

UNIVERSITAS INDONESIA

Analisa Sentimen untuk Prediksi Dukungan terhadap Calon Presiden Indonesia 2024

KARYA AKHIR

AHMAD FAUZI

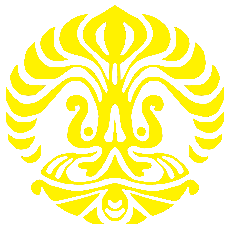
2106806630

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI MAGISTER TEKNOLOGI INFORMASI

JAKARTA

SEPTEMBER 2023



UNIVERSITAS INDONESIA

Analisa Sentimen untuk Prediksi Dukungan terhadap Calon Presiden Indonesia 2024

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

SEPTEMBER 2023

# 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 : | <nama> |
| Program Studi : | Magister Teknologi Informasi |
| Judul : | <judul> |
| Pembimbing : | …………………………………………  ………………………………………… |

Tahun 2024 merupakan tahun politik bagi bangsa Indonesia, pasalnya warga negara Indonesia akan memilih calon pemimpin yang diharapkan dapat mewujudkan aspirasi yang mereka miliki. Sejak dideklarasikannya ketiga calon presiden Indonesia 2024, masing-masing kubu mulai menganalisa situasi politik dan merumuskan strategi untuk dapat memenangkan calon yang mereka usung. Dalam proses analisa situasi dan langkah politik dibutuhkan pengukuran tingkat elektabilitas dan identifikasi faktor yang mempengaruhinya sebagai dasar acuan pengambilan keputusan. Saat ini survey tradisional merupakan cara yang paling popular digunakan, namun penerapannya dalam mengukur dinamika politik dinilai masih memiliki kendala seperti i) membutuhkan waktu yang lama, ii) biaya yang mahal, iii) prediksi yang dihasilkan kurang akurat dan iv) kurang dapat menggali opini masyarakat yang beragam. Analisa sentiment masyarakat terkait pemilu di *platform* media sosial seperti Twitter dianggap dapat menjadi solusi atas permasalah tersebut. Penelitian ini mengkomparasikan teknik *deep learning* dengan *classical machine learning* untuk menghasilkan model sistem yang dapat memprediksi tingkat elektabilitas dan teknik pemodelan topik *Latent Dirichlet Allocation* (LDA) untuk mengetahui faktor yang mempengaruhi dukungan warga negara terhadap calon presiden yang berkontestasi pada pemilihan presiden 2024. Model yang dihasilkan diharapkan bisa memberikan wawasan peta politik masing-masing kandidat dengan cepat, namun juga tetap bisa memberikan prediksi yang akurat.

Kata kunci: ……………., ……………., …

# ABSTRACT

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

<An abstract consists of three parts: problem definition; theories, methods, or data employed to solve the problem(s); findings or results.>

Keywords: ……………., ……………., …

# DAFTAR ISI

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

[**HALAMAN PERNYATAAN ORISINALITAS** ii](#_Toc15524518)

[**HALAMAN PENGESAHAN** iii](#_Toc15524519)

[**KATA PENGANTAR** iv](#_Toc15524520)

[**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS** v](#_Toc15524521)

[**ABSTRAK** vi](#_Toc15524522)

[**ABSTRACT** vii](#_Toc15524523)

[**DAFTAR ISI** viii](#_Toc15524524)

[**DAFTAR GAMBAR** ix](#_Toc15524525)

[**DAFTAR TABEL** x](#_Toc15524526)

[**DAFTAR LAMPIRAN** xi](#_Toc15524527)

[**BAB 1 PENDAHULUAN** 12](#_Toc15524528)

[1.1 Latar Belakang 12](#_Toc15524529)

[1.2 Rumusan Masalah 15](#_Toc15524530)

