terdapat 10 tahap yang diawali dengan pengumpulan data sampai dengan penarikan kesimpulan. Berikut urutan dari 11 tahapan pada penelitian ini, yang dijelaskan secara lebih mendetail pada tabel 3.2.

Tabel 3. 2 Alur Penelitian

|  |  |
| --- | --- |
| 1. **Pengumpulan Data Awal** | |
| **Input** | Undang-Undang, Literatur terkait *Electoral Integrity*, Survey BPS, Data *post-factum* Bawaslu, Berita dan Pernyataan Ahli. |
| **Process** | Observasi, Studi dokumen dan literatur |
| **Output** | Ekspektasi dan Kenyataan |

|  |  |
| --- | --- |
| **Perumusan Masalah dan Penentuan Pertanyaan Penelitian** | |
| **Input** | Ekspektasi dan Kenyataan |
| **Process** | Analisa Kesenjangan |
| **Output** | *Problem Statement* dan Pertanyaan Penelitian |

|  |  |
| --- | --- |
| 1. **Studi Literatur** | |
| **Input** | *Problem Statement* dan Pertanyaan penelitian |
| **Process** | Studi literatur terkait teori, metode dan analisa 3C2S penelitian terdahulu |
| **Output** | Rujukan teori, Rujukan metode, Sintesa analisa 3C2S, Kerangka Teoritis. |

|  |  |
| --- | --- |
| 1. **Pengumpulan Data** | |
| **Input** | Pertanyaan penelitian, Kerangka Teoritis, *Keyword* pengumpulan data. |
| **Process** | *Web Scrapping* |
| **Output** | *Unlabeled dataset* berupa *tweets* |

|  |  |
| --- | --- |
| 1. **Pelabelan Data** | |
| **Input** | *Unlabeled dataset* berupa *tweets* |
| **Process** | Anotasi dan Evaluasi hasil pelabelan menggunakan *cohen kappa coefficient*. |
| **Output** | *Labeled Dataset* (dataset dengan *dependent variable* yang telah terdefinisikan) |

|  |  |
| --- | --- |
| 1. ***Text Preprocessing*** | |
| **Input** | *Labeled Dataset* |
| **Process** | *Tokenization, Text Cleansing, Stopword Removal, Lemmatization, Normalization,* |
| **Output** | *Cleansed tokenized dataset* |

|  |  |
| --- | --- |
| 1. ***Text Representation*** | |
| **Input** | *Cleansed tokenized dataset*. |
| **Process** | Vektorisasi ke dalam bentuk *Bag of words –* TF-IDF unigram, bigram, trigram. |
| **Output** | *Vector Space Model* TF-IDF unigram, bigram, trigram. |

|  |  |
| --- | --- |
| 1. **Klasifikasi Sentimen** | |
| **Input** | *Vector space model* TF-IDF unigram, bigram, trigram. |
| **Process** | 1. Pembentukan model dengan teknik *deep learning* dan *classical machine learning* seperti SVM, *naïve bayes, decision tree,* KNN menggunakan *training dataset*. 2. Klasifikasi *testing dataset* dan evaluasi hasil klasifikasi dengan metrik *precision, recall, f1-score.* 3. Klasifikasi sentimen untuk keseluruhan data dengan model terbaik (tingkat akurasi tertinggi). |
| **Output** | 1. Nilai metrik dari masing-masing model. 2. Model dengan tingkat akurasi tertinggi. 3. Hasil klasifikasi sentimen.berupa dataset positif, negatif dan netral. |

|  |  |
| --- | --- |
| 1. **Pemodelan Topik** | |
| **Input** | Dataset positif, negatif dan netral. |
| **Process** | Pembentukan pemodelan topik dengan *Latent Dirichlet Allocation* (LDA). |
| **Output** | Topik teridentifikasi untuk setiap jenis sentimen |

|  |  |
| --- | --- |
| 1. **Penarikan Kesimpulan** | |
| **Input** | Hasil klasifikasi sentimen, Topik teridentifikasi |
| **Process** | Penarikan kesimpulan dengan pendekatan deduktif. |
| **Output** | Kesimpulan |

### Pengumpulan Data Awal

Penelitian ini diawali dengan tahapan pengumpulan data awal, dimana penulis mengumpulkan dan mempelajari dokumen dan literatur seperti Undang-Undang terkait Pemilu, Publikasi Ilmiah terkait dengan *electoral integrity,* Laporan hasil penelitian yang dilakukan oleh Bawaslu dan BPS. Kemudian observasi juga dilakukan terhadap berita dan pernyataan ahli terkait dengan penyelenggaraan pemilu 2024. Dari observasi, studi literatur dan dokumen yang telah disebutkan, maka ditemukanlah harapan dan kenyataan. Bahwa diharapkan terselenggaranya Pemilu yang luber jurdil dan berintegritas, namun kenyataannya terdapat indikasi-indikasi yang menurut para ahli, dapat menciptakan stigma dan persepsi buruk di mata masyarakat terhadap penyelenggaraan Pemilu 2024.

### Perumusan Masalah dan Penentuan Pertanyaan Penelitian

Pada tahapan ini peneliti berupaya untuk dapat mengidentifikasi permasalahan. Maka dilakukan analisa kesenjangan berdasarkan harapan dan kenyataan yang diperoleh pada tahapan sebelumnya. Hasil dari analisa kesenjangan yang berupa masalah yang menjadi fokus dalam penelitian ini adalah “Dibutuhkan pengukuran untuk dapat mengetahui sejauh mana sentimen positif dan negatif dari opini masyarakat terhadap penyelenggaraan Pemilu 2024”. Kemudian berdasarkan masalah yang telah teridentifikasi peneliti berupaya merumuskan pertanyaan, tujuan, cakupan,dan manfaat penelitian.

### Studi Literatur

Pada tahapan ini peneliti berupaya untuk menyusun *theoretical framework* yang selaras dengan pertanyaan penelitian. Dua database literatur yaitu: Scopus dan Lib UI menjadi sumber dalam pencarian rujuan teori dan metode terkait Analisis Sentimen dan Pemodelan Topik. Selain Scopus dan UI Lib, peneliti juga menggunakan database literatur seperti ACM, Google Scholar dan SpringerLink untuk mencari rujukan yang berasal dari literatur pada 2 database sebelumnya. Peneliti juga melakukan analisa 3C2S (*criticize, compare, contrast, summarize, synthesize*) dari 5 penelitian yang memiliki kesamaan dengan topik penelitian ini yaitu analisis sentimen dengan memanfaatkan teknik pemodelan *machine learning* yang didalamnya membahas *public policy* dan *political science*.

### Pengumpulan Data

Pada tahapan ini peneliti berupaya mengumpulkan data opini masyarakat terkait dengan penyelenggaraan Pemilu 2024 yang disampaikan melalui media sosial Twitter. Dengan menggunakan pendekatan *web scrapping*, proses penarikan data dilakukan pada beberapa tahapan Pemilu 2024, yaitu: masa kempanye, masa tenang dan masa pemungutan-rekapitulasi suara. Dimana ketiga tahapan tersebut berlangsung selama periode 21-Januari 2024 sampai dengan 20-Maret-2024. Digunakan *keyword* (‘Pemilu’ AND (‘Luber’ OR ‘Jurdil’ OR ‘Jujur’ OR ‘Adil’ OR ‘Demokratis’ OR ‘Curang’), untuk dapat memperoleh data yang sesuai harapan yaitu *tweet* opini masyarakat terkait dengan penyelenggaraan Pemilu 2024. Tabel 3.3 menunjukkan contoh opini masyarakat terkait dengan Pemilu yang mereka utarakan di media sosial Twitter.

