

UNIVERSITAS INDONESIA

Analisa Sentimen Masyarakat Terhadap Integritas Pemilu Indonesia 2024 pada Media Sosial Twitter

KARYA AKHIR

AHMAD FAUZI

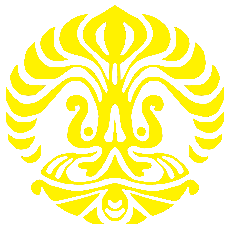
2106806630

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI

JAKARTA

JANUARI 2024



UNIVERSITAS INDONESIA

Analisa Sentimen Masyarakat Terhadap Integritas Pemilu Indonesia 2024 pada 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 : | <judul> |
| Pembimbing : | …………………………………………  ………………………………………… |

Penyelenggaraan Pemilihan Umum (Pemilu) yang luber jurdil merupakan harapan segenap bangsa sebagaimana juga tertuang dalam Undang-Undang dasar 1945. Namun pelanggaran dalam penyelenggaraan Pemilu merupakan suatu keniscayaan yang bisa saja terjadi, jika penyimpangan tersebut tidak diperbaiki secara bertahap dikhawatirkan bisa berpotensi menimbulkan dampak negatif, seperti, *political distrust* bahkan perpecahan bangsa. Kehadiran *civil society* sebagai pengawas penyelenggaraan pemilihan umum dipercaya dapat meminimalisir praktik pelanggaran Pemilu atau memberikan saran perbaikan pada kontestasi selanjutnya. Media sosial hadir sebagai sarana yang mempermudah *civil society* untuk dapat berpartisipasi dalam politik, termasuk melakukan pengawas penyelenggaraan pemilihan umum. Opini dan aspirasi yang dikemukakan *civil society* terkait integritas Pemilu dapat dijadikan bahan analisa oleh lembaga Penyelenggara Pemilu, khususnya Bawaslu yang fungsi dan tugasnya beririsan langsung dengan upaya tersebut. Penelitian ini dilakukan untuk membantu Bawaslu dalam menggali opini dan aspirasi yang dikemukakan *civil society* terkait integritas Pemilu di media sosial. Penelitian ini mengkomparasikan teknik *deep learning* dengan *classical machine learning* untuk menghasilkan model yang dapat memprediksi dengan akurat. Kemudian teknik pemodelan topik *Latent Dirichlet Allocation* (LDA) digunakan untuk mengetahui topik yang menjadi pembicaraan di tengah masyarakat yang dianalisa dalam bingkai *framework* pola indikasi pelanggaran Pemilu.

Kata kunci: Analisis Sentimen, Integritas Pemilu, Pelanggaran Pemilu

# ABSTRACT

|  |  |
| --- | --- |
| Name : | Ahmad Fauzi |
| Study Program : | Magister Teknologi Informasi |
| Title : | <titlte> |
| Counsellor : | …………………………………………  ………………………………………… |

The conduct of General Elections (Pemilu) that is transparent and just is the collective aspiration of the nation, as enshrined in the 1945 Constitution. However, violations during the conduct of elections are inevitable, and if these deviations are not rectified gradually, there is a concern that they could potentially have negative consequences, such as political distrust and even national division. The presence of civil society as overseers of the electoral process is believed to be capable of minimizing electoral violations or providing recommendations for improvements in subsequent contests. Social media serves as a facilitator enabling civil society to participate in politics, including monitoring the conduct of general elections. Opinions and aspirations expressed by civil society regarding the integrity of elections can be utilized as analytical material by the Election Organizer institutions, particularly the Election Supervisory Agency (Bawaslu), whose functions and duties directly intersect with these efforts. This research is conducted to assist Bawaslu in extracting opinions and aspirations expressed by civil society regarding the integrity of elections on social media. The study compares deep learning techniques with classical machine learning to develop a model capable of accurate prediction. Subsequently, *Latent Dirichlet Allocation* (LDA) topic modeling techniques are employed to identify the topics under discussion within the analyzed electoral integrity framework, specifically focusing on the patterns of electoral malpractice.

Keywords: Sentiment Analysis, Electoral Integrity, Electoral Malpractice

# DAFTAR ISI

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

[HALAMAN PERNYATAAN ORISINALITAS ii](#_Toc157550047)

[HALAMAN PENGESAHAN iii](#_Toc157550048)

[KATA PENGANTAR v](#_Toc157550049)

[HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS vi](#_Toc157550050)

[ABSTRAK vii](#_Toc157550051)

[ABSTRACT viii](#_Toc157550052)

[DAFTAR ISI ix](#_Toc157550053)

[DAFTAR GAMBAR xii](#_Toc157550054)

[DAFTAR TABEL xiii](#_Toc157550055)

[DAFTAR LAMPIRAN xiv](#_Toc157550056)

[BAB 1 PENDAHULUAN 15](#_Toc157550057)

[1.1 Latar Belakang 15](#_Toc157550058)

[1.2 Rumusan Masalah 21](#_Toc157550059)

[1.3 Tujuan Penelitian 30](#_Toc157550060)

[1.4 Ruang Lingkup Penelitian 30](#_Toc157550061)

[1.5 Manfaat Penelitian 31](#_Toc157550062)

[1.6 Sistematika Penulisan 31](#_Toc157550063)

[BAB 2 STUDI LITERATUR 32](#_Toc157550064)

[1.1 Pola Pelanggaran Pemilu 32](#_Toc157550065)

[1.2 Media Sosial 38](#_Toc157550066)

[2.2 Analisa Media Sosial 39](#_Toc157550067)

[2.3 CRIPS-DM 41](#_Toc157550068)

[2.4 Analisis Sentimen 42](#_Toc157550069)

[2.5 Text Mining 44](#_Toc157550070)

[2.5.1 Data Collection 44](#_Toc157550071)

[2.5.2 Text Preprocessing 45](#_Toc157550072)

[2.5.3 Text Representation 47](#_Toc157550073)

[2.5.4 Knowledge Discovery 48](#_Toc157550074)

[2.6 Penelitian Terdahulu 60](#_Toc157550075)

[2.6.1 Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi 60](#_Toc157550076)

[2.6.2 Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19 61](#_Toc157550077)

[2.6.3 Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter. 63](#_Toc157550078)

[2.6.4 Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model Machine Learning. 64](#_Toc157550079)

[2.6.5 Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia. 66](#_Toc157550080)

[2.7 Ringkasan Penelitian Terdahulu 68](#_Toc157550081)

[2.8 Perbadingan Penelitian Terdahulu 70](#_Toc157550082)

[2.9 Kerangka Penelitian 75](#_Toc157550083)

[BAB 3 METODOLOGI PENELITIAN 77](#_Toc157550084)

[3.1 Rancangan Penelitian 77](#_Toc157550085)

[3.2 Alur Peneliitian 79](#_Toc157550086)

[3.2.1 Pengumpulan Data Awal 81](#_Toc157550087)

[3.2.2 Perumusan Masalah 81](#_Toc157550088)

[3.2.3 Penentuan Pertanyaan Penelitian 82](#_Toc157550089)

[3.2.4 Studi Literatur 82](#_Toc157550090)

[3.2.5 Pengumpulan Data 82](#_Toc157550091)

[3.2.6 Membuat Korpus 82](#_Toc157550092)

[3.2.7 Data Preprocessing 83](#_Toc157550093)

[*3.2.8* *Text Representation* 84](#_Toc157550094)

[3.2.9 Klasifikasi Sentimen dan Opini Tidak Relevan 84](#_Toc157550095)

[3.2.10 Evaluasi 84](#_Toc157550096)

[3.2.11 Pemodelan Topik 85](#_Toc157550097)

[3.2.12 Penarikan Kesimpulan 85](#_Toc157550098)

[BAB 4 HASIL DAN PEMBAHASAN 86](#_Toc157550099)

[BAB 5 KESIMPULAN DAN SARAN 87](#_Toc157550100)

[DAFTAR PUSTAKA 88](#_Toc157550101)

# DAFTAR GAMBAR

# DAFTAR TABEL

# 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

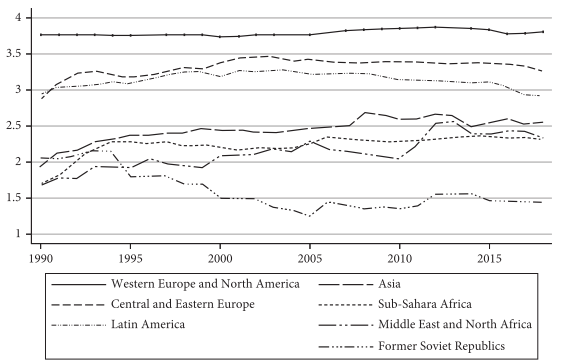
Tahun 2024 merupakan tahun politik bagi Negara Kesatuan Republik Indonesia, pasalnya, di tahun tersebut akan diselengarakannya Pemilihan Umum. Pemilihan Umum atau Pemilu adalah hak demokrasi dan perwujudan kedaulatan rakyat untuk memilih Presiden dan Wakil Presiden, anggota Dewan Perwakilan Rakyat, anggota Dewan Perwakilan Daerah dan anggota Dewan Perwakilan Rakyat Daerah (DPR RI, 2017). Berdasarkan Pancasila dan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945, Pemilihan Umum yang diselengarakan haruslah bersifat 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 ini diharapkan dapat menjadi wujud sistem ketatanegaraan yang menjunjung tinggi nilai-nilai demokrasi dan menjadi landasan hukum yang menjamin terselengaranya Pemilu yang jujur dan adil (DPR RI, 2017).

Sebagaimana yang tertuang dalam UU No 7 Tahun 2017, untuk memastikan penyelenggaraan Pemilu sesuai dengan peraturan per-undang-undang yang berlaku. Pemerintah telah membentuk Badan Pengawas Pemilu (Bawaslu) yang ditugaskan untuk melakukan pengawasan dalam penyelenggaraan pemilihan umum. Tugas Bawaslu mulai dari (a) membentuk standar tata laksana pengawasan, (b) melakukan pencegahan dan penindakan terhadap pelanggaran Pemilu, (c) mengawasi persiapan Pemilu, (d) mengawasi pelaksanaan tahapan Pemilu, (e) mengawasi netralitas ASN, TNI dan Polri, hingga (f) mengawasi segala putusan KPU, DKPP dan Gakumdu. Bersama lembaga penyelenggara Pemilu lainnya seperti Komisi Pemilihan Umum (KPU) yang bertugas menjalankan tahapan penyelenggaraan pemilihan umum dan Dewan Kehormatan Penyelenggaraan Pemilu (DKPP) yang bertugas untuk menggani pelanggaran kode etik Penyelenggara Pemilu, Bawaslu diharapkan bisa menciptakan Pemilu yang berintegritas (DPR RI, 2017).

Menurut Pippa Noris, Pemilu yang berintegritas 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).

Dalam konteks internasional, dalam upaya menyelenggarakan Pemilu yang berintegritas tidak lepas dari adanya temuan pelanggaran Pemilu. Pelanggaran di dalam Pemilu merupakan suatu hal yang tidak dapat dihindari, hal ini dikarenakan sudah menjadi bawaan, dimana mayoritas penguasa ingin melanjutkan kekuasaannya (Tan, 2023). Bahkan di 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*, manipulasi pendaftaraan pemilih, manipulasi proses pemunggutan suata, bias media, berita hoaks, penggunaan teknologi yang tidak aman, aturan *campaign finance* yang tidak adil, sampai dengan *vote buying* menjadi pelanggaran yang sering muncul di negara-negara tersebut. Walaupun skor persepsi Pemilu berintegritas 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 menunjukkan bahwa kecurangan di dalam Pemilu merupakan keniscayaan.