# 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 dimana Indonesia akan mencatatkan sejarah barunya. Pasalnya pada tahun tersebut akan berlangsung pemilihan umum presiden. Pemilihan Presiden merupakan peristiwa politik dimana setiap warga negara memilih calon pemimpin yang diharapkan dapat mewujudkan aspirasi yang mereka miliki (Gunawan et al., 2019). Sehingga hal ini membuat setiap warga negara sangat antusias dalam mengikuti perkembangan politik yang ada[[1]](#footnote-1). Terdapat 3 kandidat yang akan berkontestasi dalam ajang ini[[2]](#footnote-2). Pertama adalah Anies Baswedan. Mantan Gubernur DKI Jakarta yang menjabat pada periode 2017-2022 ini dikenal dengan ide dan gagasannya terkait perubahan. Kedua ada Ganjar Pranowo yang merupakan Gubernur aktif Jawa Tengah. Beliau merupakan calon yang didukung oleh *ruling party* PDI-Perjuangan dan terkenal akan kedekatannya dengan masyarakat umum. Ketiga ada Prabowo Subianto. Beliau merupakan Menteri Pertahanan aktif dan seorang Jendral di masa orde baru yang terkenal akan jiwa patriotismennya dan kepeduliannya tentang masalah ketahanan pangan.

Sejak dideklarasikannya ketiga calon presiden tersebut, masing-masing kubu mulai merumuskan strategi untuk dapat memenangkan calon yang mereka usung2. Dengan tujuan untuk meningkatkan elektabilitas dan memperkenalkan ide gagasan, maka langkah politik seperti: bersafari politik, menghadiri undangan debat di universitas, bersilaturahmi ke para tokoh penting terus menerus dilakukan, bahkan tukar tambah kekuatan politik dengan masuk dan keluarnya partai pengusung telah dan masih mungkin terjadi sampai masa kampanye berakhir. Peristiwa dan langkah politik tentunya dapat menghadirkan konsekuensi seperti meningkat atau menurunnya tingkat elektabilitas masing-masing calon (Ottati et al., 2016). Oleh karena itu setiap kubu harus menganalisa setiap langkah politik yang mereka ambil. Dalam proses analisa langkah politik dibutuhkan pengukuran tingkat elektabilitas dan identifikasi faktor yang mempengaruhinya sebagai dasar acuan pengambilan keputusan (Valentino et al., 2017). Hasil dari analisa tersebut bisa dijadikan bahan pertimbangan bagi para kubu dalam merumuskan kembali strategi politik kedepannya (Arafat et al., 2020).

Sejauh ini untuk mengukur tingkat elektabilitas dan mengidentifikasi faktor yang mempengaruhinya, para kubu memanfaatkan wawasan yang dihasilkan dari survey yang dilakukan[[3]](#footnote-3). Hasil survey telah mempengaruhi langkah politik masing masing kubu, hal tersebut diakui langsung oleh salah satu calon presiden yang berkontestasi pada pemilu 2024 yaitu Anies Baswedan[[4]](#footnote-4). Pada wawancara yang dilakukan dalam acara Mata Najwa, Anies menyebutkan bahwa survey elektabilitasnya yang kecil di wilayah Jawa Timur membuat kubunya harus menyusun strategi baru untuk meningkatkan peluangnya. Menjadikan keputusan untuk memilih Muhaimin Iskandar sebagai calon wakil presiden merupakan keputusan yang tepat dan rasional. Muhaimin Iskandar yang merupakkan pimpinan Partai Kebangkitan Bangsa (PKB) memang dikenal memiliki basis suara yang sangat besar di wilayah Jawa Timur.

Survey tradisional merupakan cara yang paling popular digunakan pada konteks pengukuran elektabilitas, analisa demografi pemilih dan identifikasi faktor pemilihan2. Namun dewasa ini penerapan survey tradisional dinilai kurang begitu relevan dengan kondisi yang ada pada saat ini. Robert dalam tulisannya menyebutkan ada 3 era perkembangan survey (Groves, 2011). Era pertama dan kedua disebut sebagai era survey tradisional. (1) Era pertama dikenal sebagai Era Penemuan (1930-1960). Pada era penemuan ini komponen dasar dari rancangan survey seperti proses pengumpulan data dan alat ukur yang dapat menghasilkan informasi statistik ditemukan. Pada era ini juga banyak lembaga survey yang didirikan baik yang berasal dari swasta, akademis bahkan sektor pemerintah. (2) Era kedua dikenal sebagai Era Ekspansi (1960-1990). Pada era ekspansi ini pengembangan dari sisi metode survey telah banyak dilakukan. Metode survey kuantitatif untuk mempelajari perilaku konsumen telah banyak dimanfaatkan pada era kedua ini. (3) Era ketiga dikenal sebagai Era Data Organik Sebagai Pendukung (1990 - saat ini). Pada era ketiga ini data yang tercipta dari setiap proses transaksi digunakan sebagai data pendukung.