Tabel 3. 3 Contoh opini masyarakat tentang pemilu di media sosial

|  |  |  |
| --- | --- | --- |
| **No** | **Contoh Opini Masyarakat di Twitter** | **Jenis Sentimen** |
| 1 | Siapa bilang KPU berisikan parpol gak bisa jurdil & demokrasi?  Referensi bisa dari Sistem Pemilu di Jerman. Anggota KPU nya perwalilan partai politik.  Refly Harun orang yg menolak teriak paling keras KPU dari parpol, sekarang demo tuh teriak2 pemilu curang.  #Pemilu2024  #dapse | Positif |
| 2 | Nyesek banget, kecurangan di pelupuk mata gak tampak...ya Allah...jauhkan kami dari azabMU.  Kami rakyat kecil hanya bisa berjuang dengan jari kami, ampuni jika itu belum termasuk perjuangan nyata...hiks  #pemiluterburuk #pemilucurang #pemilu2024 | Negatif |

### Pelabelan Data

Pada tahapan ini peneliti berupaya untuk dapat membangun korpus baik *training set* maupun *testing set*. Penelitian ini mengadopsi pendekatan *supervised machine learning,* dimana setiap *instance* yang akan dilatih membutuhkan tidak hanya *independent variable,* namun juga *dependent variable* yang dalam penelitian ini adalah jenis sentimen*.* Hasil dari tahapan pegumpulan data, nilai dari *dependent variable* belum ditentukan. Oleh karena itu dilakukan pelabelan untuk menentukan nilai dari *dependent variable*. Kelas sentimen merupakan repersentasi dari *dependent variable* pada penelitian ini, dimana nilai nominalnya adalah (1) positif, (2) negatif, (3) netral, (4) tidak relevan. Kemudian untuk mencegah munculnya masalah subjektifitas, maka dilakukan evaluasi hasil pelabelan dengan menggunakan *cohen kappa coefficient*. Ketika nilai evaluasi hasil anotasi menunjukan nilai koefisien *kappa* yang kurang baik maka 2 *annotator* dari sebagian data dengan hasil pelabelan berbeda akan berdiskusi dan menentukan jenis sentimen yang mereka sepakati.

### *Text Preprocessing*

Pada tahapan ini peneliti berupaya untuk membuat data teks menjadi lebih konsisten sebelum data tersebut ditransformasikan ke dalam bentuk yang lebih terstruktur. Oleh karena itu dilakukan beberapa subproses seperti *tokenization, text cleansing, stopword removal, lemmatization, normalization.* *Tokenization* dilakukan untuk mengubah data teks yang awalnya berbentuk kalimat atau paragraf menjadi kumpulan kata-kata. Hasil tokenisasi yang berupa daftar kata dijadikan parameter untuk subproses selanjutnya, sebelum pada akhirnya akan merepresntasikan sebuah fitur.

Kemudian dilakukan *text cleansing* yang merupakan upaya dalam menghilangkan konten pada teks yang tidak relevan dalam pembentukan fitur seperti URL, karakter spesial, angka dan sebagainya. *Stopword removal* merupakan upaya untuk menghilangkan kata yang tidak signifikan sesuai dengan tujuan analisis. *Lemmatization* memiliki tujuan yang sama dengan *stemming* namun dengan mekanisme yang lebih kompleks dengan analisa morfologi[[18]](#footnote-19) (seluk beluk kata) atau mendeteksi sinonim kata, seperti contoh kata dalam bahasa inggris “ate” menjadi “eat”. Terakhir dilakukan normalisasi token yang merupakan upaya untuk mentransformasi kata informal, yang menimbulkan kebisingan, ke dalam bentuk versi kata yang lebih formal atau resmi.

### *Text Representation*

Pada tahapan ini peneliti berupaya untuk mengubah data teks yang telah konsisten ke dalam bentuk terstruktur, sehingga dapat diolah ke dalam algoritma *machine learning*. Bentuk terstruktur pada penelitian ini adalah *Vector Space Model* dengan 2 sumbu atau dalam bentuk matrix, sumbu 1 merepresentasikan daftar *instance* dan sumbu 2 merepresntasikan fitur pada korpus*.* Pada penelitian ini teknik TF-IDF unigram, bigram dan trigram digunakan untuk menentukan nilai fitur. Penjelasan terkait TF-IDF dapat ditemukan pada bagian 2.5.3 *Text Representation.*

### Klasifikasi Sentimen

Pada tahapan ini peneliti berupaya untuk membentuk model klasifikasi dengan menggunakan teknik *machine learning* seperti *artificial neural network* dan *classical machine learning* seperti SVM, *naïve bayes, decision tree,* dan KNN. Hasilnya akan ada 10 model yang terbentuk, dimana masing-masing algoritma *machine learning* yang berjumlah 5 akan menghasilkan 2 model, yaitu: (1) Model yang dipergunakan untuk mengklasifikasikan data relevan dan tidak relevan dan (2) Model yang dipergunakan untuk mengklasifikan sentimen ke dalam 3 nilai nominal yaitu positif, negatif atau netral.

Sementara pembagian dataset dilakukan dengan konfigurasi 70 persen data untuk proses pelatihan dan 30 persen untuk proses pengujian. Kemudian dilanjutkan dengan memilih model terbaik dari 5 model yang terbentuk dari masing-masing jenis klasifikasi, yaitu klasifikasi relevansi data dan klasifikasi sentimen. Penelitian ini menggunakan mekanisme pengujian 10-*fold cross validation* dan pemanfaatan *confusion matrix* untuk menghitung nilai *precision, recall* dan *f1-score* dari masing-masing model. Nilai metrik kemudian dikomparasikan dan dipilih model terbaik bagi klasifikasi sentimen dan klasifikasi relevansi data. Kemudian dengan menggunakan model terbaik akan dilakukan klasifikasi terhadap dataset secara keseluruhan.

### Pemodelan Topik

Pada tahapan ini peneliti berupaya untuk mengidentifikasi topik pada masing-masing kategori sentimen dengan menggunakan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA). Topik yang teridentifikasi memberikan gambaran akan aspirasi masyarakat yang mereka ungkapkan di media sosial, baik yang berasal dari opini bersentimen positif dan opini bersentimen negatif.

### Penarikan Kesimpulan

Pada tahapan ini peneliti berupaya untuk mempresentasikan hasil penelitian yang dilakukan ke dalam bentuk laporan penelitian. Penyusunan laporan penelitian dilakukan berdasarkan pendekatan penarikan kesimpulan deduktif. Analisis sentimen pada penelitian ini memanfaatkan model *machine learning*. Model *machine learning* dapat dianalogikan sebagai aturan umum yang akan menyimpulkan *sample* yang sifatnya khusus, dalam kasus penelitian ini *sample* yang sifatnya khusus adalah *tweet* yang akan diklasifikasikan.

# BAB 4 IMPLEMENTASI

Bab ini akan menjelaskan tahapan-tahapan implementasi pada penelitian ini yang disusun ke dalam kerangka kerja *Text Mining*. Terdapat 4 tahapan pada bagian ini diantaranya adalah Pembentukan Korpus, *Text Preprocessing*, *Text Representation* dan *Knowledge Discovery*.

### Tahapan Pembentukan Korpus

Pada bagian ini peneliti akan menjelaskan urutan proses dalam membentuk korpus. Penjelasan akan mencakup (1) bagaimana peneliti mengumpulkan data, (2) apa saja atribut atau struktur data yang akan digunakan dalam penelitian dan (3) proses pelabelan dan evaluasi hasil pelabelan dalam mendefinisikan *dependent variable*.

#### Pengumpulan Data

Proses pengumpulan data mengadopsi Teknik *web scrapping* dengan menggunakan *tweet-harvest* dari Helmi Satria yang merupakan instrumen berbasis NodeJS[[19]](#footnote-20). *Web scrapping* menirukan aktifitas penelusuran dengan aplikasi *browser*. Dimana penelusuran dilakukan pada halaman demi halaman media sosial Twitter yang berisi komentar atau *tweet* berdasarkan pencarian dengan kata kunci tertentu. *Tweet* yang ditemukan pada proses penelusuran tersebut kemudian diekstrak ke dalam file *comma separated value* (CSV) secara otomatis. Peneliti menggunakan 2 parameter, yaitu: (1) kata kunci dan (2) waktu pembuatan *tweet* dan (3) dan batas jumlah *tweet* yang diambil*.* Gambar 4.1 menunjukkan contoh parameter yang digunakan untuk pencarian.



Gambar 4 1 Kata kunci pencarian *web scrapping*

Parameter (1) kata kunci digunakan untuk menemukan *tweet* yang relevan dengan topik penelitian. Kemudian parameter (2) waktu pembuatan *tweet* memungkinkan peneliti untuk melakukan penelusuran terhadap *tweet* yang telah lama diunggah. Sementara parameter (3) batas jumlah *tweet* digunakan untuk membatasi jumlah pencarian, hal ini dilakukan karena pendekatan *web scrapping* memiliki kekurangan yaitu keterbatasan kapasitas *random access memory* (RAM)yang digunakan untuk menjalankan aplikasi *browser* untuk melakukan penelusuran*.* Pencarian dilakukan terhadap *tweet* yang dibuat pada rentang waktu 21-Januari-2024 sampai dengan 30-Maret-2024, dimana rentang waktu tersebut mencakup 4 periode masa pemilu, yaitu masa kampanye, masa tenang, masa pemungutan dan perhitungan suara dan masa ajudikasi sampai dengan penetapan. Gambar 4.2 menunjukkan kode program yang digunakan untuk melakukan pengumpulan data.