Sedangkan dalam konteks Penyelenggaraan Pemilihan Umum di Indonesia. Menurut Badan Pusat Statistik (BPS), masih rendahnya tingkat kepercayaan masyarakat Indonesia terhadap proses penyelengaraan Pemilu, dimana nilai rata-rata tingkat kepercayaan untuk masing-masing provinsi adalah 69,72 poin [[1]](#footnote-2). Provinsi Papua, Banten dan Aceh menjadi provinsi dengan tingkat kepercayaan terkecil, dengan nilai 63.7 poin, 63.88 poin dan 64.05 poin. Kecilnya angka persepsi kepercayaan Pemilu di masyarakat dikhawatirkan dapat menurunkan partisipasi pemilih dalam penyelengaraan Pemilu.



Gambar 1. 1 Skor integritas Pemilu berdasarkan kawasan dari waktu ke waktu 1990-2018, rentang nilai 0-4 (Van Ham, 2020).

Pernyataan, putusan dan kabar kontroversial yang telah terjadi akhir-akhir ini, seperti: (a) Pernyataan Presiden Jokowi yang akan cawe-cawe dalam Pemilu 2024 mendatang[[2]](#footnote-3). (b) 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[[3]](#footnote-4). (c) Perhatian Bawaslu terkait adanya Menteri yang hanya mengambil cuti ketika menjadi bagian dari partai politik peserta Pemilu[[4]](#footnote-5). (d) Adanya laporan dari Koalisi Masyarakat Sipil Kawal Pemilu Bersih yang tidak ditindak lanjuti oleh DKPP terkait laporannya terhadap dugaan kecurangan proses verifikasi peserta Pemilu oleh KPU[[5]](#footnote-6). Hal-hal tersebut tentunya menambah kekhawatiran bagi publik akan tidak terlaksananya Pemilu yang luber jurdil.

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).

Kemudian berdasarkan data *post-factum* Pemilu sebelumnya, 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. Tabel 1.1 menunjukkan sebaran tingkat kerawanan berdasarkan jenis dimensi untuk tingkat provinsi. Sementara, Tabel 1.2 dan 1.3 menunjukkan daerah yang masuk ke dalam kategori kerawanan tinggi yang dipetakan berdasarkan dimensi kerawanannya.

Tabel 1. 1 Distribusi kategori kerawanan berdasarkan jenis dimensi untuk tingkat provinsi

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **Kategori** | **Dimensi Sosial Politik** | **Dimensi Penyelenggaraan Pemilu** | **Dimensi Kontestasi** | **Dimensi Partisipasi** |
| 1 | Tinggi | 14,71% | 17,65% | 17,65% | 17,65% |
| 2 | Sedang | 64,71% | 58,82% | 55,88% | 82,35% |
| 3 | Rendah | 20,59% | 23,53% | 26,47% |  |

Tabel 1. 2 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) |

Tabel 1. 3 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) |

Pada tingkat provinsi, mayoritas berada pada kategori sedang. Untuk sebaran kategori tinggi, terjadi di seluruh pulau besar di Indonesia. Sedangkan untuk tingkat kabupaten/kota, mayoritas berasal dari provinsi Papua, temuan ini mengindikasikan dibutuhkannya strategi mitigasi khusus yang melibatkan para tokoh adat di Papua (Bagja et al., 2023). Lebih lanjut lagi penilaian *Electoral Integtiry Project* menyebutkan Indonesia masuk ke dalam kategori moderat sebagai negara yang telah mengadopsi nilai-nilai pemilu yang berintegritas (Garnett et al., 2023). Sehingga pengawasan penyelenggaraan Pemilu dalam upaya meningkatkan kualitas integritas Pemilu di Indonesia masih sangat relevan.

Penelitian tentang *democratic satisfaction* dan *political trust*, menunjukkan betapa pentingnya Pemilu yang berintegritas di mata masyarakat. Norris mengungkapkan bahwa Pemilu yang berintegritas 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 terseleenggara dengan baik dan berintegritas 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(Mauk, 2022).

Ketika ditemukannya kejanggalan, manipulasi, pelanggaran dalam penyelenggaraan pemilihan umum maka telah rusaklah proses translasi dari preferensi menjadi dukungan, dukungan menjadi suara, suara menjadi representasi, representasi menjadi kekuatan politik (Van Ham, 2020). 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 perpecahan bangsa, terlebih di negara dengan struktur sosial yang beragam (Mauk, 2022; Norris, 2019).

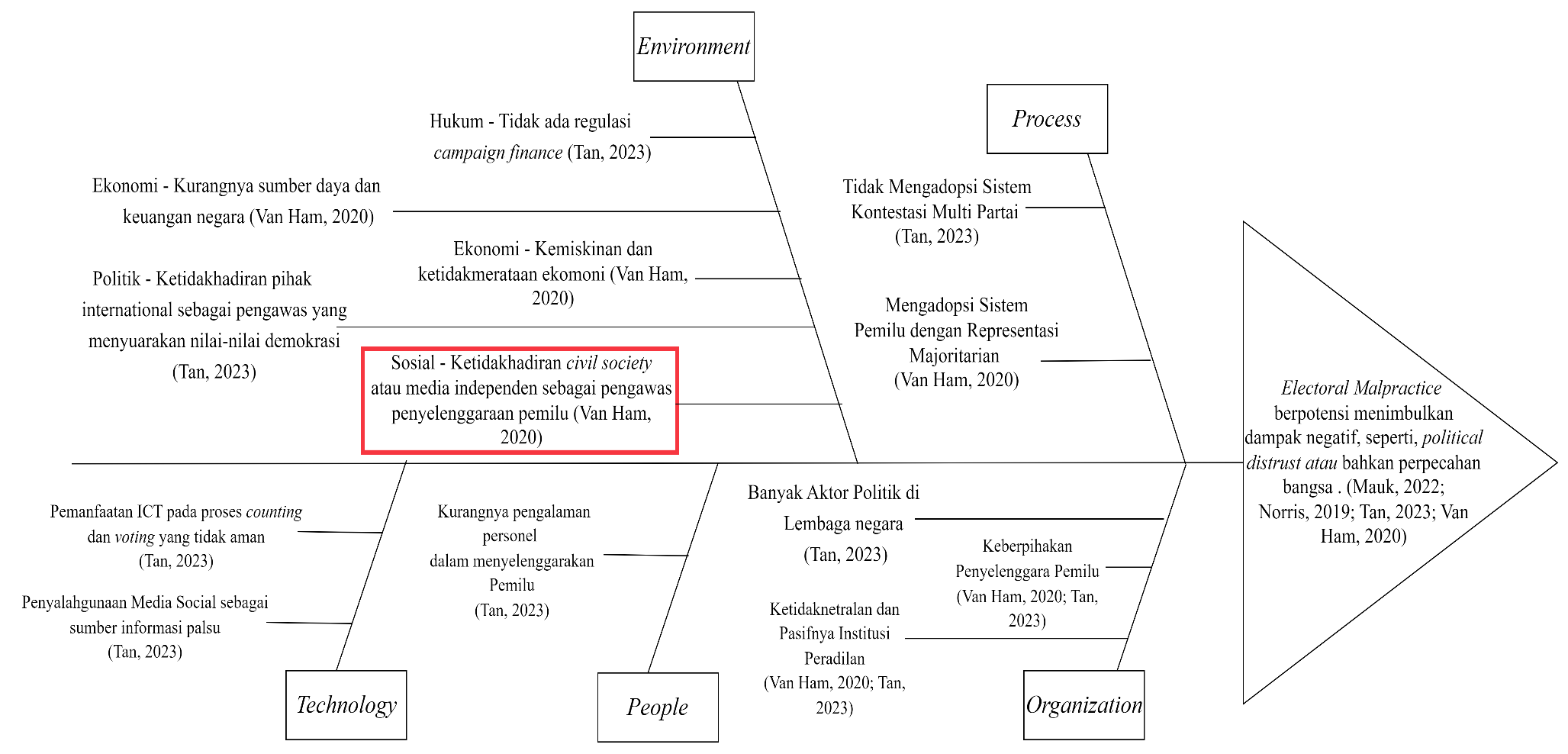
## Rumusan Masalah

Analisa kesenjangan, yang tampilkan pada tabel 1.4, dilakukan berdasarkan hasil studi literatur dan analisa data skunder untuk menemukan (a) harapan atas jalannya penyelenggaraan Pemilu di Indonesia yang luber jurdil dan (b) kenyataan akan adanya indikasi-indikasi pelanggaran dalam tahapan penyelenggaraan Pemilu. Kemudian masalah umum diidentifikasi berdasarkan perbandingan kedua hal tersebut.

Tabel 1. 4 Harapan, Kenyataan dan Masalah Umum

|  |  |  |
| --- | --- | --- |
| **Harapan** | **Kenyataan** | **Masalah Umum** |
| Terselenggaranya pemilihan umum yang luber jurdil sebagai bentuk sarana kedaulatan dan hak demokrasi rakyat (DPR RI, 2017). | Berdasarkan data *post-factum* Pemilu sebelumnya pada IKP 2024 dan opini yang tengah berkembang di Masyarakat, adanya kerawanan dalam penyelenggaraan Pemilu 2024 (Bagja et al., 2023). | Pelanggaran dalam penyelenggaeraan Pemilu bisa saja terjadi dan merupakan suatu keniscayaan, namun jika penyimpangan tersebut tidak diperbaiki secara bertahap dikhawatirkan bisa berpotensi menimbulkan dampak negatif, seperti, *political distrust* atau bahkan lebih dari itu perpecahan bangsa. (Mauk, 2022; Norris, 2019; Tan, 2023; Van Ham, 2020) |

Dari masalah umum yang berhasil teridentifikasi pada analisis kesenjangan, kemudian dilanjutkan dengan mengidentifikasi dan menguraikan faktor-faktor yang menyebabkan masalah umum tersebut. Untuk dapat mengidentifikasi faktor-faktor pendorong atau penyebab masalah umum, dilakukan studi literatur terkait domain *electoral integtiry* dan *electoral malpractice.* Faktor-faktor yang teridentifikasi kemudian dipetakan ke dalam 5 domain masalah yaitu *environment, people, process, organization,* dan *technology* (Li & Lee, 2011) yang direpresentasikan dengan *fishbone analysis diagram* sebagaimana ditunjukkan pada Gambar 1.2*.*



Gambar 1. 2 Analisa faktor pendorong dalam pelanggaran Pemilu dengan Fishbone analysis diagram

Berikut adalah uraian faktor-faktor masalah yang teridentifikasi dan dikelompokkan ke dalam 4 domain:

1. *Environment*

Pada domain ini akan dijabarkan faktor-faktor pendorong pelanggaran Pemilu yang dilihat dalam konteks ekonomi, sosial, politik dan hukum.