Pada 2 era pertama yang juga disebut sebagai era survey tradisional, tantangan seperti menurunnya *response rate* dan meningkatnya *coverage error* menjadi faktor pendorong yang melahirkan era selanjutnya (Groves, 2011).

1. *Response rate* yang semakin menurun menyebabkan semakin tingginya biaya yang dibutuhkan untuk melakukan *survey.* Khususnya *Household Survey* yang saat itu dianggap sebagai metode dengan *response rate* yang tinggi. Hal ini dikarenakan semakin tinggi *non-response* yang dialami maka semakin tinggi usaha yang dibutuhkan untuk menghubungi dan mewawancara responden.
2. *Coverage error* yang meningkat disebabkan oleh pergeseran penggunaan alat komunikasi dari telpon rumah ke telpon genggam. Tidak diregulasinya lokasi penggunaan telpon genggam, membuat hilangnya informasi penting lokasi geografis dari responden. Hal ini akan memberikan dampak yaitu sulitnya menganalisa pola demografi berdasarkan area geografis.

Ditambah lagi dalam praktiknya cukup banyak hasil survey yang tidak tepat dalam memprediksi hasil pemilihan umum. Salah satu contohnya adalah yang terjadi di pemilu Amerika Serikat tahun 2016. Pada pemilu tersebut Nicolas et. al menyatakan bahwa ada 2 *low-tech* faktor yang menjadi penyebab mengapa survey yang dilakukan gagal memprediksi hasil pemilu yang sebenarnya (Valentino et al., 2017). 2 faktor tersebut adalah *sampling error* dan *inaccuracy likely-voters model.*

1. *Sampling error* yang dimaksud adalah *measurement bias,* dimana responden tidak jujur dalam menyatakan preferensinya dalam pemilihan.
2. Sedangkan *inaccuracy likely model* merupakan ketidaktepatan pembobotan yang diberikan terhadap kategori sample tertentu. Pembobotan diberikan berdasarkan kecendrungan populasi dari suatu kategori sample memberikan suaranya ketika pemilu berlangsung. Pada pemilu Amerika Serikat 2016, kelompok berpendidikan rendah dan kaum pekerja kulit putih kurang mendapatkan bobot yang semestinya, karena kecendrungan 2 kelompok tersebut untuk sedikit memberikan suaranya ketika pemilu berlangsung.

Nicolas juga menyebutkan perbaikan untuk meminilalisir kesalahan yang sama di masa depan telah dilakukan. Untuk mencegah *measurement bias,* usaha seperti memilih pewawancara dengan *gender* dan ras yang sama dengan responden dianggap cukup efektif. Usaha ini dianggap dapat memberikan rasa lebih aman dan nyaman bagi responden, sehingga mereka dapat memberikan pandangan yang sebenarnya mereka miliki, tanpa perlu khawatir pandangan yang mereka berikan dapat menyinggung pewawancara (Valentino et al., 2017).