A screenshot of a computer code

Description automatically generated

Gambar 4 2 Potongan kode pengumpulan data

Pada potongan kode diatas terlihat untuk dapat menjalankan instrumen *tweet-harvest* dibutuhkan *authentication token* (authToken) yang merupakan parameter ke-3 dan didapat pada bagian *cookie browser* ketika seorang user telah login ke dalam aplikasi Twitter. Pada akhir bagian ini dilakukan penyaringan untuk menghapus data opini yang duplikat. Kemudian dilanjutkan dengan mengkonversi file csv ke dalam bentuk file excel sehingga dapat mudah digunakan dalam proses anotasi yang dilakukan dengan menggunakan instrumen *google spreadsheet*. Gambar 4.3 menunjukkan kode program yang digunakan untuk melakukan menghapus data duplikat dan konversi ke file excel.

A screen shot of a computer code

Description automatically generated

Gambar 4 3 Potongan kode penyaringan data duplikat

#### Struktur Data

Instrumen *web scrapping* yang digunakan dalam penelitian ini menghasilkan beberapa atribut yang merepresentasikan *tweet*. Penelitian ini hanya memanfaatkan sebagian dari atribut yang tersedia, sebagimana ditunjukkan pada tabel 4.1.

Tabel 4 1 Struktur data hasil dari web scrapping

|  |  |  |
| --- | --- | --- |
| **Atribut** | **Deskripsi** | **Contoh** |
| Created\_at | Nilai berformat *timestamp* GMT +7*,* yang merupakan kapan *tweet* dibuat. Nilai pada atribut ini digunakan untuk mengidentifikasi 4 masa periode tahapan pemilu yang akan digunakan untuk analisa tren sentimen dari masing-masing periode masa. |  |
| id | Nilai berformat number, yang merupakan identifier unik dari masing-masing *tweet.* |  |
| Full\_text | Nilai berformat teks, merupakan opini terkait penyelenggaraan pemilu yang akan dianalisa. |  |

#### Pelabelan Data

Untuk dapat menjalankan tugas prediksi klasifikasi dengan menggunakan pendekatan *supervised machine learning*, dibutuhkan data latih dan data uji dengan nilai variable dependen atau atribut kelas yang sudah terdefinisi. Pada data latih, nilai atribut kelas yang sudah terdefinisi akan digunakan dalam upaya pembelajaran dan pembentukan model prediksi klasifikasi. Sedangkan pada data uji, nilai tersebut digunakan untuk mengevaluasi seberapa akurat hasil prediksi yang dilakukan oleh model. Maka untuk dapat membentuk data latih dan data uji dengan nilai atribut kelas yang sudah terdefinisi dibutuhkan usaha untuk menganotasikan *tweet* secara manual.

Dalam penelitian ini, jenis sentimen menjadi atribut kelas, dimana nilai nominalnya adalah positif, negatif, netral dan tidak relevan. Proses anaotasi dilakukan dengan menggunakan bantuan instrumen *google spreadsheet.* Untuk satu *tweet* peneliti melibatkan 8 *orang annotator* dimana 2 orang akan ditugaskan untuk mengklasifikasikan setiap *tweet* secara manual.

Proses anotasi diawali dengan penulis melakukan pelabelan awal untuk mengidentifikasi karakteristik dari masing-masing jenis sentimen. Pelabelan awal dilakukan terhadap 500 *tweet.* Kemudian berdasarkan karakteristik jenis sentimen yang teridentifikasi, peneliti menjabarkan secara mendetail karakteristik dari masing-masing jenis sentimen ke dalam dokumen panduan pelabelan (Lampiran 1). Dokumen panduan pelabelan menjadi rujukan bagi annotator dalam melakukan tugas klasifikasi manual. Dimana selama proses anotasi berlangsung peneliti secara rutin berkoordinasi dengan para *annotator* dengan bantuan aplikasi *google meet*.

Kemudian dilakukan evaluasi hasil anotasi dengan menggunakan koefisien *kappa*. Ketika nilai evaluasi hasil anotasi menunjukan nilai koefisien *kappa* yang kurang baik maka 2 *annotator* dari sebagian data dengan hasil pelabelan berbeda akan berdiskusi dan menentukan jenis sentimen yang mereka sepakati. Hal ini dilakukan sampai nilai evaluasi koefisien *kappa* menunjukan ambang batas yang diharapkan yaitu“*Good Agreement*” (sebagai mana dijelaskan pada Bab 2)*.* Tabel 4.2 menunjukkan nilai evaluasi hasil anotasi dan tabel 4.3 menunjukkan komposisi jumlah dari masing-masing jenis sentimen pada hasil pelabelan terakhir.

Tabel 4 2 Hasil evaluasi proses pelabelan.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | ***Annotator* 2** | | | |  |
| TR | Positif | Netral | Negatif |
| ***Annotator* 1** | TR | 5000 | 0 | 0 | 0 | 1.0 |
| Positif | 0 | 2000 | 0 | 0 | 1.0 |
| Netral | 0 | 0 | 1000 | 0 | 1.0 |
| Negatif | 0 | 0 | 0 | 2000 | 1.0 |
|  | | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| N | 10000 | | | | | |
| Po | 1.0 | | | | | |
| Pe | 0.0 | | | | | |
| *Kappa* | 1.0 | | | | | |

Tabel 4 3 Komposisi jumlah dataset hasil anotasi.

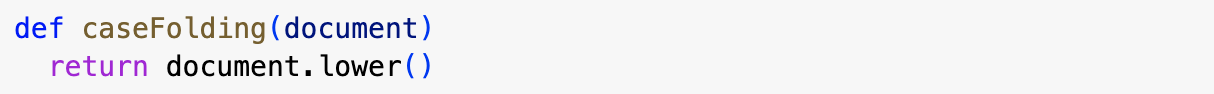
|  |  |
| --- | --- |
| **Jenis Sentimen** | **Jumlah *tweet*** |
| Tidak Relevan | 5000 |
| Relevan | 5000 |
| Positif | 2000 |
| Netral | 1000 |
| Negatif | 2000 |
| **Total** | **10000** |

### Tahapan *Text Preprocessing*

Tahapan ini dilakukan dengan tujuan agar data menjadi lebih konsisten sebelum diubah ke dalam bentuk *vector space model* yang menjadi dasar pembentukan model *machine learning*. Data yang lebih konsisten diharapkan dapat memberikan hasil klasifikasi yang lebih akurat, selain itu jumlah data yang terreduksi dapat menghemat penggunaan sumber daya pada saat proses komputasi dan penyimpanan. Sehingga proses dapat berjalan lebih cepat dan membutuhkan kapasitas RAM dan *storage* yang lebih sedikit.

#### *Case Folding*

Proses *case folding* menjadi penting dilakukan karena tidak ada standar baku yang harus ditaati pengguna dalam menuliskan cuitan. Pengguna dapat menggunakan huruf besar dan huruf kecil yang mereka inginkan dalam menuliskan cuitan. Hal ini jika tidak distandarisasi akan membentuk lebih dari satu token yang sebetulnya merepresentasikan satu kata yang sama. Gambar 4.4 menunjukkan potongan kode yang digunakan untuk mengubah teks menjadi huruf kecil.



Gambar 4 4 Potongan kode proses *case folding*.

Proses ini menggunakan fungsi .lower() dari bahasa pemprograman python. Tabel x menunjukkan perbandingan antara input dan output dari proses ini.