1. Ekonomi - Kurangnya sumber daya atau anggaran yang dimiliki negara bisa mempegaruhi kapasitas negara tersebut dalam melaksanakan administrasi Pemilu yang kompleks dan membutuhkan biaya logistik yang sangat besar (Van Ham, 2020).
2. Ekonomi - Tingginya angka kemiskinan dan ketidakmerataan ekonomi membuat aktor politik berada di atas angin, hal ini memudahkan mereka untuk melakukan praktik *vote buying* atau klientalisme (Van Ham, 2020). Bahkan pada kasus di negara dengan kesenjangan ekonomi yang ekstrim, faktor ekonomi bisa menjadi penghalang akses kontestasi politik. Kemudian ketika motivasi untuk masuk ke dalam pemerintahan adalah untuk menguasai ekonomi (kontrak proyek, sumber daya alam, atau bahkan legislatif untuk menjaga kekayaan), maka para aktor politik akan lebih termotivasi untuk dapat memenangkan kontestasi dengan segala macam cara (Van Ham, 2020).
3. Sosial - Ketidakhadiran *civil society* dan media dalam melaporkan setiap pelanggaran dalam Pemilu (Tan, 2023). Menurut Netina Tan, Indonesia menjadi negara yang berhasil meningkatkan kualitas Pemilunya secara bertahap dengan kehadiran 2 elemen tersebut, walaupun belum secara maksimal.
4. Sosial – Keberagaman sosial baik itu etnis, agama, suku, ras dan bahasa. Jika tidak ditanggani dengan tepat, hal ini dapat berdampak pada makin tingginya tensi kompetisi dan pertaruhan politik, yang pada akhirnya meningkatkan motivasi para aktor politik untuk memenangkan Pemilu dengan berbagai macam cara (Van Ham, 2020).
5. Politik – Ketidakhadiran pihak internasional sebagai pengawas penyelenggaraan Pemilu. Walaupun penyelenggaraan Pemilu merupakan hak demokrasi suatu negara, namun kehadiran pihak internasional sebagai pengawas yang independen, berguna untuk memastikan penyelenggaraan Pemilu di suatu negara berlangsung dengan tidak mengabaikan norma dan standar yang disepakati secara international (Tan, 2023). Kecaman dari pihak luar ketika adanya pelanggaran Pemilu yang meruntuhkan sistem demokrasi, menjadi bahan pertimbangan bagi mereka yang komit akan melakukan kecurangan.
6. Hukum – Ketidakhadiran regulasi yang efektif terkait *campaign finance* telah memungkinkan praktik pembelian suara merajalela dan menyebabkan tingkat pengeluaran yang sangat tinggi dari para kandidat (Tan, 2023).
7. *Organization / Institutional*

Pada domain ini membahas faktor-faktor yang disebabkan oleh ketidaknetralan dari Lembaga negara.

1. Keberpihakan Penyelenggara Pemilu, hal ini dapat dikarenakan postur atau kapasitas lembaga Penelenggara Pemilu mayoritas diisi oleh partisan peserta Pemilu (Tan, 2023). Salah satu cara yang dinilai cukup berhasil dalam menganggulangi keberpihakan Penglenggara Pemilu adalah dengan memecah fungsi dan kewenanggan lembaga tersebut dan menciptakan lembaga yang menjalankan peran pengawasan (Van Ham, 2020). Seperti yang telah terjadi di Indonesia, dimana lembaga penyelenggara Pemilu dibagi menjadi 3 lembaga terpisah, yaitu: (1) KPU yang bertugas melaksanakan tahapan Pemilu, (2) Bawaslu yang bertugas melakukan pengawasan penyelenggaraan Pemilu dan (3) DKPP yang bertugas mengawasi kedua lembaga yang sebelumnya disebutkan (Tan, 2023). Menurut Netina Tan, sistem penyelenggaraan Pemilu yang terpecah tersebut, cukup efektif dalam mengurangi dugaan kecurangan Pemilu seperti manipulasi proses pemungutan dan penghitungan suara, yang kerap terjadi di era Presiden Suharto (Tan, 2023).
2. Ketidaknetralan atau Pasifnya Institusi Peradilan, selain Penyelenggara Pemilu, lembaga lain yang cukup *powerful* dalam menjaga Pemilu yang berintegritas adalah Lembaga Peradilan, Lembaga Peradilan dianggap dapat mempengaruhi secara sistematis proses electoral yang berlangsung (Tan, 2023). Mulai dari menggeluarkan peraturan perundang-undangan yang menguntungkan salah satu pihak, sampai dengan lambatnya proses penegakan hukum ketika terjadi kasus pidana dalam proses Pemilu.
3. Banyaknya aktor politik di institusi negara, menjadi salah satu pemicu adanya birokrasi politik. Dimana aktor politik, yang banyak memegang kendali atas institusi negara, memanfaatkan kewenangannya dan menggunakan sarana dan prasarana milik negara untuk kepentingan kelompoknya (Sirait et al., 2022; Tan, 2023).
4. *Technology*

Pada domain ini menjelaskan faktor-faktor yang berkaitan dengan implementasi *technology* dalam partisipasi atau menunjang proses penyelenggaraan Pemilu.

1. Pemanfaatan *Information Communication Technology* (ICT) dalam proses pemungutan dan penghitungan suara. Walupun dianggap dapat membuat penyelenggaraan Pemilu menjadi lebih efisien, namun penggunaan ICT dalam proses *voting* dan *counting* tentunya menghadirkan celah baru untuk melakukan kecurangan (Van Ham, 2020). Transparansi menjadi isu penting dalam pemanfaatan ICT. Walaupun lebih efisien, kehadirannya ICT tidak dapat menggantikan aspek transparansi dari proses pemungutan dan penghitungan suara secara manual, dimana dalam prosesnya dapat diamati secara langsung oleh siapa saja (Tan, 2023).
2. Pemanfaatan Media Sosial dalam proses kampanye. Kemudahan yang diberikan media sosial dalam nyebarkan informasi, dapat dimanfaatkan oleh oknum-oknum tidak bertanggung jawab untuk menyebarkan berita palsu yang mendeskriditkan pihak awan (Van Ham, 2020). Selain itu keterjangkauan media sosial memungkinkan pihak asing untuk menyebarkan berita palsu. Sebagaimana Pemilu yang terjadi di Jerman, Inggris, Spanyol, dan Amerika Serikat, dimana aktor dari Rusia menyebarkan berita palsu untuk menyesatkan preferensi pemilih (Van Ham, 2020).
3. *Process*

Pada domain ini menjelaskan faktor-faktor yang berkaitan dengan sistem penyelenggaraan pemilihan umum.

1. Sistem Pemilu dengan Representasi Majoritarian dinilai memiliki keterkaitan dengan banyakkan temuan pelanggaran Pemilu, hal ini dikarenakan sistem majoritarian membuat taruan dalam kontestasi menjadi lebih besar, dimana partai dengan suara terbanyak menguasai seluruh kursi daerah pemilihan, yang kemudian memotivasi pada aktor politik untuk memenangkan kontestasi dengan segala macam cara. Menurut Sarah Birch (2011) yang dikonfirmasi oleh penelitian yang dilakukan oleh Lehoucq dan Kolev (2015), nilai rapot integritas Pemilu cendrung rendah di negara yang mengadopsi sistem representasi majoritarian dan tinggi di negara yang menerapkan sistem representasi proporsional (Van Ham, 2020).
2. Sistem Kontestasi Bukan Multi Partai, Michael Seeberg mengemukakan bagaimana sistem kontestasi multi partai bisa mempertahankan sistem *check and balance.* Sebagaimana yang terjadi di Indonesia setelah kejatuhan pemerintahan Orde Baru, dimana jumlah partai cukup terfragmentasi sehingga menciptakan ketidakpastian akan partai mana yang akan menang (Tan, 2023).
3. *People*

Pada domain ini menjelaskan faktor yang berkaitan dengan potensi *electoral malpractice* yang disebabkan oleh prilaku dari oknum yang tidak merepresentasikan suatu lembaga. Kurangnya pengalaman dalam menyelenggarakan Pemilu. Pelanggaran dapat terjadi secara tidak sengaja dan tidak dimaksudkan untuk merugikan kelompok, partai, atau kandidat tertentu. Pelanggaran ini seringkali merupakan hasil dari kurangnya kapasitas administratif atau pengalaman dalam menjalankan pemilihan (Tan, 2023).

Dari faktor-faktor pendorong yang dikelompokan ke dalam 5 domain pada *fishbone analysis* di atas, penulis mencoba menarik salah satu faktor yang menjadi fokus pada penelitian ini. Faktor yang dipilih berasal dari domain *environment*, yaitu ketidakhadiran partisipasi *civil society* atau media independen yang berperan sebagai pengawas partisipatif penyelenggaraan pemilu. Sebagaimana termaktub dalam Peraturan KPU No 10 Tahun 2018 Pasal 18 dan Pasal 21, masyarakat berhak ikut serta dalam mengevaluasi, mengawasi, dan memberikan masukan perbaikan terhadap tahapan Pemilu yang sudah dijalankan (KPU, 2018). Selanjutnya Bawaslu, sebagaimana tercantum pada UU No 7 Tahun 2017 Pasal 93, juga mendorong kelompok masyarakat untuk aktif dalam berbagai bentuk pengawasan, hal ini bertujuan untuk menjaga ruang partisipasi yang adil bagi semua pihak (Bagja et al., 2023; DPR RI, 2017).

Salah satu alternatif media yang dapat digunakan masyarakat secara luas untuk melaporkan indikasi pelanggaran Pemilu dan berpertisipasi dalam menjaga integritas Pemilu adalah media sosial. M. Saud mengungkapkan bahwa kolompok pemuda, yang merupakan kelompok yang memiliki porsi suara terbesar dalam Pemilu 2024 di Indonesia, sudah terbiasa dengan partisipasi politik di ruang digital khususnya media sosial, dimana media sosial menjadi sarana demokrasi yang tidak hanya menyediakan ruang untuk berkomunikasi dan berekspresi, namun juga memperkaya pengetahuan dan keterlibatan mereka terhadap situasi politik yang ada (Saud et al., 2023). Pada penelitian yang berbeda M. Saud menemukan adanya korelasi positif antara partisipasi politik di media sosial dan aksi protes kebijakan penyelenggara negara (Saud & Margono, 2021). Sementara M. Mashud mengemukakan media sosial sebagai faktor penentu dalam membentuk pandangan politik bagi para pemuda, selain itu media sosial juga dianggap sebagai sarana yang berguna dalam akses keterlibatan secara komunitas dan melaksanakan hak politik (Mashud et al., 2023).

Penggunaan media sosial untuk partisipasi politik di Indonesia, tidak terlepas dari normalnya penggunaan internet 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[[6]](#footnote-7). 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. Informasi lebih detail lagi menyebutkan bahwa Banten sebagai kota dengan tingkat penggunaan internet tertinggi yaitu 89,10 persen, diikuti oleh DKI Jakarta dengan 86,96 persen, Jawa Barat dengan 81,26 persen, Bali dengan 80,88 persen, Jambi dengan 80,48 persen dan Sumatra Barat dengan 80,31 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[[7]](#footnote-8). 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[[8]](#footnote-9).

Sementara itu salah satu media sosial yang paling banyak digunakan dengan tujuan untuk komunikasi politik 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[[9]](#footnote-10). [**ARGUMENT TENTANG TWITTER MERUPAKAN SOSIAL MEDIA YANG SERING DI GUNAKAN UNTUK PARTISIPASI POLITIK**]

Pelanggaran dalam penyelenggaraan Pemilu merupakan suatu keniscayaan yang bisa saja terjadi. Namun jika penyimpangan tersebut tidak diperbaiki secara bertahap dikhawatirkan bisa berpotensi menimbulkan dampak negatif, seperti, *political distrust* bahkan perpecahan bangsa. (Mauk, 2022; Norris, 2019; Tan, 2023; Van Ham, 2020). Kehadiran *civil society* sebagai pengawas penyelenggaraan pemilihan umum dipercaya dapat meminimalisir praktik pelanggaran Pemilu atau memberikan saran perbaikan secara bertahap pada kontestasi selanjutnya. Media sosial hadir sebagai sarana yang mempermudah *civil society* untuk dapat berpartisipasi dalam politik, termasuk melakukan pengawas penyelenggaraan pemilihan umum. Opini dan aspirasi yang dikemukakan *civil society* terkait integritas Pemilu haruslah mendapat perhatian oleh lembaga Penyelenggara Pemilu, khususnya Bawaslu yang fungsi dan tugasnya beririsan langsung dengan upaya tersebut. Penelitian ini dilakukan untuk membantu Bawaslu dalam menggali opini dan aspirasi yang dikemukakan *civil society* terkait integritas Pemilu di media sosial. Oleh karena itu penelitian ini, sebagai salah satu wujud dari terlaksananya partisipasi publik pada penyelenggaraan Pemilu, mengajukan 2 pertanyaan penelitian, yaitu:

**RQ1:** “Bagaimana pola sentimen publik terhadap integritas penyelenggaraan Pemilu 2024 di Indonesia berdasarkan data teks komentar masyarakat di media sosial Twitter ?”