Lebih lanjut lagi terkait dengan kendala survey tradisional diungkapkan oleh Reveilhac et. al dalam penelitiannya bertajuk tinjauan literatur secara sistematis terhadap pemanfaatan data media sosial sebagai pelengkap survey tradisional (Reveilhac et al., 2022). Reveilhac menyatakan bahwa:

1. Kekurangan survey tradisional dibandingkan dengan analisa menggunakan *social media data* adalah survey tradisional membutuhkan waktu yang tidak sebentar dalam prosesnya sehingga sulit bagi peneliti untuk dapat menangkap dinamika perubahan opini dengan cepat.
2. Dalam menggali *public opinion,* survey tradisional kerap dibatasi dengan lingkup pertanyaan. Pembatasan ini memberikan ruang sempit untuk beropini secara bebas (kecuali untuk pertanyaan *open-ended*), membuat keragaman jawaban terbatasi oleh sekumpulan data yang bersifat diskrit.

Dari sejumlah kendala yang disebutkan sebelumnya menjelaskan bagaimana teknik survey tradisional tidak lagi cukup efektif untuk dapat membantu para kubu yang berkontestasi dalam menganalisa peta politik. Sehingga cara baru dibutuhkan untuk dapat membantu mereka dalam mengukur elektabilitas calon presiden dan mengidentifikasi faktor yang mempengaruhinya, agar tetap selaras dengan dinamika politik yang ada.

## Rumusan Masalah

Analisa kesenjangan, yang tampilkan pada tabel 1.2, dilakukan berdasarkan fakta empiris hasil penelitian tentang survey tradisional yang dipaparkan pada bagian latar belakang. Tujuannya untuk dapat mengidentifikasi masalah umum dengan membandingkan kenyataan yang saat ini ada dengan harapan yang belum tercapai terhadap praktik pengukuran elektabilitas calon presiden dan identifikasi faktor yang mempengaruhinya.

|  |  |  |
| --- | --- | --- |
| Harapan | Kenyataan | Masalah Umum |
| Dalam proses analisa langkah politik dibutuhkan pengukuran tingkat elektabilitas dan identifikasi faktor yang mempengaruhinya sebagai dasar acuan pengambilan keputusan (Valentino et al., 2017). Hasil dari analisa tersebut bisa dijadikan bahan pertimbangan bagi para kubu dalam merumuskan kembali strategi politik kedepannya (Arafat et al., 2020) | Pengukuran elektabilias calon presiden dan faktor yang mempengaruhinya masih dilakukan dengan cara survey tradisional yang dianggap masih memiliki beberapa kelemahan dalam konteks pengukuran dinamika politik. | Penerapan survey tradisional untuk mengukur dinamika politik yang cepat berubah dan opini yang beragam memiliki beberapa kekurangan. Seperti i) membutuhkan waktu yang lama, ii) biaya yang lebih mahal, iii) prediksi yang dihasilkan kurang akurat dan iv) kurang kaya dalam menggali opiniyang beragam. |

Dari masalah umum yang berhasil teridentifikasi pada analisis kesenjangan, kemudian proses penguraian faktor yang menyebabkan masalah umum tersebut dilakukan dengan menggunakan teknik *fishbone analysis* *diagram*.Menerut Li, faktor-faktor penyebab masalah umum dapat dikelompokkan ke dalam 4 domain yaitu proses, manusia, teknologi dan kebijakan (Li & Lee, 2011). Untuk itu penulis mencoba untuk mengklasifikasikan potensi akar masalah yang berhasil teridentifikasi ke dalam domain yang penulis anggap relevan.

A black background with white text

Description automatically generated

Gambar 1.2 Fishbone analysis diagram

Berikut adalah uraian masalah teridentifikasi yang dikelompokkan ke dalam 3 domain:

1. Proses

Pada domain proses terdapat 3 masalah yang sering muncul ketika melakukan penelitian dengan menggunakan metode survey tradisional dalam konteks mengukur elektabilitas calon presiden. (1) Proses survey tradisional membutuhkan biaya yang tidak sedikit, hal ini dikarenakan *response rate* yang terus menurun sehingga biaya dan waktu tambahan dibutuhkan untuk menghubungi dan mewawancara responden. (2) Proses survey tradisional yang membutuhkan waktu yang lama sehingga membuatnya sulit untuk dilakukan dalam frekuensi yang sering dan rentang waktu yang pendek. Hal ini membuatnya kurang cocok digunakan untuk menangkap dinamika politik yang sering berubah. (3) Survey tradisional kurang bisa menggali opini publik yang beragam karena formatnya yang *close-ended* dan dibatasi dengan pilihan jawaban yang bersifat diskrit.