Tabel 4 4 Perbedaan input dan output dari proses *case folding*.

|  |  |
| --- | --- |
| **Input** | **Output** |
| Pokoknya!!! <h1>Netijen+62 Toplah</h1> 👍 kawal terus Pemilu yang jujur hingga pecundang yang ingin curang menjerit jerit karena kehilangan ruang gerak [http://pemiluindonesia.com/kawal-pemilu-ber](http://pemiluindonesia.com/kawal-pemilu-berintegritas) | pokoknya!!! <h1>netijen+62 toplah</h1> 👍 kawal terus pemilu yang jujur hingga pecundang yang ingin curang menjerit jerit karena kehilangan ruang gerak [http://pemiluindonesia.com/kawal-pemilu-ber](http://pemiluindonesia.com/kawal-pemilu-berintegritas) |

#### *Tokenization*

Proses ini dilakukan untuk mengubah teks menjadi daftar unit terkecil yaitu daftar kata yang disebut token. Token yang dihasilkan akan dijadikan input untuk proses selanjutnya yaitu mengurangi *noise*, menghapus *stopword*, menghilangkan imbuhan danidentifikasi sinonim kata (*lemmatization*). Gambar 4.4 menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

A computer screen shot of a computer code

Description automatically generated

Gambar 4 5 Potongan kode tokenisasi

Proses ini memanfaatkan fungsi regex dari bahasa pemograman python untuk mengidentifikasi setiap token. Tabel 4.5 menunjukkan perbandingan input dan output dari proses ini.

Tabel 4 5 Perbedaan input dan output dari proses tokenisasi.

|  |  |
| --- | --- |
| **Input** | **Output** |
| pokoknya!!! <h1>netijen+62 toplah</h1> 👍 kawal terus pemilu yang jujur hingga pecundang yang ingin curang menjerit jerit karena kehilangan ruang gerak [http://pemiluindonesia.com/kawal-pemilu-ber](http://pemiluindonesia.com/kawal-pemilu-berintegritas) | ['pokoknya', '!', '!', '!', '<h1>', 'netijen', '+', '62', 'toplah', '</h1>', '👍', 'kawal', 'terus', 'pemilu', 'yang', 'jujur', 'hingga', 'pecundang', 'yang', 'ingin', 'curang', 'menjerit', 'jerit', 'karena', 'kehilangan', 'ruang', 'gerak', '[http://pemiluindonesia.com/kawal-pemilu-ber](http://pemiluindonesia.com/kawal-pemilu-berintegritas)'] |

#### *Text Cleansing*

Proses ini diperlukan untuk meminimalisir kebisingan dalam menganalisa token yang disebabkan oleh keberadaan URL, tag-html dan karakter yang tidak signifikan keberadaannya dalam menentukan hasil klasifikasi *machine learning*. Karakter seperti tanda baca (*punctuation*), dan angka. Kemudian pada penelitian ini dimana ditujukan untuk menggali wawasan yang berasal dari opini masyarakat terkait dengan penyelenggaraan Pemilu 2024, maka ada beberapa kombinasi angka yang diharapkan dapat memberikan wawasan pada hasil penelitian. Seperti 1,2,3,01,02,03 menjadi kombinasi angka yang merepresentasikan nomor urut kontestan pemilu pasangan calon presiden dan wakil presiden. Maka dibutuhkan pengecualian untuk beberapa kombinasi angka tersebut. Gambar 4.5 menunjukkan potongan kode yang digunakan untuk mengeksekusi proses ini. [hapus regex pendeteksi emoticon dari gambar]

A computer code with text

Description automatically generated with medium confidence

Gambar 4 6 Potongan kode menghapus URL, panctuation, tag-html dan angka

Proses ini menggunakan *regex* dari Bahasa pemprograman python untuk mengidentifikasi konten noise pada teks. Tabel 4.6 menunjukkan perbandingan antara input token dan output token setelah menjalankan proses ini.

Tabel 4 6 Perbandingan input dan output dari proses *text cleansing*.

|  |  |
| --- | --- |
| Input | Output |
| ['pokoknya', '!', '!', '!', '<h1>', 'netijen', '+', '62', 'toplah', '</h1>', '👍', 'kawal', 'terus', 'pemilu', 'yang', 'jujur', 'hingga', 'pecundang', 'yang', 'ingin', 'curang', 'menjerit', 'jerit', 'karena', 'kehilangan', 'ruang', 'gerak', '[http://pemiluindonesia.com/kawal-pemilu-ber](http://pemiluindonesia.com/kawal-pemilu-berintegritas)'] | ['pokoknya', 'netijen', 'toplah', ':thumbs\_up:', 'kawal', 'terus', 'pemilu', 'yang', 'jujur', 'hingga', 'pecundang', 'yang', 'ingin', 'curang', 'menjerit', 'jerit', 'karena', 'kehilangan', 'ruang', 'gerak'] |

#### *Stopword Removal*

Proses ini dilakukan untuk mengurangi kebisingan pada teks yang disebabkan oleh keberadaan dari token yang tidak memiliki signifikansi dalam menentukan hasil penilaian klasifikasi atau yang disebut *stopword.* Dengan menggunakan token hasil dari proses tokenisasi, proses ini menggunakan bahasa pemprograman python dengan bantuan *library nlp\_id*. Gambar 4.6 menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

A close-up of a computer code

Description automatically generated

Gambar 4 7 Potongan kodemenghapus *stopword*

Tabel 3.4 menunjukkan perbedaan input token dan output token seteleh menjalankan proses ini. Terdapat beberapa token yang dihapus, diantaranya adalah: ‘terus’, ‘yang’, ‘ingin’, ‘karena’.

Tabel 4 7 Perbandingan input dan output proses *stopword removal.*

|  |  |
| --- | --- |
| Input | Output |
| ['pokoknya', 'netijen', 'toplah', ':thumbs\_up:', 'kawal', 'terus', 'pemilu', 'yang', 'jujur', 'hingga', 'pecundang', 'yang', 'ingin', 'curang', 'menjerit', 'jerit', 'karena', 'kehilangan', 'ruang', 'gerak'] | ['pokoknya', 'netijen', 'toplah', ':thumbs\_up:', 'kawal', 'pemilu', 'jujur', 'hingga', 'pecundang', 'curang', 'menjerit', 'jerit', 'kehilangan', 'ruang', 'gerak'] |

#### *Lematization*

Proses ini bertujuan untuk mereduksi jumlah token dengan mengidentifikasi token dengan makna yang sama namun dengan representasi yang berbeda dan mengtransformasikannya agar lebih standar atau konsisten. Hal ini akan membuat nilai yang merepresentasikan token menjadi lebih aktual dan pada akhirnya dapat meningkatkan performa klasifikasi. Proses ini menggunakan bahasa pemprograman python dengan bantuan library *nlp\_id*. Gambar 4.7 menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

A close-up of a computer screen

Description automatically generated

Gambar 4 8 Potongan kode proses lemmatisasi*.*

Tabel 4.8 menunjukkan perbedaan token input dan output yang dihasilkan setelah menjalankan proses ini. Terdapat beberapa token yang ditransformasi setelah menjalankan proses ini, diantaranya adalah: ‘top’ menjadi ‘toplah’, ‘pemilu’ menjadi ‘milu’, ‘pecundang’ menjadi ‘cundang’, ‘menjerit’ menjadi ‘jerit’ dan ‘kehilangan’ menjadi ‘hilang’.

Tabel 4 8 Perbedaan input dan output dari proses lematisasi.

|  |  |
| --- | --- |
| **Input** | **Output** |
| ['pokoknya', 'netijen', 'toplah', ':thumbs\_up:', 'kawal', 'pemilu', 'jujur', 'hingga', 'pecundang', 'curang', 'menjerit', 'jerit', 'kehilangan', 'ruang', 'gerak'] | ['pokok', 'netijen', 'top', 'thumbs up', 'kawal', 'milu', 'jujur', 'hingga', 'cundang', 'curang', 'jerit', 'jerit', 'hilang', 'ruang', 'gerak'] |

#### *Normalization*

Proses ini mentransformasi token yang merepresentasikan leksikal yang bervariasi atau informal ke dalam bentuk formal atau resminya. Untuk itu proses ini menggunakan kamus normalisasi leksikal yang dikelola oleh Salsabila[[20]](#footnote-21). Selain itu proses ini juga menyasar token dengan pengulanggan karakter. Gambar x menunjukkan potongan kode yang digunakan dalam menjalankan proses ini. Proses transformasi untuk token dengan pengulanggan karakter dilakukan dengan menggunakan *regex* dalam bahasa pemprograman *python.* Sedangkan untuk mentransformasi token berdasarkan kamus normalisasi leksikal, digunakan library *pandas* untuk memuat data*.* Tabel x menunjukkan perbedaan token input dan output yang dihasilkan setelah menjalankan proses ini.