**RQ2:** “Bagaimanakan pola jenis pelanggaran Pemilu 2024 di Indonensia berdasarkan data teks komentar masyarakat di media sosial Twitter ?”

## Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk dapat menghasilkan model sistem yang dapat mengali pola jenis pelanggaran dan sentimen publik terhadap integritas penyelenggaraan pemilihan umum 2024 di Indonesia. Sehingga dapat memberikan manfaaat kepada para pihak yang berkepentingan seperti Bawaslu untuk memberikan informasi yang berguna dalam memformulasikan strategi pencegahan pelanggaran Pemilu pada kontestasi yang akan datang.

## Ruang Lingkup Penelitian

Batasan pada penelitian ini adalah sebagai berikut:

1. Pada Penelitian ini akan hanya menggunakan data pada sosial media twitter.
2. Data yang diambil merupakan data *tweet* yang berupa komentar masyarakat terhadap Integritas Penyelenggaraan Pemilu 2024.
3. Data *tweet* yang dikumpulkan akan dikempokkan ke dalam 2 jenis klasifinasi. Jenis klasifikasi pertama adalah jenis sentimen *tweet* apakah bersentimen positif, negative, netral atau tidak relevan. Lalu jenis klasifikasi yang kedua adalah klasifikasi pola atau jenis pelanggaran Pemilu sebagaimana akan dijelaskan pada BAB 2 bagian “Pola Pelanggaran Pemilu”.
4. Rentang waktu pengambilan data dilakukan ketika **[TBA]**.
5. Data Tweet yang digunakan pada penelitian ini merupakan tweet berbahasa Indonesia.

## 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 pemilihan umum 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) Bawaslu, dimana informasi yang diberikan berguna dalam memformulasikan strategi pencegahan pelanggaran Pemilu pada kontestasi yang akan datang. (2) Media **[TAMBAHKAN INFORMASI]**.

## Sistematika Penulisan

Sistematika penulisan pada makalah ilmiah ini adalah sebagai berikut:

|  |  |
| --- | --- |
| BAB 1 | PENDAHULUAN |
| BAB 2 | TINJAUAN PUSTAKA |
| BAB 3 | METODOLOGI PENELITIAN |
| BAB 4 | IMPLEMENTASI |
| BAB 5 | HASIL DAN PEMBAHASAN |
| BAB 6 | KESIMPULAN DAN SARAN |

# 

# 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 3CS2. Pada akhir bagian akan ditampilkan kerangka teoritis yang digunakan pada penelitian ini.

## Pola Pelanggaran Pemilu

Terdapat beberapa *framework* yang dapat digunakan dalam menganalisis pola indikasi pelanggaran Pemilu. Diantranya adalah: (1) Jenis Pelanggaran Pemilu yang dikemukakan Sarah Birch pada bukunya *Electoral Malpractice* dan Tingkat Keparahan Pelanggaran Pemilu dari Netina Tanpada bukunya *Electoral Malpractice in Asia*, dan (2) Domain, Variable, dan Indikasi Indeks Kerawanan Pemilu 2024 dari Bawaslu pada dokumen Indeks Kerawanan Pemilu (IKP). Berikut elaborasi *framework* yang telah disebutkan di atas.

* + 1. Jenis dan Tingkat Keparahan Pelanggaran Pemilu

Sarah Birch (2011) mengklasifikasikan berbagai cara pelanggaran Pemilu ke dalam 3 jenis. Pertama, pelanggaran dalam bentuk manipulasi peraturan perundang-undangan. Pelanggaran ini dapat menyasar hak partisipasi warga negara dan hak kontestasi peserta Pemilu. (a) Aturan yang menghalang-halangi warga negara dalam menjalankan haknya, seperti contoh mewajibkan warga negara mendaftar secara manual (Tan, 2023). (b) Aturan yang memanipulasi proporsi suara yang tidak proporsional seperti pembagian daerah pemilihan yang tidak seimbang antara jumlah pemilih dengan jumlah kursi di parlemen akibat aturan yang ugal-ugalan, hal ini merusak prinsip kesetaraan dimana suara sebagian warga negara jauh lebih berbobot ketimbang yang lain (Tan, 2023). (c) Aturan yang membatasi atau melonggarkan hak kontestasi calon peserta Pemilu untuk kepentingan sebagian kelompok (Tan, 2023). Menjadi contoh dari pelanggaran dalam bentuk manipulsi peraturan perundang-undangan.

Kedua, pelanggaran dalam bentuk manipulasi preferensi pemilih. Pelanggaran ini mencedrai prinsip integritas Pemilu, dimana setiap pemilih berhak memberikan suara mereka secara bebas dan rahasia (Tan, 2023). Keputusan pemilih harus terlepas dari adanya sangsi atau imbalan yang akan diterima. Tindak kekesaran atau premanisme yang memaksa pemilih untuk mengubah sikap dukungannya merupakan contoh dari manipulasi preferensi pemilih (Tan, 2023). Tidak hanya menargetkan pemilih, namun juga kandidat yang berkontestasi, kritikus politik, panitia acara politik, atau bahkan jurnalis. Kemudian, tidak seperti kekerasan atau premanisme yang dapat memberikan dampak buruk karena secara kasar melanggar hak pemilih, pelanggaran dalam bentuk membeli suara dukungan (*vote buying*) lebih dapat dimaklumi (Tan, 2023). Selain premanisme dan *vote buying,* mengeksploitasi fasilitas negara untuk kepentingan kelompoknya merupakan contoh lain dari pelanggaran manipulasi preferensi pemilih yang sering ditemukan (Tan, 2023). (a) Mempolitisasi bantuan sosial yang menggunakan dana APBN. (b) Memaksa ASN untuk mendukung calon tertentu karena adanya relasi kuasa. (c) Memberikan prioritas layanan birokrasi bagi kelompok yang mendukung peserta Pemilu yang diusung. (d) Melakukan praktik klientalisme dimana adanya hubungan transaksional antara penyelenggara negara dan pihak swasta yang memiliki kontrak proyek untuk mendukung calon tertentu. (e) Meregulasi untuk menguasai dan mengendalikan arus informasi. Eksploitasi fasilitas negara dapat terjadi diakibatkan oleh banyaknya aktor politik yang menduduki posisi penting dalam pemerintahan (Tan, 2023).

Ketiga, pelanggaran yang dilakukan pada proses pemungutan dan perhitungan suara. Pelanggaran ini dapat disebabkan oleh penyelenggara Pemilu yang tidak netral. Adanya kertas suara yang sudah dicoblos, panitia Pemilu tidak mencatat perolehan suara kandidat yang tidak mereka dukung, pemalsuan hasil akhir penghitungan suara merupakan contoh dari adanya ketidaknetralan Penyelenggara Pemilu (Tan, 2023). Selain itu pelanggaran yang masuk ke dalam jenis ini juga dapat disebabkan oleh adanya upaya dalam menghalang-halangi hak pemilih untuk memberikan suaranya dan hak *civil society* sebagai pengawas ketika proses pemungutan suara berlangsung (Tan, 2023).

Sementara *framework* lainnya adalah tingkat keparahan kecurangan Pemilu. Netina Tan mengelompokkannya ke dalam 4 tingkat. (a) Tingkat 1: pelanggaran secara paksa dalam merebut kekuasaan seperti pembunuhan lawan politik, kudeta, atau pembatalan hasil Pemilu, (b) Tingkat 2: pelanggaran dalam proses pemungutan dan perhitungan suara, (c) Tingkat 3: pelanggaran dalam bentuk mengubah peraturan per-undang-undangan, (d) Tingkat 4: pelanggaran dalam bentuk menguasai sarana dan prasarana informasi (Tan, 2023). Menurut Netina Tan, ketika kecurangan dalam Pemilu merupakan suatu keniscayaan dan tidak dapat dihindari, yang terpenting adalah kadar kecurangannya tidak sampai merusak nilai demokrasi yang sudah sekian lama dibangun, dimana potensi kerusakan itu ada ketika kecurangan tingkat 1 dan 2 terjadi. Berdasarkan definisi di atas dapat dipetakan kerangka jenis pelanggaran Pemilu dari Sarah Birch dan kerangka Tingkat Keparahan Pelanggaran Pemilu dari Netina Tan. Tabel 2.1 mengambarkan pemetaan yang dilakukan.

Tabel 2. 1 Pemetaan Tingkat Keparahan Pemilu dan Jenis Pelanggaran Pemilu

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Tingkat Keparahan Pelanggaran Pemilu (Netina Tan, 2023)** | **Jenis Pelanggaran Pemilu (Sarah Birch, 2011)** | **Contoh Pelanggaran Pemilu** |
| 1 | Level 1 | - | - |
| 2 | Level 2 | Manipulasi Proses Pemunngutan dan Penghitungan Suara | Adanya ketidaknetralan Penyelenggara Pemilu |
| Adanya upaya dalam menghalang-halangi hak pemilih untuk memberikan suaranya dan hak *civil society* sebagai pengawas ketika proses pemungutan suara berlangsung |
| 3 | Level 3 | Manipulasi Regulasi | Aturan yang menyulitkan warna negara mendapatkan hak suara. |
| Aturan yang memanipulasi proporsi hak suara. |
| Aturan yang membatasi atau melonggarkan hak kontestasi calon peserta Pemilu. |
| Manipulasi Preferensi Pemilih | Mengeksploitasi fasilitas, sarana dan prasarana negara untuk kepentingan kelompok tertentu. |
| 4 | Level 4 | Manipulasi Preferensi Pemilh | Mengendalikan arus informasi |

Penelitian ini menjadikan *framework* jenis pelanggaran Pemilu dan tingkat keparahan pelanggaran sebagai rujukan dalam menganalisa pola indikasi pelanggaran Pemilu yang dibicarakan masyarakat di media sosial.

* + 1. Domain, Variable, dan Indikasi Indeks Kerawanan Pemilu 2024 dari Bawaslu

Indeks Kerawanan Pemilu (IKP) merupakan produk hasil riset yang dilakukan Bawaslu dengan dibantu pihak terkait yang di dalamnya memetakan potensi kerawanan pelanggaran dalam penyelenggaraan Pemilu pada masing-masing daerah (Bagja et al., 2023). IKP merupakan wujud tanggung jawab Bawaslu dalam menjalankan fungsinya sebagai lembaga yang ditugaskan untuk tidak hanya melakukan penindakan namun juga pencegahan potensi pelanggaran Pemilu, sesuai dengan apa yang termaktub dalam UU No 7 Tahun 2017 Pasal 94. Dalam internal Bawaslu, IKP berguna sebagai pedoman Bawaslu dalam menyusun perencanaan strategi dan program kerja pencegahan pelanggaran Pemilu. Sedangkan bagi pihak eksternal yang berkepentingan seperti KPU, DKPP, Aparat penegak hukum, media dan masyarakat umum, informasi yang diberikan IKP dapat berguna sebagai referensi dalam upaya pencegahan pelanggaran Pemilu.