1. Manusia

Pada domain manusia terdapat masalah *sampling error* seperti (1) *Measurement bias.* Hal initerjadi ketika responden tidak memberikan informasi yang jujur, sehingga membuat hasil analisa survey menjadi tidak akurat dan (2) Kemungkinan kesalahan yang Pembobotan likely-voter yang salah, hal ini disebabkan kurang tepatnya peneliti dalam mementukan nilai bobot suatu kategori sample karena dianggap kurang berpengaruh sehingga membuat hasil perhitungan menjadi kurang akurat.

1. Teknologi

Pada domain teknologi pergeseran adopsi alat komunikasi dari telpon rumah ke telpon genggam menimbulkan masalah baru terkait dengan *coverage error*. Penggunaan telpon genggam yang tidak diregulasi berdasarkan lokasi, membuat survey yang dilakukan sulit dianggap realiable dalam mengidentifikasi demografi lokasi sampel.

Dari permasalahan yang dikelompokan ke dalam 3 domain pada *fishbone analysis* di atas, penulis mencoba menarik akar masalah. Masalah yang dijadikan penulis sebagai akar masalah berasal dari domain proses, yaitu proses survey tradisional membutuhkan waktu yang lama. Proses survey yang lama menghambat para pihak yang berkepentingan untuk dapat tetap selaras dengan kondisi politik yang sangat dinamis (Reveilhac et al., 2022). Salah satu alternatif dari survey tradisional yang dianggap bisa memberikan waktu yang lebih cepat dalam proses analisanya adalah analisa sentimen aspirasi masyarakat di media sosial.

Menurut Robert, era ketiga dalam perkembangan survey merupakan bentuk adaptasi dari tantangan yang muncul pada era survey tradisional (Groves, 2011). Dimana kemuculan internet memberikan gairah baru pada dunia penelitian dan memunculkan harapan untuk dapat beradaptasi dari tantangan yang ada. Internet menghadirkan data organik yang dapat dijadikan alternatif bahan penelitian. Data organik dapat tercipta dengan sendirinya, dimana peneliti tidak secara sengaja merancang pertanyaan untuk mendapatkan data tersebut. Contoh produk dari teknologi internet adalah Media Sosial. Media Sosial dapat berperan sebagai *self*-*administer instrument,* dimana memungkinkan pengguna menyatakan pandangan dan ekspresinya secara suka rela, sehingga menghasilkan data yang organik. Ia juga menambahkan dengan media sosial memungkinkan pengumpulan data dilakukan dengan biaya yang murah dan dalam waktu yang cepat.

Sejalan dengan Robert, Ceron dalam penelitiannya mengutip Xin et. al yang menyebutkan bahwa analisis sosial media membutuhkan waktu lebih cepat dan biaya yang lebih murah dibandingkan dengan survey tradisional, dan kemampuan untuk dilakukan pemantauan secara *day-by-day* dan *realtime* memungkinkan peneliti untuk menangkap opini publik terhadap setiap kabar berita yang muncul (Ceron et al., 2014). Dengan demikian membuat analisis media social sangat relevan digunakan pada periode kampenye politik untuk menangkap opini publik. Selain itu analisa sentiment di media sosial juga dapat menjadi solusi masalah *measurement bias.* Kasus *measurement bias* pada pemilu Amerika Serikat 2016 yang dikemukakan oleh Nicolas pada bagian latar belakang. Media sosial memungkinkan pengguna untuk mengeskpresikan dan memberikan pendapat mereka secara *genuine* (Valentino et al., 2017).