A computer screen shot of a code

Description automatically generated

Gambar 4 9Potongan kode normalisasi token*.*

Tabel 4 9 Perbedaan input dan output dari proses normalisasi teks.

|  |  |  |
| --- | --- | --- |
| **Teks Asli** | **Input** | **Output** |
| Pemilu ini adlh pemilu yg tdk mencerminkan kehendak rakyat, curranggggg seeemuuuaaa!!! | ['milu', 'adlh', 'milu', 'tdk', 'cermin', 'hendak', 'rakyat', 'curranggggg', 'seeemuuuaaa'] | ['milu', 'adalah', 'milu', 'tidak', 'cermin', 'hendak', 'rakyat', 'curang', 'semua'] |

### Tahapan *Text Representation*

Pada tahapan ini, akan dibentuk *vector space model* atau matrik *n x m* yang merepresentasikan daftar cuitan pada dataset dengan fitur-fitur berserta nilai yang dimiliki. Dimana nilai *n* pada matrik adalah jumlah *instance* atau cuitan pada korpus. Dan nilai *m* pada matriks adalah jumlah fitur yang merupakan keseluruhan token pada korpus hasil tahapan sebelumnya (*text* preprocessing). Setiap entry pada *vector space model* merepresentasikan nilai fitur*.* Penelitian ini menggunakan TF-IDF dengan konfigurasi unigram, bigram, trigram dan kombinasi ketiganya untuk menentukan nilai dari setiap fitur.Proses ini menggunakan bahasa pemograman *python* dengan bantuan library *sklearn.* Gambar 4.x menunjukkan potongan kode yang digunakan untuk mengeksekusi proses ini.

A screen shot of a computer

Description automatically generated

Gambar 4 10 Potongan kode untuk *text representation* dengan TF-IDF

Fungsi TfidfVectorizer dari sklern membutuhkan satu parameter yaitu n\_gram\_range. Nilai ngram\_range menentukan konfigurasi bentuk TF-IDF apakah unigram, bigram, trigram atau kombinasi ketiganya. Tabel x menunjukkan perbedaan input dan output dari proses *text representation*.

Tabel 4 10 Perbedaan input dan output proses *text representation*.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Contoh Input | | | | | Hasil *Text Preprocessing* | | | |
| Instance 1: pemilu kali ini sulit dikatakan pemilu yang luber jurdil | | | | | milu', 'sulit', 'milu', 'luber', 'jurdil' | | | |
| Instance 2: pemilu berlangsung aman, damai dan luber jurdil | | | | | milu', 'aman', 'damai', 'luber', 'jurdil' | | | |
| *Text Representation* | | | | | | | | |
| TF-IDF Unigram | | | TF-IDF Bigram | | | TF-IDF Trigram | | |
| Ins 1 | Ins 2 | Fitur | Ins 1 | Ins 2 | Fitur | Ins 1 | Ins 2 | Fitur |
| 0.000 | 0.533 | aman | 0.000 | 0.534 | aman damai | 0.000 | 0.577 | aman damai luber |
| 0.000 | 0.533 | damai | 0.000 | 0.534 | damai luber | 0.000 | 0.577 | damai luber jurdil |
| 0.354 | 0.379 | jurdil | 0.380 | 0.380 | luber jurdil | 0.000 | 0.577 | milu aman damai |
| 0.354 | 0.379 | luber | 0.000 | 0.534 | milu aman | 0.577 | 0.000 | milu luber jurdil |
| 0.708 | 0.379 | milu | 0.534 | 0.000 | milu luber | 0.577 | 0.000 | milu sulit milu |
| 0.498 | 0.000 | sulit | 0.534 | 0.000 | milu sulit | 0.577 | 0.000 | sulit milu luber |
|  |  |  | 0.534 | 0.000 | sulit milu |  |  |  |

### Tahapan *Knowledge Discovery*

Secara garis besar akan ada 2 proses yang dilakukan pada tahapan ini, yaitu klasifikasi sentimen dan pemodelan topik. Hal ini dilakukan dalam upaya menemukan pola menarik yang memiliki nilai bisnis atau wawasan yang berharga.

#### Tahapan Klasifikasi Sentimen

Tahapan klasifikasi sentimen secara garis besar dapat dibagi menjadi 2, yaitu: Identifikasi Model Terbaik dan Pemanfaatan Model Terbaik. Berikut adalah penjelasan dari kedua tahapan tersebut.

##### Identifikasi Model Terbaik

Penelitian ini menggunakan pendekatan *supervised machine learning* dan mengkomparasikan 4 model, yaitu: *artificial neural network, random forest,* SVM dan *logistic regression*. Sebelum mengkomparasikan kelima model tersebut peneliti berupaya untuk mengidentifikasi hyperparameter paling optimal dari masing-masing algoritma. Oleh karena itu peneliti menggunakan modul *HalvingGridSearchCV* dari *sklearn. HalvingGridSeachCV* akan membagi proses evaluasi ke dalam beberapa jenjang dan hanya sebagian kandidat dengan performa terbaik akan kembali dievaluasi pada jenjang berikutnya.

Dibandingkan dengan GridSearchCV, *HalvingGridSeachCV* memberikan waktu hyperparameter tuning yang lebih cepat [[21]](#footnote-22). Hal ini dikarenakan pendekatan ini pada awalnya hanya menggunakan sebagian dataset, yang jumlahnya akan bertambah secara bertahap pada jenjang berikutnya. Selain *hyperparameter* dari masing-masing algoritma, dalam pencarian menggunakan *HalvingGridSearchCV* kombinasi vektorisasi TFIDF yang berbeda (unigram, bigram, trigram, dan kombinasi ketiganya) juga akan dievaluasi. Kemudian untuk mengatasi permasalah *oversampling* atau data tidak seimbang, penelitian ini menggunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE), yang membuat secara sintetis *instance* baru bagi kategori atau label minoritas.

Sedangkan untuk perbandingan data latih dan data uji sebesar adalah sebesar 70 persen berbanding 30 persen. Kemudian pencarian juga mengadopsi pendekatan evaluasi 5-*fold cross validation,* dimana masing-masing kandidat akan dievaluasi sebanyak 5 kali dan akan diuturkan berdasarkan nilai rata-rata dari metrik yang ditentukan (*f1-score*). Gambar 4.x menunjukkan potongan kode yang digunakan pada proses *hyperparameter tunning*. Sedangkan Gambar x-x menunjukkan potongan kode fungsi umum yang digunakan untuk menjalankan hyperparameter tuning untuk masing-masing algoritma *machine learning*.

A screen shot of a computer code

Description automatically generated

Gambar 4 11 Potongan kode proses *hyperparamter tuning.*

Selain konfigurasi umum yang sudah dijelaskan pada bagian sebelumnya, setiap algoritma memiliki parameter dan cara evaluasi yang berbeda. Hal ini dikarenakan perbedaan penerapan kombinasi dan perbedaan waktu training yang dibutuhkan oleh setiap algoritma *machine learning*. Sehingga proses pencarian untuk masing-masing algoritma *machine learning* harus dijalankan secara terpisah*.* Berikut adalah penjelasan strategi hyperparameter tunning dari masing-masing algoritma yang diujicobakan.

4.4.1.1.1 Evaluasi Algoritma SVM

Terdapat 3 hyperparameter dari algoritma SVM yang coba diidentifikasi nilai paling optimalnya dalam menghasilkan prediksi paling akurat, yaitu: C, gamma dan karnel. Tabel x menjelaskan 3 parameter tersebut berserta dengan daftar nilai dari masing-masing parameter yang diujicobakan. Pemilihan parameter dan rentang nilai didasari dari penelitian analysis sentimen yang dilakukan oleh Kusumanigrum (2021), penelitian tersebut mengkomparasikan berbagai algoritma *classical machine learning* dan *deep learning* (Kusumaningrum et al., 2021). Berdasarkan rentang nilai yang ada, algoritma SVM menghasilkan 600 kandidat / kombinasi konfigurasi yang akan dievaluasi.

|  |  |  |  |
| --- | --- | --- | --- |
| Kombinasi 1 *halving grid search* – SVM | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | C | 0.1, 1, 10, 100, 1000 | 5 |
| Gamma | 1, 0.1, 0.01, 0.001, 0.0001 | 5 |
| kernel | 'rbf', 'poly', 'sigmoid', 'linear' | 4 |
| Vektorisasi TF-IDF | | Unigram, bigram, trigram, unibigram, unitrigram, bitrigram | 6 |
| Jumlah Kandidat | | 600 (5x5x4x6) kandidat | |

Untuk algoritma SVM penelitian ini hanya menjalankan satu skenario proses *halving grid search,* dimana terdapat 3 tahapan di dalamnya sebagaimana yang ditunjukkan pada Tabel X. Pada tahapan pertama digunakan 1000 *instance* yang dipilih secara acak untuk dapat mengevaluasi 600 kandidat kombinasi *hyperparameter* dan vektorisasi*.* Tahapan pertama menghasilkan 200 kandidat terbaik, dimana pada tahapan kedua dilakukan evaluasi dari 200 kandidat terbaik tersebut dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 3000 *instance*. Pada tahapan akhir terdapat 66 kandidat yang dievaluasi dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 5000 *instance*. Gambar x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini. Hasil dari proses ini akan ditampilkan pada Bagian “5.1.1 Hasil Evaluasi Algoritma SVM”.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Iterasi Ke-1 | Iterasi Ke-2 | Iterasi Ke-3 |
| Jumlah Kandidat | 600 | 200 | 66 |
| Jumlah Dataset | 1000 | 3000 | 5000 |

A screen shot of a computer code

Description automatically generated

Gambar 4 15Potongan kode *hyperparameter tuning* untuk algoritma SVM.