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 (Bagja et al., 2023). Tabel 2.2 menjelaskan secara mendetail indikasi-indikasi pelanggaran Pemilu yang dikelompokkan ke dalam subdimensi dan dimensi IKP. Penelitian ini menjadikan IKP sebagai rujukan dalam mengidentifikasi pola indikasi pelanggaran Pemilu yang dibicarakan masyarakat di media sosial.

Tabel 2. 2 Pemetaan Indikasi Pelanggaran Pemilu berdasarkan subdimensi dan dimensinya (Sumber: IKP Bawaslu, 2023).

|  |  |  |  |
| --- | --- | --- | --- |
| **Dimensi** | **Subdimensi** | **No** | **Indikator** |
|
| KONTEKS SOSIAL DAN POLITIK | KEAMANAN | 1 | Adanya kekerasan/kerusuhan berbasis SARA dalam isu-isu politik dan kePemiluan |
| 2 | Adanya kekerasan/kerusuhan yang melibatkan tokoh publik/politik/aparat keamanan |
| 3 | Adanya bencana alam yang mengganggu tahapan Pemilu/Pilkada |
| 4 | Adanya bencana non alam (pandemi covid 19) yang mengganggu tahapan Pemilu/Pilkada |
| 5 | Adanya Intimidasi Terhadap Peserta Pemilu dalam proses pelaksanaan Pemilu/Pilkada |
| 6 | Adanya Intimidasi Terhadap Penyelenggara Pemilu dalam proses pelaksanaan Pemilu/Pilkada |
| 7 | Adanya Intimidasi Terhadap Pemilih dalam proses pelaksanaan Pemilu/Pilkada |
| 8 | Adanya perusakan fasilitas penyelenggaraan Pemilu |
| OTORITAS PENYELENGGARA PEMILU | 9 | Adanya putusan DKPP yang ditujukan kepada Jajaran KPU dan/atau Bawaslu |
| 10 | Adanya Rekomendasi dan/atau Putusan Bawaslu yang tidak ditindaklanjuti oleh KPU |
| 11 | Adanya rekomendasi Bawaslu terkait dengan perubahan suara pada proses rekapitulasi suara |
| PENYELENGGARA NEGARA | 12 | Adanya Rekomendasi Bawaslu terkait ketidaknetralitan ASN/TNI/POLRI |
| 13 | Adanya penyalahgunaan anggaran Pemilu/Pilkada |
| 14 | Adanya intimidasi kepada penyelenggara Pemilu yang dilakukan oleh Kepala Daerah |
| 15 | Adanya himbauan dan/atau tindakan untuk memilih calon tertentu dari pemerintah lokal |
| 16 | Adanya himbauan dan/atau tindakan untuk menolak calon tertentu dari pemerintah lokal atau tokoh masyarakat |
| PENYELENGGARAAN PEMILU | HAK MEMILIH | 17 | Adanya pemilih memenuhi syarat tetapi tidak terdaftar dalam daftar pemilih tetap |
| 18 | Adanya pemilih tidak memenuhi syarat terdaftar dalam pemilih tetap |
| 19 | Adanya penduduk potensial memilih tetapi tidak memiliki KTP-Elektronik |
| 20 | Adanya pemilih ganda dalam daftar pemilih (logistik) |
| 21 | Adanya penggelembungan jumlah data pemilih |
| 22 | Adanya Pemilih Pindah Memilih (DPTB) yang Tidak Dapat Memberikan Hak Suaranya |
| PELAKSANAAN KAMPANYE | 23 | Adanya ketidakprofesionalan Penyelenggara Pemilu yang merugikan kampanye calon |
| 24 | Adanya kampanye yang difasilitasi KPU tidak sesuai dengan ketentuan |
| 25 | Adanya penyelenggara Pemilu yang menunjukan sikap keberpihakan dalam tahapan kampanye |
| 26 | Adanya penyelenggara Pemilu yang melakukan pengabaian terhadap protokol kesehatan dalam kegiatan kampanye |
| PELAKSANAAN PEMUNGUTAN SUARA | 27 | Adanya perlengkapan pemungutan suara yang tidak sesuai dengan ketentuan |
| 28 | Ada keterlambatan perlengkapan (logistik) pemungutan suara |
| 29 | Adanya logistik berupa surat suara pemungutan suara yang tertukar |
| 30 | Adanya informasi tidak terdapat TPS untuk menfasilitasi pemungutan suara bagi pemilih di LAPAS/RS/SEKOLAH DINAS/ |
| 31 | Adanya informasi pelanggaraan saat pemungutan suara di Pemilu/Pilkada |
| 32 | Adanya pemungutan suara ulang di Pemilu/Pilkada |
| 33 | Adanya penghitungan suara ulang di Pemilu/Pilkada |
| 34 | Adanya Pemungutan suara susulan di Pemilu/Pilkada |
| 35 | Adanya saran perbaikan (catatan khusus) dari pengawas saat pemungutan suara |
| 36 | Adanya komplain dari saksi saat pemungutan atau penghitungan suara |
| AJUDIKASI DAN KEBERATAN | 37 | Adanya gugatan atas hasil Pemilu/Pilkada |
| 38 | Adanya keberatan dan/atau sengketa proses Pemilu/Pilkada |
| PENGAWASAN PEMILU | 39 | Adanya TPS yang tidak dihadiri oleh Pengawas TPS pada Pemilu/Pilkada |
| 40 | Adanya Informasi ketiadaan Pengawas Pemilu Tingkat Kelurahan/Desa |
| KONTESTASI | HAK DIPILIH | 41 | Adanya Penolakan terhadap calon perempuan di Pemilu/Pilkada |
| 42 | Adanya Penolakan terhadap calon berlatarbelakang etnis, Suku, dan Agama tertentu di Pemilu/Pilkada |
| 43 | Adanya dokumen palsu dalam proses pencalonan di Pemilu dan Pilkada |
| KAMPANYE CALON | 44 | Adanya informasi kampanye di luar jadwal yang dilakukan oleh peserta Pemilu |
| 45 | Adanya iklan kampanye di luar jadwal yang dilakukan oleh peserta Pemilu |
| 46 | Adanya materi kampanye yang bermuatan SARA di tempat umum |
| 47 | Adanya materi kampanye yang bermuata SARA di sosial media akun lokal (Grup WA / Facebook) |
| 48 | Adanya materi kampanye Hoax di tempat umum |
| 49 | Adanya materi kampanye Hoax di sosial media (Grup WA / Facebook) |
| 50 | Adanya materi kampanye ujaran kebencian di tempat umum |
| 51 | Adanya materi kampanye ujaran kebencian di sosial media (Grup WA / Facebook) |
| 52 | Adanya pelanggaran lokasi kampanye yang dilakukan oleh peserta Pemilu/Pilkada |
| 53 | Adanya konflik antar pendukung peserta Pemilu/pasangan calon |
| 54 | Adanya Laporan tentang Politik Uang yang dilakukan oleh Peserta/Tim Sukses/Tim Kampanye Pemilu |
| 55 | Adanya peserta Pemilu/calon yang tidak melaporkan dana kampanye |
| PARTISIPASI | PARTISIPASI PEMILIH | 56 | Adanya upaya untuk menghalang-halangi Pemilih memberikan suara di TPS |
| 57 | Adanya mobilisasi pemilih tambahan secara mendadak |
| 58 | Adanya pemilih tambahan melebihi jumlah surat suara cadangan 2% |
| PARTISIPASI KELOMPOK MASYARAKAT | 59 | Adanya laporan masyarakat yang tidak ditindaklanjuti oleh Pengawas Pemilu |
| 60 | Adanya Informasi tentang ketiadaan pemantau Pemilu |
| 61 | Adanya mobilisasi penolakan terhadap pelaksanaan Pemilu/pilkada dari kelompok masyarakat |

## Media Sosial

Menurut Kaplan dan Haenlein (2010) media sosial adalah suatu istilah yang terlahir dari 2 perpaduan 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.* (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 professional. Dari ketiga kriteria tersebut kita dapat memisahkan beberapa jenis konten dan informasi yang tidak masuk kedalam kategori *user-generated content* seperti: (1) email dan *instant messager* yang tidak lolos dalam kreteria pertama. (2) Memposting berita tanppa memodifikasinya dan menyebarkannya ke internet juga bukan masuk kedalam *user generated content* berdasarkan kreteria kedua dan (3) Semua konten yang mengandung tujuan komersial tidak masuk kreteria ketiga. Dengan demikian dapat disimpulkan bahwa merupakan plikasebasis 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 ketika 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

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

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, dan *engagement tracking* di media sosial*.*

1. Application

Analisa *apps layer* dilakukan dengan tujuan untuk mengukur dan mengoptimalkan kepuasan pengguna terhadap penggunaan aplikasi media sosial.

1. Hyperlinks

Analisa *hyperlinks layer* dilakukan dengan menganalisa dan menginterpretasikan keberadaan *hyperlink* pada media sosial, baik *in-links* maupun *out-links*.

1. Locations

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

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

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* (saat ini)*, diagnostic analytics* (masa lalu)*,* dan *predictive analytics* (masa depan). Contoh dari penerapannya adalah *recommender system, decision modeling* dan *expert system.*

## 2.3 CRIPS-DM

*Cross Industries Standard Process- Data Mining* (CRIPS-DM) adalah prosedur standar dari proses ekstraksi informasi penting yang berasal dari data dengan jumlah yang besar (*data mining*) yang dapat digunakan di seluruh sektor industri (Schröer et al., 2021). Prosedur standar yang ditemukan pada tahun 2000 ini memiliki 6 tahapan yang bersifat iteratif, yaitu *business understanding, data understanding, data preparation, modelling, evaluation,* dan *deployment.* Dimana ke 6 tahapan tersebut bisa diulangi ketika hasil yang didapat belum sesuai dengan yang diekspektasikan. Selain itu kerangka kerja CRIPS-DM juga dapat dikombinasikan dengan metodologi *agile* (Grady et al., 2017). Hal ini tentunya dapat memecahkan masalah yang berasal dari penerapan pendekatan yang sangat liniear dari metodologi *waterfall*. Berikut adalah penjelasan dari keenam tahapan pada kerangka kerja CRIPS-DM:

1. Business Understanding

Pada tahapan ini ditentukan tujuan utama dari proses *data mining.* Kemudian jenis *data mining,* apakah jenis *classification* atau *regression* yang akan dilakukan*.* Kemudiankriteria kesuksesan proyek, seperti nilai ambang batas akurasi yang diharapkan*.* Selain itu penilaian terkait situasi bisnis dilakukan untuk mengukur sumber daya yang dibutuhkan dalam mengeksekusi proyek. *Outcome* dari tahapan ini adalah *project plan* yang meliputi schedule dan sumber daya yang dibutuhkan.

1. Data Understanding

Pada tahapan ini dilakukan perencanaan data yang akan dikumpulkan seperti menentukan kriteria, sumber data, dan period waktu pengumpulan. Kemudian proses pengumpulan data dieksekusi. Data yang dikumpulkan akan diperiksa kualitasnya dan ditentukan atribute yang akan digunakan. Yang terakhir memberikan label / deskripsi pada data yang telah disaring.