Teknologi internet, sebagai *enabler* dari era survey yang ketiga, bukanlah barang yang langka di Indonesia. Survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) selama periode 10 Januari -27 Januari 2023 menyatakan bahwa terjadi peningkatan kebutuhan internet di Indonesia terlebih sejak terjadi pandemi Covid-19 pada 2020 lalu[[5]](#footnote-5). Hal ini juga dikarenakan masih banyaknya perusahaan yang tetap menerapkan kebijakan *Work from Home* (WFH) walaupun Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) sudah dicabut oleh pemerintah, sehingga kebutuhan jaringan internet masing tetap tinggi layaknya ketika kebijakan PPKM masih diberlakukan. APJII mencatat 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. 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 yang merupakan sumber dari data organik yang dapat dijadikan bahan untuk penelitian 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[[6]](#footnote-6). 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 bermain 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[[7]](#footnote-7).

Media sosial juga dianggap sebagai *channel* penting dalam komunikasi politik. Survey yang dilakukan oleh litbang Kompas menyatakan media sosial menjadi media yang paling sering digunakan oleh kostituen partai politik[[8]](#footnote-8). Tercatat konstituen dari partai seperti PSK, Demokrat, Gerindra dan Perindo telah menjadikan media sosial sebagai pilihan utamanya. Sebanyak 63 persen konstituen dari partai PKS, 46,3 persen dari partai Demokrat, 40 persen dari partai Gerindra dan 42,9 persen dari partai Perindo telah menggunakan media sosial setiap harinya. Namun tidak semua mayoritas konstituen partai menjadikan media sosial sebagai *channel* utama mereka. Partai yang cukup mapan seperti Golkar, PPP, PKB dan PDI-P tercatat hanya sekitar 4 persen dari konstituennya yang menggunakan media sosial. Menariknya demografi usia responden yang tidak memilih media sosial tergolong dalam usia tua, sebaliknya untuk responden yang memilih media sosial tergolong dalam usia muda. Temuaan ini mengindikasikan adanya korelasi antara usia pemilih dan jenis media yang digunakan. Demografi pemilu 2024 yang mayoritas didominasi oleh pemilih usia muda menjadi alasan, mengapa media sosial sangatlah penting sebagai alat komunikasi politik.

Gambar 1.1 Demografi pemilih pada Pemilihan Presiden Indonesia 2024

(Sumber: Komisi Pemilihan Umum (KPU), 2023)

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-9). Berdasarkan penelitian dari Reuters yang dilakukan pada 48 negara termasuk Indonesia, twitter merupakan media sosial yang paling banyak digunakan dengan tujuan untuk mencari berita terbaru dengan presntase 25 persen [[10]](#footnote-10). Angka persentasenya lebih tinggi dari beberapa media sosial lainnya, dimana angkanya hanya berkisar 11 sampai dengan 16 persen. Sedangkan 20 persen pengguna mengakses *platform* milik Elon Musk tersebut dengan tujuan melihat debat dan 17 persen menggunakannya untuk mencari perspektif yang berbeda dari berita *mainstream*. Dengan data seperti itu, tahun politik berpotensi menjadikan Twitter sebagai media sosial yang akan paling sering dikunjungi oleh masyarakat Indonesia.

Untuk saat ini Twitter memungkinkan penggunanya untuk mengekspresikan pandangan mereka melalui kiriman teks dengan jumlah maksimal 280 karakter[[11]](#footnote-11). Twitter memungkinkan terjadinya komunikasi 2 arah dimana pengguna lain dapat mengkomentari kiriman yang telah dibuat. sehingga memungkinkan terjadinya komunikasi politik yang sehat. Dalam konteks politik, Twitter sangat cocok digunakan oleh para politisi pengusung masing-masing calon presiden untuk menyebarkan pesan kampanye mereka[[12]](#footnote-12). Sifatnya yang 2 arah memungkinkan politisi secara real-time menganalisa tanggapan masyarakat melalui komentar yang diberikan. Media sosial Twitter sendiri telah dimanfaatkan oleh calon Presiden Indonesia pada pemilu 2019 yang lalu[[13]](#footnote-13). Dimana saat itu calon presiden yang berkontestasi Jokowi dan Prabowo menggunakan twitter dengan strategi yang berbeda. Presiden yang akhirnya terpilih Jokowi menggunakan twitter untuk membagikan informasi terkait program pemerintah yang telah dan akan dijalankan. Sedangkan Prabowo cendrung menggunakan twitter untuk membalas komentar masyarakat pada kirimannya. Ini tidak terlepas dari strategi Prabowo untuk bisa lebih dekat dengan masyarakat umum.