4.4.1.1.2 Evaluasi Algoritma *Random Forest*

Terdapat 2 hyperparameter dari algoritma *random forest* yang coba diidentifikasi nilai paling optimalnya dalam menghasilkan prediksi paling akurat, yaitu: n\_estimator dan max\_depth. Tabel x menjelaskan 2 parameter tersebut berserta dengan daftar nilai dari masing-masing parameter yang diujicobakan. Pemilihan parameter dan rentang nilai didasari dari penelitian analysis sentimen yang dilakukan oleh Parmar (2014), penelitian tersebut mengevaluasi berbagai kombinasi *hyperparameter* algoritma random forest. Berdasarkan rentang nilai yang ada, algoritma *random forest* menghasilkan 36 kandidat / kombinasi konfigurasi yang akan dievaluasikan.

|  |  |  |  |
| --- | --- | --- | --- |
| Kombinasi 1 *halving grid search* – Random Forest | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | n\_estimators | 100, 500, 1000 | 3 |
| max\_depth | 100, None (Unlimited) | 2 |
| Vektorisasi TF-IDF | | Unigram, bigram, trigram, unibigram, unitrigram, bitrigram | 6 |
| Jumlah Kandidat | | 36 (3x2x6) kandidat | |

Untuk algoritma *random forest* penelitian ini hanya menjalankan satu skenario proses *halving grid search,* dimana terdapat 3 tahapan di dalamnya sebagaimana yang ditunjukkan pada Tabel X. Pada tahapan pertama digunakan 1000 *instance* yang dipilih secara acak untuk dapat mengevaluasi 36 kandidat kombinasi *hyperparameter* dan vektorisasi*.* Tahapan pertama menghasilkan 12 kandidat terbaik, dimana pada tahapan kedua dilakukan evaluasi dari 12 kandidat terbaik tersebut dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 3000 *instance*. Pada tahapan akhir terdapat 4 kandidat yang dievaluasi dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 5000 *instance*. Gambar x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini. Hasil dari proses ini akan ditampilkan pada Bagian “5.1.2 Hasil Evaluasi Algoritma *Random Forest*”.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Iterasi Ke-1 | Iterasi Ke-2 | Iterasi Ke-3 |
| Jumlah Kandidat | 36 | 12 | 4 |
| Jumlah Dataset | 1000 | 3000 | 5000 |

A screenshot of a computer code

Description automatically generated

Gambar 4 12 Potongan kode proses *hyperparamter tuning* algoritma *random forest*.

4.4.1.1.3 Evaluasi Algoritma *Logistic Regression*

Terdapat 3 *hyperparameter* utama pada algoritma *logistic regression* yang coba diidentifikasi nilai paling optimalnya dalam menghasilkan prediksi paling akurat, yaitu: solver, pinalty dan C. Namun terdapat 1 *hyperparameter* tambahan dibutuhkan, yaitu l1\_ratio, ketika nilai ‘elasticnet’ digunakan sebagai nilai *hyperparameter*. Tidak seperti *hyperparameter* pada algoritma SVM dan *random forest,* dimana semua nilai *hyperparameter*-nya dapat dikombinasikan, pada algoritma *logistic reggression* tidak semua nilai *hyperparameter* dapat dikombinasikan. Seperti contoh nilai ‘liblinear’ pada *hyperparameter* ‘solver’ hanya bisa dikombinasikan dengan nilai ‘l1’ dan ‘l2’ pada *hyperparameter* ‘pinalty’, sehingga dibutuhkan 3 konfigurasi kombinasi yang berbeda.

Tabel x menjelaskan 4 *hyperparameter* tersebut berserta dengan daftar nilai dari masing-masing parameter yang diujicobakan. Pemilihan parameter dan rentang nilai didasari dari penelitian analysis sentimen yang dilakukan oleh Kusumanigrum (2021), penelitian tersebut mengkomparasikan berbagai algoritma *classical machine learning* dan *deep learning* (Kusumaningrum et al., 2021). Berdasarkan rentang nilai yang ada, algoritma *logistic regression* menghasilkan 570 kandidat / kombinasi konfigurasi yang akan dievaluasi.

|  |  |  |  |
| --- | --- | --- | --- |
| Kombinasi 1 *halving grid search* – Random Forest | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | solver | newton-cg, lbfgs, sag, saga | 4 |
| pinalty | none, l2 | 2 |
| C | 100, 10, 1.0, 0.1, 0.01 | 5 |
| Vektorisasi TF-IDF | | Unigram, … , bitrigram | 6 |
| Jumlah Kandidat | | 240 (4x2x5x6) kandidat | |
| Kombinasi 2 *halving grid search* – Random Forest | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | solver | saga | 1 |
| pinalty | elasticnet | 1 |
| C | 100, 10, 1.0, 0.1, 0.01 | 5 |
| l1\_ratio | 0.1, …., 1.0 | 10 |
| Vektorisasi TF-IDF | | Unigram, …. , bitrigram | 6 |
| Jumlah Kandidat | | 300 (1x1x5x10x6) kandidat | |
| Kombinasi 3 *halving grid search* – Random Forest | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | solver | liblinear | 1 |
| pinalty | l2, l1 | 1 |
| C | 100, 10, 1.0, 0.1, 0.01 | 5 |
| Vektorisasi TF-IDF | | Unigram, …. , bitrigram | 6 |
| Jumlah Kandidat | | 30 (1x1x5x6) kandidat | |
| **Total Kandidat** | | 240 + 300 + 30 = 570 Kandidat | |

Untuk algoritma *random forest* penelitian ini hanya menjalankan satu skenario proses *halving grid search* dengan 3 konfigurasi kombinasi yang berbeda*.* Dimana dalam 1 proses tersebut terdapat 3 tahapan di dalamnya sebagaimana yang ditunjukkan pada Tabel X. Pada tahapan pertama digunakan 1000 *instance* yang dipilih secara acak untuk dapat mengevaluasi 570 kandidat kombinasi *hyperparameter* dan vektorisasi*.* Tahapan pertama menghasilkan 190 kandidat terbaik, dimana pada tahapan kedua dilakukan evaluasi dari 190 kandidat terbaik tersebut dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 3000 *instance*. Pada tahapan akhir terdapat 63 kandidat yang dievaluasi dengan menggunakan jumlah dataset yang ditingkatkan, yaitu 5000 *instance*. Gambar x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini. Hasil dari proses ini akan ditampilkan pada Bagian “5.1.3 Hasil Evaluasi Algoritma *Logistic Regression*”.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Iterasi Ke-1 | Iterasi Ke-2 | Iterasi Ke-3 |
| Jumlah Kandidat | 570 | 190 | 63 |
| Jumlah Dataset | 1000 | 3000 | 5000 |

A screenshot of a computer code

Description automatically generated

Gambar 4 13 Potongan kode proses *hyperparameter tuning* algoritma *logistic regression*.