1. Data Preparation

Pada tahapan ini dimulai dengan mendefinisikan kriteria yang digunakan untuk menseleksi data. Proses *cleansing data* dilakukan untuk mengeliminasi data dengan kualitas yang buruk, menormalisasi data agar data lebih konsisten dan memberikan hasil yang lebih optimal.

1. Modelling

Pada tahapan ini diawali dengan memilih teknik pemodelan yang akan digunakan, membuat skenario pengujian dan membangun model itu sendiri. Pemilihan teknik pemodelan tentunya harus selaras dengan tujuan proses *data mining* yang ditentukan pada tahapan *business understanding.*

1. Evaluation

Pada tahapan ini hasil prediksi dari model akan divalidasi terhadap *business objective* yang telah ditentukan pada tahapan *business understanding*. Apakah prediksi yang dihasilkan memenuhi syarat ambang batas tingkat akurasi yang diharapkan. Ketika tingkat akurasi yang dihasilkan belum mencapai standar yang ditentukan maka dibutuhkan penyesuaian. Ketika dibutuhkan penyesuaian maka proses akan diulang dari tahapan yang sebelumnya telah dilakukan.

1. Deployment

Tahapan ini dilakukan ketika model yang dihasilkan telah berhasil memberikan hasil tingkat akurasi yang diharapkan. Model yang dihasilkan akan ditanam pada *environment* khusus yang digunakan untuk keperluan *production.* Tahapan ini terdiri dari perencanaan *deployment*, eksekusi *deployment,* pemantauan dan pengelolaan model pada *environment production*.

Penelitian ini akan menggunakan kerangka kerja CRIPS-DM dikarenakan cocok dengan tujuan penelitian yaitu menggali wawasan tentang sentimen dan pola diskusi masyarakat terkait integritas Pemilu yang datanya berupa teks, memiliki ukuran yang besar dan bersumber dari sosial media.

## 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 penelitiana 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, 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 integritas penyelenggaraan Pemilu di media sosial.

## 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 *data collection,* menurut Zhai (2013) terdapat 3 fase dalam proses *text analytics*, yaitu *text preprocessing*, *text representation*, dan *knowledge discovery.* Berikut penjelasan lebih detail terkait dengan ke 4 fase tersebut:

### Data Collection

Akuisisi data adalah tahap awal dalam setiap studi dan pengembangan *data science*. 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 (Khder, 2021). Sementara itu ada 2 cara yang sering digunakan untuk mengumpulkan data mentah dari media sosial. i) menggunakan *Application Programming Interface* (API) yang disediakan oleh situs media sosial atau ii) menggambil informasi menggunakan proses *web scraping* (Zafarani et al., 2014). *Web scraping* atau *web crawling* yang juga dikenal sebagai ekstraksi *web*, adalah teknik untuk mengekstrak data dari halaman web dan menyimpannya ke dalam sistem file atau database untuk pengambilan atau 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 request HTTP untuk mengakuisis data dari website yang dituju. Request HTTP yang digunakan dapat berisikan sebuah URL ditambah dengan query metode GET atau POST yang dibutuhkan.
2. Mengirimkan request HTTP 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 s eperti 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).

### 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 *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[[10]](#footnote-11),[[11]](#footnote-12).

2.5.2.3 Stemming

*Stemming*, secara harfiah, adalah proses memotong cabang-cabang pohon hingga hanya tersisa batangnya saja. Dalam konteks *text mining*, *stemming* adalah teknik untuk menghilangkan imbuhan dari sebuah kata, sehingga hanya tersisa kata dasarnya (Hardeniya et al., 2016). 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.

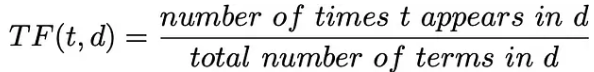
2.5.2.4 Lematisasi

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, 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 "ate" akan menghasilkan 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.*

### 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.*

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 atau 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).

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 kesediaan 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 membangun model matematis dari dataset yang mencakup input yang berupa fitur-fitur dan variable dependen (sering disebut kelas atribut) yang nilainya sudah diketahui. 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 (Laurie A. Schintler; Connie L. McNeely, 2022). Maka proses *supervised machine learning* bisa dibagi menjadi 2, yaitu proses pembentukan model yang disebut sebagai proses induksi dan proses pemanfaatan model yang terbentuk 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 menggunakan *supervised machine learning* hanya digunakan untuk tugas *classification,* yaitu untuk memprediksi sentiment masyarakat terhadap integritas penyelenggaraan Pemilu. 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 *predictive analytics* tertua (Hardeniya et al., 2016). Berdasarkan fitur yang tersedia, algoritma ini akan mengolahnya dan membuat pohon logika. Lebih dari satu *decision tree* dapat dibentuk berdasarkan dataset yang sama walaupun hasilnya memiliki struktur logika yang berbeda. *Decision tree* yang berbeda struktur dapat memberikan hasil yang prediksi yang sama jika dihasilkan dari data latih yang sama (Zafarani et al., 2014). Gambar 2.1 memberikan contoh *decision tree* yang terbentuk dari data latih yang tunjukkan pada Tabel 2.4. *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 coda diprediksi. Prediksi dihasilkan dengan mengikuti jalur logika yang dimulai dari *root node* dan berakhir sampai dengan *leaf node*. Jalur logika yang dilewati, berdasarkan nilai dari 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 |

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

Gambar 2. 1 Dua struktur *decision tree* yang berbeda yang terbentuk dari dataset yang sama

*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 sebagi *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 untuk *predictive analytics*, 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. 2 Perbandingan kemungkinan hyperplane yang dapat digunakan dengan hyperplane yang optimal pada ruang 2 dimensi

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[[12]](#footnote-13). 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 *Y* 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 penentuan 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. 3 Contoh prediksi dengan algoritma KNN

KNN banyak digunakan dalam implementasi *predictive analytics* data besar, hal tersebut dikarenakan selain mudah untuk terapkan dan cukup sederhana, KNN juga dapat memberikan hasil *predictive analytics* 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[[13]](#footnote-14)

2.5.3.2.5 *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:

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. 4 Struktur model deep learning

#### 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*. *Confusion matrix* mengkalkulasi 4 kemungkinan prediksi, yaitu:

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

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | *Prediction* | | | |
| Positif | | Negative | |
| *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 2.6 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 baik suatu model ketika memprediksi *instance* ke dalam kategori positif, secara aktual kategori *instance* tersebut memang positif. |  |
| *Recall* menunjukkan seberapa baik suatu model dalam menemukan *instance* yang secara aktual memiliki kategori positif |  |
| *F1-measure* merupakan harmonisasi dari *precision* dan *rall* |  |

Ketiga *performance metrics* tersebut memiliki rentang nilai [0, 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 pelatihan 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 *k* partisi (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 valudation.*
* *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. 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 (Jelodar et al., 2019). 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. 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 (Jelodar et al., 2019):

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

## Penelitian Terdahulu

Bagian ini menjabarkan beberapa penelitian terdahulu yang 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 sentiment 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

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* 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 negative. 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 Twetter. 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, negative, 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

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 negatef sebanyak 1553 *tweet.* Selain itu penelitian ini juga memaparkan fitur yang dominan muncul, fitur tersebut dapat dimanfaaatkan 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, negative, netral dan tidak relevan. Proses ini menghasilkan 2697 *tweet* positif, 1565 *tweet* negative, 799 *tweet* netral dan 10063 *tweet* yang tidak relevan dengan topik.

* + 1. Preprocessing

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. Modelling

Penelitian ini mengkomparasikan metode SVMdan *Multilayer Perceptron* (*Atrificial 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 yang hanya menyentuh angka 1.565 *tweet*.

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

Penelitian yang dilakukan oleh Blablablabla 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 crawling* 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 *posting-*an Pemprov DKI di media sosial Facebook dan berita tentang PPKM di Jakarta dari beberapa *channel* TV nasional.

* + - 1. Anotasi

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

* + - 1. Preprocessing

Beberapa teknik *preprocessing* teks 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 kometar 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. Modelling

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* ke 4.4% *point*. Penelitian ini juga menghasilkan wawasan berupa opini *concern* 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 demograsi dukungan masyarakat pada kontestasi pemilihan presiden Indonesia 2019. Analisis demografi dan sentimen yang dilakukan memanfaatkan pendekatan *machine learning predictive analytics* 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. Pengumpullan 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

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. Modelling

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 hasilnya adalah: untuk 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 namun lebih aktif dibandingkan dengan pendukung paslon 02.

## Ringkasan Penelitian Terdahulu

Tabel 2.7 meringkas daftar penelitian terdahulu yang disebutkan pada bagian sebelumnya. Ringkasan disajikan dengan mengkomparasikan dataset, *text representasi*, metode klasifikasi, metode evaluasi model dan metode pembentukan topik.

Tabel 2. 7 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. | 894 *tweet* berbahasa Indonesia.  *dependen variable* datasetmemiliki 3 nilai nominal: positif, negatif dan netral | 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.* | *Latent Dirichlet Allocation* (LDA). |
| 2 | Analisis Sentimen Terhadap Upaya Pemerintah dalam Mengatasi Masalah Pengangguran di Indonesia Selama Masa Pandemi COVID-19 | 6859 *tweet* berbahasa Indonesia  *dependen variable* datasetmemiliki 4 nilai nominal: positif, negatif, netral, dan tidak relevant. | Text Representation:   * TF-IDF (unigram dan bigram).   Metode Klasifikasi:   * *Naive Bayes* * *Decision Tree* * *Random Forest* | Metriks: *accuracy, precision, recall.*  5-*fold cross validation.* | [TBA] |
| 3 | Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter. | 15.000 *tweet* berbahasa Indonesia.  *dependen variable* datasetmemiliki 4 nilai nominal: positif, negatif, netral, dan tidak relevant. | Text Representation:   * TF-IDF (unigram dan bigram).   Metode Klasifikasi:   * *Artificial Neural Network* * SVM | Metriks: *accuracy, precision, recall* dan *F1-score.* | [TBA] |
| 4 | Klasifikasi Sentimen Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) di Jakarta Menggunakan Model Machine Learning. | 3.583 komentar pada *platform* Facebook dan YouTube berbahasa Indonesia.  *dependen variable* datasetmemiliki 3 nilai nominal: positif, negatif, dan netral. | 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*  10-*fold cross validation* | [TBA] |
| 5 | Analisis Demografi Pendukung Calon Presiden dan Calon Wakil Presiden Indonesia 2019 Pada Twitter Berbahasa Indonesia. | 199,727 *tweet* berbahasa Indonesia dari total 16,170 akun. | Text Representation:   * TF-IDF (unigram)   Metode klasifikasi usia, lokasi dan jenis kelamin:   * *Naïve Bayes* * *Logistic Regression* * SVM * *Random Forest*   Metode klasifikasi sentimen: *SentiStrength* | *F1-score* | [TBA] |