Sementara penelitian yang dilakukan oleh Nicolas menyebutkan pada pemilu presiden Amerika Serikat tahun 2016, Donald Trump menjadi kontestan pertama pada level pemilihan kepala negara telah berhasil memobilisasi dan pada akhirnya mengubah peta dukungan masyarakat Amerika dengan berkampanye di Twitter (Valentino et al., 2017). Banyak kalangan menilai pesan politik yang dibagikan oleh Trump saat itu dalam kampanyenya telah menjadi salah satu faktor yang mengubah keputusan *swing voters,* khususnya mereka yang berasal dari kalangan kaum pekerja kulit putih yang merupakan kelompok sample yang mendapatkan pembobotan kecil pada likely-voter model yang digunakan pada metode survey tradisional. Dengan demikian, seringnya media sosial Twitter digunakan sebagai alat kampanye dan komunikasi politik menjadikannya tempat yang valid untuk mengukur elektabilitas danmengidentifikasi faktor yang mempengaruhi keputusan pemilih untuk mendukung calon presiden yang berkontestasi.

Berdasarkan uraian identifikasi pemasalahan pada latar belakang, penarikan akar masalah dengan *fishbone analysis* dan potensi media sosial sebagai solusi untuk bisa memberikan *predictive analysis* yang cepat maka penulis mengajukan pertanyaan sebagai berikut dalam penelitiannya kali ini:

“Bagaimanakan model yang dapat mengukur elektabilitas calon presiden Indonesia 2024 dan mengidentifikasi faktor yang mempengaruhinya berdasarkan data yang terdapat pada media sosial Twitter?”

## Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk dapat menghasilkan model sistem yang dapat memprediksi dengan cepat tingkat elektabilitas dan faktor yang mempengaruhi dukungan warga negara terhadap calon presiden yang berkontestasi pada pemilihan presiden 2024. Sehingga dapat memberikan manfaaat kepada para phak yang berkepentingan seperti penggerak lembaga survey untuk memberikan informasi dan berita kepada masyarakat maupun masing-masing kubu yang berkontestasi untuk dapat memformulasikan strategi politik untuk memenangkan calon yang mereka usung. Disamping itu masih berpotensinya masing-masing calon presiden pada pemilihan presiden periode yang akan datang (2029-2034) menjadikan data atau model yang dihasilkan diharapkan masih dapat dimanfaatkan pada kontestasi yang akan mendatang.

## 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 para calon presiden pada kontestasi pemilu 2024.
3. Data *tweet* yang dikumpulkan akan dikempokkan ke dalam 7 kategori, dimana 6 kategori digunakan untuk melabeli *tweet* terhadap jenis sentimen positif atau negatif (2 nilai nominal) untuk masing masing calon presiden: Anies Baswedan, Ganjar Pranowo dan Prabowo Subianto (3 kandidat). Dan label yang terakhir digunakan untuk komentar masyarakat yang dinilai netral atau tidak relevan.
4. Rentang waktu pengambilan data dilakukan sampai dengan masa kampanye berakhir.
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 mengukur elektabilitas calon presiden dan mengidentifikasi faktor yang mempengaruhinya ini diharapkan (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) Pengelola lembaga survey, dimana model yang dikembangkan atau hasil penelitian bisa dimanfaatkan untuk bisa memberikan informasi dan berita terkait elektabilitas calon presiden kepada masyarakat. (2) Politisi atau Analis partai politik, model yang dikembangkan atau hasil penelitian bisa dimanfaatkan untuk bisa membantu dalam menganalisa peta politik dengan memprediksi elektabilias dan faktor yang mempengaruhinya sehingga dapat membantu dalam penyusunan stategi dan langkah politik.