4.4.1.1.4 Evaluasi Algoritma *Deep Learning*

Terdapat 3 *hyperparameter* utama pada algoritma *deep learning* yang coba diidentifikasi nilai paling optimalnya dalam menghasilkan prediksi paling akurat, yaitu: activation, solver dan hidden\_layer\_sizes. Tidak seperti evaluasi *hyperparameter* pada algoritma SVM, *random forest* dan *logistic regression,* dimana pencarian kombinasi *hyperparameter* dilakukan secara serentak, pada algoritma *deep learning* identifikasi nilai *hyperparameter* dengan menggunakan *halving grid search* dipecah untuk masing-masing *hyperparameter*. Hal ini dilakukan dengan alasan, waktu evaluasi yang dibutuhkan oleh algoritma *deep learning* jauh lebih lama dibandingkan ketiga algoritma *classical machine learning* yang diujikan.

Sebelum melakukan evaluasi *hyperparameter* algoritma *deep learning,* evaluasi terhadap jenis vektorisasi dilakukan, sebagaimana ditunjukkan pada Tabel x. Hasil evaluasi jenis vektorisasi, dengan nilai *hyperparameter* default, menghasilkan unitrigam sebagai jenis vektorisasi dengan nilai akurasi tertinggi. Tabel x-x menjelaskan 3 *hyperparameter* berserta dengan daftar nilai yang diujicobakan. Pemilihan parameter dan rentang nilai didasari dari penelitian analysis sentimen yang dilakukan oleh Kusumanigrum (2021), penelitian tersebut mengkomparasikan berbagai algoritma *classical machine learning* dan *deep learning* (Kusumaningrum et al., 2021).

|  |  |  |
| --- | --- | --- |
| Scenario 1 - Kombinasi 1 *halving grid search* – Deep learning | | |
|  | Rentang Nilai | Jumlah Nilai |
| Vektorisasi TF-IDF | Unigram, bigram, trigram, unibigram, unitrigram, bitrigram | 6 |
| Jumlah Kandidat | 6 kandidat | |

|  |  |  |  |
| --- | --- | --- | --- |
| Scenario 2 - Kombinasi 1 *halving grid search* – Deep learning | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | activation | identity, logistic, tanh, relu | 4 |
| Vektorisasi TF-IDF | | unitrigram | 1 |
| Jumlah Kandidat | | 4 (4x1) kandidat | |

|  |  |  |  |
| --- | --- | --- | --- |
| Scenario 3 - Kombinasi 1 *halving grid search* – Deep learning | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | solver | lbfgs, sgd, adam | 3 |
| Vektorisasi TF-IDF | | unitrigram | 1 |
| Jumlah Kandidat | | 3 (3x1) kandidat | |

|  |  |  |  |
| --- | --- | --- | --- |
| Scenario 4 - Kombinasi 3 *halving grid search* – Deep learning | | | |
|  | | Rentang Nilai | Jumlah Nilai |
| *Classifer hyperparameter* | hidden\_layer\_sizes | (100) | 4 |
| (100, 50,) |
| (100, 50, 10) |
| (100, 70, 40, 10) |
| Vektorisasi TF-IDF | | unitrigram | 1 |
| Jumlah Kandidat | | 4 (4x1) kandidat | |

Berdasarkan rentang nilai yang ada, algoritma *deep learning* menghasilkan 4 kandidat untuk evaluasi parameter ‘activation’, 3 kandidat untuk evaluasi parameter ‘solver’ dan 4 kandidat untuk evaluasi parameter *hidden layer.* Pada evaluasi masing-masing *hyperparameter* hanya dilakukan dalam 1 tahap / jenjang, dimana seluruh *trainingset* sejumlah 5798 *instance* digunakan untuk masing-masing proses evaluasi. Gambar x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

A screen shot of a computer code

Description automatically generated

Gambar 4 14 Potongan kode *hyperparameter tuning* untuk algoritma *deep learning.*

##### Pemanfaatan Model Terbaik

Hasil *hyperparameter tunning* dari masing-masing algoritma kemudian dibandingkan berdasarkan metrik f1-score. Hasil dari perbandingan tersebut menghasilkan algoritma dengan performa terbaik yang akan digunakan untuk mengklasifikasikan secara otomatis jenis sentimen dari dataset dengan kelas sentimen yang belum terdefinisi. Gambar x.x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

### Tahapan Pemodelan Topik

Tujuan dari tahapan ini adalah untuk mengidentifikasi topik atau tema-tema aspirasi masyarakat baik yang berasal dari opini bersentimen positif maupun opini bersentimen negatif. Oleh karena itu input pada tahapan ini adalah hasil dari tahapan sebelumnya, yaitu pemanfaatan model *machine learning* terbaik, yang berupa seluruh dataset dengan jenis sentimen atau *dependent variable* yang sudah terdefinisi. Teknik pemodelan topik yang digunakan pada tahapan ini adalah *latent derichlet allocation* (LDA)*.* Gambar x menunjukkan potongan kode yang digunakan untuk menjalankan proses ini.

# BAB 5 HASIL DAN PEMBAHASAN

Pada bagian ini akan menjelaskan beberapa temuan yang menjadi jawaban atas Pertanyaan Penelitian, yang diantaranya adalah (1) Hasil Evaluasi Model Machine Learning, (2) Hasil Klasifikasi Sentimen, (3) Hasil Pemodelan Topik.

5.1 Hasil Evaluasi Model Machine Learning

Dalam proses mengidentifikasi model yang dapat memberikan tingkat akurasi tertinggi, penelitian ini mengadopsi proses *hyperparameter tunning* dengan strategi *halving grid search*.Penjelasan tentang *halving grid search* dapat ditemukan pada bagian “4.4.1.1 Identifikasi Model Terbaik”. Proses tersebut dipecah untuk masing-masing algoritma *machine learning.* Dalam satu proses *halving grid search,* selain untuk menemukan *classifier* dengan kombinasi *hyperparameter* paling optimal, pencarian juga dilakukan untuk menemukan bentuk vektorisasi TF-IDF (unigram, bigram, trigram, atau kombinasi ketiganya) paling optimal untuk dikombinasikan dengan *classifier* tersebut. Berikut penjelasan hasil evaluasi menggunakan halving grid search untuk masing masing algoritma.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | Jenis Vektorisasi | C | Gamma | Kernel | F1-score |
| #1 |  |  |  |  |  |
| #2 |  |  |  |  |  |
| #3 |  |  |  |  |  |
| #4 |  |  |  |  |  |
| #5 |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Jenis Vektorisasi | n\_estimator | max\_depth | F1-score |
| #1 |  |  |  |  |
| #2 |  |  |  |  |
| #3 |  |  |  |  |
| #4 |  |  |  |  |
| #5 |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | Jenis Vektorisasi | solver | pinalty | C | l1\_ratio | F1-score |
| #1 |  |  |  |  |  |  |
| #2 |  |  |  |  |  |  |
| #3 |  |  |  |  |  |  |
| #4 |  |  |  |  |  |  |
| #5 |  |  |  |  |  |  |

5.1 Hasil Klasifikasi Sentimen

5.3 Hasil Pemodelan Topik

# DAFTAR PUSTAKA

Afra, D. I. N. (2022). *Analisis Sentimen dan Pemodelan Topik dengan Data Media Sosial Twitter: Studi Kasus Komisi Pemberantasan Korupsi*. https://lib.ui.ac.id/detail?id=20523321

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In *Mining Text Data* (Vol. 9781461432). https://doi.org/10.1007/978-1-4614-3223-4

Arafat, T. A., Budi, I., Mahendra, R., & Salehah, D. A. (2020). Demographic Analysis of Candidates Supporter in Twitter During Indonesian Presidential Election 2019. *2020 International Conference on ICT for Smart Society (ICISS)*, 1–6. https://doi.org/10.1109/ICISS50791.2020.9307598

Bagja, Suhenti, & Puadi. (2023). *Buku Ikp Pemilu Dan Pemilihan Serentak 2024*. 14–18.