## Perbadingan Penelitian Terdahulu

Tabel 2.8 menunjukkan hasil analisis *compare, contrast, critisize, syntesize, summarize* (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. 8 Analisis 3C2S penelitian tentang Analisis Sentimen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Judul Penelitian | *Compare* | *Contrast* | *Criticize* | *Syntesize* |
| 1 | Analisis Sentimen dan Pemodelan Topik Studi Kasus Komisi Pemberantasan Korupsi | 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 | 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.* | [TBA] |
| 3 | Analisis Sentimen Terhadap Kembalinya Pembelajaran dengan Sistem Tatap Muka melalui Media Sosial Twitter. | 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 memlatih 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 ANN yang dianggap lebih *powerful* ketimbang *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. | 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. | 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 dll. | (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:***  Studi literatur dari beberapa penelitian terkait memberikan banyak masukan dari penulis. Seperti tidak selalu *machine learning* yang dianggap lebih *powerful* seperti ANN memberikan hasil lebih baik dari *classical machine learning*, namun hal tersebut tampaknya juga terpengaruh dari arsitektur ANN dan teknik representasi teks yang digunakan. Eksperimentasi seperti membentuk fitur baru berdasarkan teori atau rekomendasi penelitian sebelumnya bisa dilakukan untuk meningkatkan hasil akurasi prediksi. Teknik seperti LDA dan frekuensi (*wordcloud* dan *scatterplot*) menjadi teknik yang paling sering digunakan untuk pemodelan topik. Selain itu untuk dapat memberikan wawasan yang lebih bernilai analisa untuk menggali pola demografi juga dapat dilakukan. | | | | | |

## Kerangka Penelitian

Berdasarkan studi lituratur peneliti menyusun kerangka penelitian yang dipresentasikan pada Gambar 2.5. Komponen pada kerangka penelitian yang disusun berdasarkan tahapan *predictive analytics* pada kerangka kerja *Text Mining*. Komponen *data collection* menjelaskan proses pengumpulan data yang dilakukan akan menggunakan pendekatan *web crawling* dengan memanfaatkan Twitter API. Komponen ini berhuubungan dengan komponen tujuan penelitian “Analisis sentimen dan pola indikasi pelanggaran integritas penyelenggaraan Pemilihan Umum Indonesia 2024” dimana komponen tujuan penelitian menentukan kriteria dalam proses pengumpulan data termasuk di dalamnya anotasi. Komponen *text preprocessing* menjelaskan data opini masyarakat yang berupa *text* akan akan melalui tahapan penghapusan *links* dan kerakter *non-alphanumeric, tokenization, stemming, lemmatization, stopword removal* dan *normalization*. Tujuannya adalah untuk membuat data *text* yang lebih konsisten agar dapat diolah pada tahapan berikutnya. Komponen ini terhubung dengan komponen tujuan penelitian, dimana pendekatan *text preprocessing* yang dipilih haruslah sesuai dengan domain dan tujuan penelitian. Komponen *text representation* menjelaskan bentuk terstruktur dari dataset yang awalnya berupa *text* agar dapat diteliti secara kuantitatif dengan formula matemarika dan statistis. *Bag of Words* TF-IDF dalam bentuk unigram, bigram dan trigram menjadi pendekatan *text representation* yang dipilih. Komponen *knowledge discovery* menjelaskan bagaimana data yang sudah terstruktur dianalisa agar menghasilkan suatu wawasan. Pada komponen *knowledge discovery* terdapat subkomponen, yaitu: *topic modelling, classification model* dan *evaluation.* Komponen *topic modelling* menjelaskan penelitian ini akan berupaya mengidentifikasi *concern* masyarakat, dengan menggunakan teknik *Latent Dirichlet Allocation*, terkait dengan integritas penyelenggaraan Pemilihan Umum 2024. 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 *predictive analytics* dari *unlabeled dataset* yang akan diambil pada tahapan pengumpulan data fase kedua, untuk dapat menentukan model dengan tingkat akurasi paling tinggi digunakan beberapa metrik yang terdapat pada subkomponen *evaluation* seperti *precision, recall* dan *f1-score*.

A group of white business cards

Description automatically generated

Gambar 2. 5 Theoretical framework

# BAB 3 METODOLOGI PENELITIAN

Pada bab ini penulis menjelaskan metodologi penelitian yang digunakan. Bab disusun ke dalam beberapa sub bab diantaranya adalah: Rancangan Penelitian, Alur Penelitian dan Instrumen Penelitian.

## Rancangan Penelitian

Penelitian ini menggali opini masyarakat terkait dengan integritas penyelenggaraan Pemilihan Umum 2024 di Indonesia dengan mengukur sentimen masyarakat, pola indikasi pelanggaran dan mengidentifikasi *concern* 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 data *preprocessing, text representation,* klasifikasi menggunakan *machine learning,* evaluasi model dan pemodelan topik. Lebih sepesifik lagi penelitian ini dapat diklasifikasikan sebagai penelitian eksprimental, dikarenakan penelitian ini berupaya mengidentifikasi hubungan sebab akibat dari variable-variable yang menjadi objek penelitian dan juga menjawab tantangan yang muncul ketika menggunakan pendekatan metode penelitian observasi seperti survey, dimana dengan penelitian eksperimental, dilakukan penyesuaian yang diperlukan ketika terdapat faktor pembeda yang muncul (Recker, 2013). 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 pembobotan terhadap suatu fitur atau independen variable yang menjadi faktor pembeda.

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, pola indikasi pelanggaran dan menggali *concern* masyarakat terkait dengan integritas penyelenggaraan Pemilihan Umum. |
| 4 | Hasil Penelitian | Rekomendasi bagi Bawaslu berdasarkan opini masyarakat di media sosial. |
| 5 | Analisis Data | Kuantitatif |
| 6 | Jenis Data | Data Sekunder |
| 7 | Metode Pengumpulan Data | *Web Crawling* |
| 8 | Metode Pengolahan Data | *Data 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 sekunder yang merupakan opini masyarakat terkait integritas penyelenggaraan Pemilihan Umum 2024 dari media sosial twitter*.* Dengan menggunakan teknik *web crawling* melalui twitter API sebagai cara untuk pengumpulan datanya. Metode pengolahan data seperti *tokenization, stopword removal, stemming, lemmatization, normalization* dilakukan pada tahapan *data preprocessing.* Teknik *bag of word* dengan TF-IDF unigram, bigram, dan trigram digunakan untuk *text representation,* hal ini dilakukan supaya data bisa lebih terstruktur dan dapat dioleh pada tahapan pembentukan *machine learning model.* 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, accuracy, recall*, *f1-score* dari masisng-masing model *machine learning*. Selain analisis sentimen dan pola indikasi pelanggaran, dengan tujuan untuk dapat mengidentifikasi *concern* masyarakat baik yang berasal dari sentimen positif maupun yang negatif, maka pemodelan topik juga dilakukan dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Table 3.1 menjelaskan secara ringkas rancangan penelitian yang akan dilakukan.

## Alur Peneliitian

Pada penelitian ini, terdapat 11 tahap yang diawali dengan Pengumpulan data sampai dengan Laporan akhir. Berikut detail dari 11 tahapan yang ada, yang dijelaskan secara lebih pendetail pada tabel 3.2.

Tabel 3. 2 Alur Penelitian

|  |  |
| --- | --- |
| 1. **Pengumpulan Data Awal** | |
| **Input** | Data kasus XYZ |
| **Process** | Studi dokumen dan observasi |
| **Output** | Ekspektasi dan Kenyataan |

|  |  |
| --- | --- |
| 1. **Perumusan Masalah** | |
| **Input** | Ekspektasi dan Kenyataan |
| **Process** | Analisa Kesenjangan |
| **Output** | *Problem Statement* |

|  |  |
| --- | --- |
| 1. **Penentuan Pertanyaan Penelitian** | |
| **Input** | *Problem Statement* |
| **Process** | Analisa akar masalah |
| **Output** | Akar masalah dan Pertanyaan penelitian |

|  |  |
| --- | --- |
| 1. **Studi Literatur** | |
| **Input** | Akar masalah dan Pertanyaan penelitian |
| **Process** | Studi literatur teori, metode penelitian dan analisa 3C+2S penelitian terdahulu |
| **Output** | Akar masalah, Pertanyaan penelitian, Rujukan teori, Rujukan metode dan Sintesa analisa 3C+2S, Kerangka Teoritis. |

|  |  |
| --- | --- |
| 1. **Pengumpulan Data** | |
| **Input** | Akar masalah, Pertanyaan penelitian, *Keyword* pengumpulan data. |
| **Process** | *Web Crawling* – Twitter API |
| **Output** | *Unlabeled Dataset* berupa tweets |

|  |  |
| --- | --- |
| 1. **Membuat Korpus** | |
| **Input** | *Unlabeled Dataset* berupa tweets |
| **Process** | Anotasi |
| **Output** | *Labeled Dataset* (dataset dengan *dependen variable*) |

|  |  |
| --- | --- |
| 1. **Data Preprocessing** | |
| **Input** | *Labeled Dataset* (dataset dengan *dependen variable*) |
| **Process** | *Tokenization, Stemming, Lemmatization, Normalization, Stopword Removal, Text Cleansing* |
| **Output** | *Cleansed tokenized dataset* |

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

|  |  |
| --- | --- |
| 1. **Klasifikasi Sentimen dan Opini tidak relevan** | |
| **Input** | *Space Model Vector* TF-IDF unigram, bigram, trigram (training dataset). |
| **Process** | Pembentukan model dengan *artificial netural network* dan *classical machine learning* seperti SVM, *Naïve Bayes, Decision Tree,* KNN. |
| **Output** | Model Klasifikasi |

|  |  |
| --- | --- |
| 1. **Evaluasi** | |
| **Input** | Model Klasifikasi, testing dataset |
| **Process** | Klasifikasi testing dataset dan evaluasi hasil klasifikasi dengan metrik *accuracy, precision, recall, f1-score.* |
| **Output** | Hasil klasifikasi sentimen, Nilai metrik dari masing-masing model, Model dengan tingkat akurasi tertinggi. |

|  |  |
| --- | --- |
| 1. **Pemodelan Topik** | |
| **Input** | *Space Model Vector* TF-IDF unigram, bigram, trigram (training dan testing dataset). |
| **Process** | Pembentukan pemodelan topik dengan *Latent Dirichlet Allocation* (LDA). |
| **Output** | Topik teridentifikasi |

|  |  |
| --- | --- |
| 1. **Penarikan Kesimpulan** | |
| **Input** | Hasil klasifikasi sentimen, Topik teridentifikasi |
| **Process** | Penarikan kesimpulan dengan |
| **Output** | Saran dan Masukan bagi lembaga ABC. |

### Pengumpulan Data Awal

Penelitian ini diawali dengan tahapan pengumpulan data awal, dimana penulis mengumpulkan dan mempelajari literatur terkait dengan integritas penyelenggaraan Pemilihan Umum. Observasi dan studi literatur yang dilakukan terhadap dokumen terkait menghasilkan kesimpulan bahwa, *civil society* diharapkan untuk dapat berpartisipasi dalam proses pengawasan dan memberikan usulan perbaikan dalam penyelenggaraan Pemilihan Umum. Kemudian, Bawaslu diharapkan dapat menangkap aspirasi, opini dan laporan masyarakat terutama terkait hal tersebut yang mereka sampaikan melalui media sosial *twitter.*

[**ARGUMENT YANG MENJELASKAN BAHWA BAWASLU PERLU MENGANALISA OPINI MASYARAKAT TERKAIT DENGAN INTEGRITAS PEMILU YANG MEREKA SAMPAIKAN DI MEDA SOSIAL TWITTER**]

### Perumusan Masalah

Pada tahapan ini peneliti berupaya untuk dapat mengidentifikasi permasalahan umum yang muncul dalam penyelenggaraan Pemilihan Umum. Maka dilakukan analisa kesenjangan berdasarkan harapan dan kenyataan yang diperoleh pada tahapan sebelumnya. Sehingga dapat disimpulkan bahwa permasalahan utama pada penyelenggaraan Pemilihan Umum adalah “Adanya pelanggaran yang merupakan suatu keniscayaan, namun jika penyimpangan tersebut tidak diperbaiki secara bertahap dikhawatirkan bisa berpotensi menimbulkan dampak negatif, seperti, *political distrust* bahkan perpecahan bangsa (Mauk, 2022; Norris, 2019; Tan, 2023; Van Ham, 2020) ”.