DataIndonesia.id. (2023). *BPS: Kepercayaan Masyarakat RI terhadap Pemilu Masih Rendah*. https://dataindonesia.id/data-pemilu/detail/bps-kepercayaan-masyarakat-ri-terhadap-pemilu-masih-rendah

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143–148. https://doi.org/10.1016/j.neucom.2015.08.112

Dennis, D., Nugraha, Y., Sulasikin, A., Irawan, F. A., Febria Finola, C., Kurniawan, F., Kanggrawan, J. I., & Suherman, A. L. (2021). Sentiment Classification Against the Public Activity Restrictions Policy in Jakarta Using Machine Learning Models. *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, 5–9. https://doi.org/10.1109/ICAIBDA53487.2021.9689761

DPR RI. (2017). Undang-Undang Nomor 7 Tahun 2017 tentang Pemilihan Umum. *Undang-Undang Pemilu*, 1–317. http://rumahpemilu.org/wp-content/uploads/2017/08/UU-No.7-Tahun-2017-tentang-Pemilu.pdf

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707–714. https://doi.org/10.1016/j.procs.2019.11.174

Garnett, H. A., James, T. S., MacGregor, M., & Caal-Lam, S. (2023). *Perceptions of Electoral Integrity, (PEI-9.0)* (E. I. P. A.-S. S. and H. R. Council (ed.); V1 ed.). Harvard Dataverse. https://doi.org/doi:10.7910/DVN/2MFQ9K

Haenlein, M., & Kaplan, A. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. http://search.proquest.com/docview/195360323?accountid=15920

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). *Natural Language Processing: Python and NLTK*. 755. http://www.nltk.org/book

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviewsJain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment an. *Computer Science Review*, *41*. https://doi.org/10.1016/j.cosrev.2021.100413

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kanugrahan, G., & Wicaksono, A. F. (2021). Sentiment Analysis of Face-to-face Learning during Covid-19 Pandemic using Twitter Data. *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*. https://doi.org/10.1109/ICAICTA53211.2021.9640282

Kelleher, J. D., Namee, B. Mac, & D’Arcy, A. (2020). Fundamentals of Machine Learning for Predictive Data Analytics (Second Edition). In *The MIT Press* (Issue 1).

Khan, D. G. F. (2018). *Creating Value With Social Media Analytics: Managing, Aligning, and Mining Social Media Text, Networks, Actions, Location, Aps, Hyperlinks, Multimedia, & Search Engines Data*. 510. https://www.amazon.com/Creating-Value-Social-Media-Analytics/dp/1977543979

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 144–168. https://doi.org/10.15849/ijasca.211128.11

Kompas. (2023, November 13). *Dari Putusan MK yang Kontroversial hingga Netralitas Aparat Jadi ”PR” KPU Pulihkan Kepercayaan Publik*. https://www.kompas.id/baca/polhuk/2023/11/13/pulihkan-kepercayaan-publik-terhadap-proses-pemilu

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kusumaningrum, R., Nisa, I. Z., Nawangsari, R. P., & Wibowo, A. (2021). Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning. *International Journal of Advances in Intelligent Informatics*, *7*(3), 292–303. https://doi.org/10.26555/ijain.v7i3.737

Laurie A. Schintler; Connie L. McNeely. (2022). Encyclopedia of Big Data. *Encyclopedia of Big Data*. https://doi.org/10.1007/978-3-319-32010-6

Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. In *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. https://doi.org/10.1017/CBO9781139084789

Mauk, M. (2022). Electoral integrity matters: how electoral process conditions the relationship between political losing and political trust. *Quality and Quantity*, *56*(3), 1709–1728. https://doi.org/10.1007/s11135-020-01050-1

Maulana, P., Budi, I., & Budi Santoso, A. (2022). Sentiment Analysis of Indonesian Government’s Effort to Overcome the Unemployment Problem during COVID-19 Pandemic. *ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding*, 144–149. https://doi.org/10.1109/ICOIACT55506.2022.9971853

Mosquera, A., Gutiérrez, Y., & Moreda, P. (2017). On evaluating the contribution of text normalisation techniques to sentiment analysis on informal Web 2.0 texts. *Procesamiento Del Lenguaje Natural*, *58*, 29–36.

Norris, P. (2019). Do perceptions of electoral malpractice undermine democratic satisfaction? The US in comparative perspective. *International Political Science Review*, *40*(1), 5–22. https://doi.org/10.1177/0192512118806783

Perkins, J. (2014). Python 3 Text Processing With NLTK 3 Cookbook. In *Python 3 Text Processing With NLTK 3 Cookbook*. http://www.amazon.com/s/ref=nb\_sb\_noss?url=search-alias%3Daps&field-keywords=9781782167853

Recker, J. (2013). Scientific Research in Information Systems. In *Scientific Research in Information Systems*. https://doi.org/10.1007/978-3-642-30048-6

Rodríguez, S., Allende-Cid, H., Palma, W., Alfaro, R., Gonzalez, C., Elortegui, C., & Santander, P. (2018). Forecasting the Chilean electoral year: Using twitter to predict the presidential elections of 2017. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10914 LNCS*, 298–314. https://doi.org/10.1007/978-3-319-91485-5\_23

Saud, M., Ida, R., Mashud, M., Yousaf, F. N., & Ashfaq, A. (2023). Cultural dynamics of digital space: Democracy, civic engagement and youth participation in virtual spheres. *International Journal of Intercultural Relations*, *97*. https://doi.org/10.1016/j.ijintrel.2023.101904

Saud, M., & Margono, H. (2021). Indonesia’s rise in digital democracy and youth’s political participation. *Journal of Information Technology and Politics*, *18*(4), 443–454. https://doi.org/10.1080/19331681.2021.1900019

Sirait, F. E. T., Pratiwi, D., Kusuma, R. D., Habibi, M., Manan, L., Setiawan, H., & Destavino, I. (2022). *Netralitas ASN: Problematika dan Studi Kontemporer*.

Tan, N. (2023). Electoral Malpractice in Asia. *Electoral Malpractice in Asia*. https://doi.org/10.1515/9781685852931

Tang, W., Hu, J., Zhang, H., Wu, P., & He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai Archives of Psychiatry*, *27*(1), 62–67. https://doi.org/10.11919/j.issn.1002-0829.215010

Van Ham, C. (2020). ELECTORAL INTEGRITY. In *The Oxford Handbook of Political Representation in Liberal Democracies* (pp. 113–134). https://doi.org/10.1093/oxfordhb/9780198825081.013.5

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: An introduction. In *Social Media Mining: An Introduction* (Vol. 9781107018). https://doi.org/10.1017/CBO9781139088510

Lampiran 1

1. https://id.wikipedia.org/wiki/Pembatasan\_daerah\_pemilihan [↑](#footnote-ref-2)
2. https://www.kompas.id/baca/polhuk/2023/11/13/pulihkan-kepercayaan-publik-terhadap-proses-pemilu [↑](#footnote-ref-3)
3. https://www.bbc.com/indonesia/articles/cpw7enedn39o [↑](#footnote-ref-4)
4. https://www.cnnindonesia.com/teknologi/20240212064143-192-1061254/dirty-vote-trending-di-x-netizen-teriak-hancur-negara-ini [↑](#footnote-ref-5)
5. https://www.metrotvnews.com/read/bVDCgQ4g-kpu-harus-segera-beri-solusi-soal-masalah-sirekap [↑](#footnote-ref-6)
6. https://www.bbc.com/indonesia/articles/c97vmg0z0y1o [↑](#footnote-ref-7)
7. https://apjii.or.id/berita/d/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang [↑](#footnote-ref-8)
8. https://dataindonesia.id/internet/detail/pengguna-media-sosial-di-indonesia-sebanyak-167-juta-pada-2023 [↑](#footnote-ref-9)
9. https://data.goodstats.id/statistic/agneszefanyayonatan/menilik-pengguna-media-sosial-indonesia-2017-2026-xUAlp [↑](#footnote-ref-10)
10. https://datareportal.com/reports/digital-2023-indonesia [↑](#footnote-ref-11)
11. https://en.wikipedia.org/wiki/User-generated\_content [↑](#footnote-ref-12)
12. https://www.botify.com/learn/basics/internal-links [↑](#footnote-ref-13)
13. https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist [↑](#footnote-ref-14)
14. https://rahmadya.com/2019/04/24/stopword-berbahasa-indonesia/ [↑](#footnote-ref-15)
15. https://en.wikipedia.org/wiki/Conditional\_probability [↑](#footnote-ref-16)
16. https://en.wikipedia.org/wiki/Lazy\_learning [↑](#footnote-ref-17)
17. https://lib.ui.ac.id/home [↑](#footnote-ref-18)
18. https://www.detik.com/jabar/berita/d-6241560/morfologi-adalah-pengertian-proses-jenis-dan-contoh [↑](#footnote-ref-19)
19. https://github.com/helmisatria/tweet-harvest [↑](#footnote-ref-20)
20. https://github.com/nasalsabila/kamus-alay [↑](#footnote-ref-21)
21. https://scikit-learn.org/stable/modules/grid\_search.html [↑](#footnote-ref-22)