### Penentuan Pertanyaan Penelitian

Pada tahapan ini peneliti berupaya untuk dapat mengidentifikasi cakupan, tujuan, hasil dan pertanyaan penelitian. Berdasarkan permasalahan umum yang ada, peneliti membaginya ke dalam 13 faktor yang berpotensi menjadi akar masalah yang dikelompokkan ke dalam 5 domain, yaitu: *environment, process, technology, people* dan *organization*. Lalu dilakukan analisa akar masalah dengan menggunakan *fishbone analysis diagram.* Hasilnya disimpulkan bahws “Partisipasi *civil society* sebagai pengawas independen dalam penyelenggaraan Pemilu dapat meminimalisir dampak buruk kecurangan Pemilu yang merupakan suatu keniscayaan”. Berdasarkan akar masalah yang telah teridentifikasi peneliti menyusun cakupan penelitian, hasil penelitian, tujuan penelitian dan pertanyaan penelitian.

### Studi Literatur

Pada tahapan ini peneliti berupaya untuk menyusun *theoretical framework* yang selaras dengan pertanyaan penelitian. Database literatur seperti Scopus, Springer, ACM menjadi sumber dalam pencarian rujuan teori dan metode terkait dengan Integritas Pemilu, Pelanggaran Pemilu, Partisilasi Politik dan analisis sentimen. Peneliti juga melakukan analisa 3C+2S (*criticize, compare, contrast, summarize, synthesize*) dari 5 penelitian yang memiliki kesamaan dengan topik penelitian ini yaitu analisis sentimen dengan memanfaatkan pemodelan *machine learning*.

### Pengumpulan Data

Pada tahapan ini peneliti berupaya mengumpulkan data opini masyarakat terkait dengan integritas Pemilu yang sampaikan melalui media sosial twitter. Dengan menggunakan pendekatan *web crawling* melalui Twitter API, proses penarikan data dilakukan pada periode Januari 2024 sampai dengan Februari 2024. Selain itu digunakan *keyword* (‘Pemilu’ AND (‘Luber’ OR ‘Jurdil’ OR ‘Jujur’ OR ‘Adil’ OR ‘Rahasia’ OR ‘Curang’ OR ‘Bawaslu’), untuk dapat memperoleh data yang sesuai harapan yaitu *tweet* terkait dengan Integritas Pemilu 2024. Tabel 3.3 menunjukkan contoh opini masyarakat terkait dengan integritas Pemilu yang mereka utarakan di media sosial *twitter*.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Komentar Masyarakat | Sentimen | Kategori indikasi kerawanan Pemilu berdasarkan dimensi dan subdimensi IKP 2024 Bawaslu | |
| Dimensi | Subdimensi |
| 1 | Hanya karena poster dukungan ke Ganjar saat kunjungan Jokowi, terduga oknum aparat lakukan kekerasan di Wonosari Gunung Kidul, DIY. | Negatif | Sosial Politik | Penyelenggara Negara |
| 2 | Pantesan Pakde mesti berjerih-lelah turun tangan, sementara yang di promosikan terlihat lebih banyak duduk manis. | Negatif | Sosial Politik | Penyelenggara Negara |
| 3 | Ayo selamatkan 8 orang Guru yang kabarnya gaji ditahan gara-gara tidak mau dukung Paslon 02. Terlalu. | Negatif | Sosial Politik | Otoritas Penyelenggara Negara |
| 4 | Tokoh-tokoh senior turun gunung bahas Pemilu 2024, indikasi NKRI sedang tidak baik-baik saja. | Negatif |  |  |
| 5 | dah ga usah ada pemilu aja @KPU\_ID dah @prabowo sm @gibran\_tweet menang pemilu penunjukan langsung. dah..ga usah ada pemilu aja...dah juri udh jadi pemain...susah dah.. @bawaslu\_RI hanya bisa diam. | Negatif | Sosial Politik | Otoritas Penyelenggara Pemilu |
| 6 | gn percaya apa kata Presiden jokowi menjelang pilpres skrg. Dia wanti2 semua pjs dan pejabat di daerah agar netral tp dibelakang dia koordinasi dgn relawan nya agar gibran didukung spy menang. #Pemilucurang di depan mata | Negatif | Sosial Politik | Otoritas Penyelenggara Negara |
| 7 | Mohon edarkan agar masing2 kita mengecek TSP nya di data KPU dan cocokkan dg C1 yg kita miliki utk hindari #pemilucurang | Negatif | Penyelenggaraan Pemilu | Pelaksanaan Pemungutan Suara |
| 8 | Pak mohon infokan ke BPN, di web KPU sekarang C1 sudah jarang di upload. Dimungkinkan ada penggelembungan. Masyarakat awam tidak bisa melihat transparaansinya. | Negatif | Penyelenggaraan Pemilu | Pelaksanaan Pemungutan Suara |

### Membuat Korpus

Pada tahapan ini peneliti berupaya untuk dapat membangun korpus baik *training set* maupun *testing set*. Korpus merupakan sekumpulan *instance* yang terdiri dari *independen variable* dan *dependen variable.* Hasil dari tahapan pegumpulan data, nilai dari *dependen variable* belum ditentukan. Oleh karena itu dilakukan pelabelan untuk menentukan nilai dari *dependen variable*. Kelas sentimen merupakan repersentasi dari *dependen variable* pada penelitian ini, dimana nilai nominalnya adalah (1) positif, (2) negatif, (3) netral, (4) tidak relevan. Maka dari itu peneliti berupaya untuk menciptakan korpus dengan komposisi seperti yang ditunjukkan pada Tabel 3.4.

Tabel 3. 4 Jumlah tweet pelabelan dari proses pelabelan.

|  |  |
| --- | --- |
| **Nilai Nonimal** | **Jumlah Dataset** |
| Opini relevan | 5.000 |
| Sentimen Positif | 2.000 |
| Sentimen Negatif | 2.000 |
| Sentimen Netral | 1.000 |
| Opini tidak relevan | 5.000 |
| Total | 10.000 |

### Data 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, normalization, lemmatization, stemming, stopword removal* dan *text cleansing*. *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. *Normalization* di penelitian ini adalah upaya untuk mengubah menjadi lebih konsisten seperti contoh kata yang dituliskan “senaaangggg” menjadi “senang”. *Stemming* merupakan proses yang mengubah kata menjadi bentuk dasarnya dengan menghapus imbuhan pada kata tersebut, seperti kata “memberikan” menjadi “beri”. *Lemmatization* memiliki tujuan yang sama dengan *stemming* namun dengan mekanisme yang lebih kompleks dengan analisa morfologi atau mendeteksi sinonim kata, seperti contoh kata “ate” menjadi “eat”. *Stopword removal* merupakan upaya untuk menghilangkan kata yang tidak signifikan sesuai dengan tujuan analisis. *Text cleansing* merupakan upaya dalam menghilangkan konten pada teks yang tidak relevan dalam pembentukan fitur seperti URL, karakter spesial, emoticon yang tidak relevan dan sebagainya.

### *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 Processing.*

### Klasifikasi Sentimen dan Opini Tidak Relevan

Pada tahapan ini peneliti berupaya untuk membentuk model *predictive* 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 teknik 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 utnuk 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. Penjelasan masing-masing teknik *machine learning* dapat ditemukan pada bagian 2.5.4 *Knowledge Discovery.*

### Evaluasi

Pada tahapan ini peneliti berupaya untuk 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 *accuracy, precision, recall* dan *f1-score* dari masing-masing model untuk kemudian dikomparasikan dan dipilih model terbaik bagi klasifikasi sentimen dan klasifikasi relevansi data.

### Pemodelan Topik

Pada tahapan ini peneliti berupaya untuk mengidentifikasi topik pada masing-masing kategori sentimen (positif, negatif, netral) dengan menggunakan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA). Topik yang teridentifikasi memberikan gambaran akan *concern* masyarakat yang mereka ungkapkan di media sosial. Topik yang terbentuk akan disesuaikan dengan pola indikasi pelanggaran Pemilu sesuai dengan *framework* yang sering digunakan dalam literatur terkait integritas Pemilu. 3 *framework* yang digunakan pada penelitian ini adalah: (1) Jenis pelanggaran Pemilu dari Sarah Birch (2011), (2) Tingkat keparahan pelanggaran Pemilu dari Netina Tan (2023), (3) Domain, Subdomain, Indikator pelanggaran Pemilu dari IKP 2024 Bawaslu (2023).

### Penarikan Kesimpulan

Pada tahapan ini peneliti berupaya untuk mereprenstasikan hasil penelitian yang dilakukan ke dalam bentuk laporan penelitian. Penyusunan laporan penelitian dilakukan berdasarkan pendekatan penarikan kesimpulan deduktif. Analisa 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 HASIL DAN PEMBAHASAN

# BAB 5 KESIMPULAN DAN SARAN

# DAFTAR PUSTAKA

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

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

Birch, S. (2011). *Electoral malpractice*. Oxford University Press, USA.

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

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

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

Grady, N. W., Payne, J. A., & Parker, H. (2017). Agile big data analytics: AnalyticsOps for data science. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018*-*Janua*, 2331–2339. https://doi.org/10.1109/BigData.2017.8258187

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

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

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

PERATURAN KPU RI NOMOR 10 TAHUN 2018 Tentang Sosialisasi, Pendidikan Pemilih dan Partisipasi Masyarakat dalam Penyelenggaraan Pemilihan Umum, (2018).

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

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

Li, S. S., & Lee, L. C. (2011). Using fishbone analysis to improve the quality of proposals for science and technology programs. *Research Evaluation*, *20*(4), 275–282. https://doi.org/10.3152/095820211X13176484436050

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

Mashud, M., Ida, R., & Saud, M. (2023). Political discussions lead to political efficacy among students in Indonesia. *Asian Journal of Comparative Politics*, *8*(1), 184–200. https://doi.org/10.1177/20578911221143674

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

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

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

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(January), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

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

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://dataindonesia.id/data-Pemilu/detail/bps-kepercayaan-masyarakat-ri-terhadap-Pemilu-masih-rendah [↑](#footnote-ref-2)
2. [berita Jokowi caew-cawe] [↑](#footnote-ref-3)
3. [berita putusan MKMK] [↑](#footnote-ref-4)
4. [berita menteri bagian partai politik hanya mengambil cuti] [↑](#footnote-ref-5)
5. [berita laporan kecurangan vesifikasi parpol yang dilakukan KPU] [↑](#footnote-ref-6)
6. https://apjii.or.id/berita/d/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang [↑](#footnote-ref-7)
7. https://dataindonesia.id/internet/detail/pengguna-media-sosial-di-indonesia-sebanyak-167-juta-pada-2023 [↑](#footnote-ref-8)
8. https://data.goodstats.id/statistic/agneszefanyayonatan/menilik-pengguna-media-sosial-indonesia-2017-2026-xUAlp [↑](#footnote-ref-9)
9. https://datareportal.com/reports/digital-2023-indonesia [↑](#footnote-ref-10)
10. https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist [↑](#footnote-ref-11)
11. https://rahmadya.com/2019/04/24/stopword-berbahasa-indonesia/ [↑](#footnote-ref-12)
12. https://en.wikipedia.org/wiki/Conditional\_probability [↑](#footnote-ref-13)
13. https://en.wikipedia.org/wiki/Lazy\_learning [↑](#footnote-ref-14)