## 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

## Media Social

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 tanpa 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 Kaplan dan Haenlein menyimpulkan bahwa media sosial merupakan sekumpulan aplikasi berbasis internet yang dibangun berlandaskan fondasi ideologi dan teknologi dari *web 2.0* dan memungkinkan penggunanya untuk membuat dan membagikan *user-generated content.*

Kaplan dan Haenlein juga membagi sosial media ke dalam 6 kategori berdasarkan 2 dimensi yaitu social presence / media richness dan self-presentation / self-disclosure. Social presence dan media richness diambil dari domain media research. Social presence dipengaruhi oleh 2 hal, yang pertama adalah seberapa intim suatu interaksi sosial berlangsung, seperti contoh komunikasi tatap muka dinilai lebih intim dari komunikasi melalui jaringan telpon, lalu yang kedua seberapa langsung (imidiacy) interaksi sosial dilakukan, seperti contoh komunikasi sinkronus menggunakan live chat lebih baik dalam konteks imidiacy ketimbang komunikasi asinkronus menggunakan email. Sedangkan media richness mengambarkan tingkat keberagaman jenis konten yang dipertukarkan. Sedangkan untuk dimensi self-presentation / self-disclosure diambil dari domain social process, dimana mengambarkan tingkat keterbukaan informasi personal, yang dalam konteks media sosial adalah informasi personal dari si pengguna. Tabel 2.1 menggambarkan keenam kategori yang dimaksud.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Social presence / Media richness | | |
| Low | Medium | High |
| Self-presentation / Self-disclosure | High | Blog | Social Networking sites (e.g: Facebook) | Virtual social worlds (e.g: Second Life) |
| Low | Collaborative projects (e.g: Wikipedia) | Content communities (e.g: Youtube) | Virtual game worlds (e.g: World of Warcraft) |

Berikut penjelasan dari keenam kategori media sosial:

1. Collaborative Projects

Ide utama dari *collaborative projects* adalah menekankan pada kemampuan untuk membuat setiap pengguna dapat secara bersamaan / simultan dalam membuat dan mengubah konten dan informasi dalam satu *platform*. *Collaborative project* masuk dalam karegori *self-presentation / self-disclosure* rendah karena isi konten yang disiarkan spesifik terhadap domain tertentu dan tidak menekankan pada keterbukaan informasi personal dari si pengguna. Sedangkan pada sisi *social presence / media richness* masuk ke dalam kategori rendah karena jenis konten yang ditertukarkan umumnya berbasis teks. Salah satu contoh dari *collaborative project* adalah Wikipedia.

1. Blogs

Media sosial berbentuk halaman web personal, *blog* biasanya membahas banyak hal seperti: diari personal yang menceritakan kehidupan sehari hari dari si penulis sampai dengan pembahasan tentang domain yang disukai atau digeluti si penulis. Secara *social presence / media richness, blog* sama seperti *collaborative projects.* Namun secara *self-presentation / self-disclosure, blog* masuk ke dalam karegori tinggi karena biasanya konten dan informasi yang disiarkan benar-benar menceritakan informasi personal dari si penulis.

1. Content communities

Ide utama dari *content communities* adalah sebagai *platform* yang memungkinkan pengguna untuk bertukar konten dengan jenis yang beragam. Seperti contoh *platform BookCrossing* untuk konten berjenis teks, *Flickr* untuk konten gambar, *Youtube* untuk konten video dan *SlideShare* untuk konten slide *powerpoint*. *Content* communities masuk dalam kategori rendah untuk aspek *self-presentation / self-disclosure* dikarenakan dikarenakan media sosial berjenis ini tidak mewajibkan penggunanya untuk membuat halaman profil personal dari pengguna, jika pun ada biasanya tidak lebih dari informasi *basic* seperti informasi tanggal bergabung atau jumlah konten yang dibagikan. Untuk aspek *social presence / media richness, content communities* masuk ke tingkat sedang karena walaupun cukup interaktif dan dengan jenis konten yang beragam, namun *content communities* tidak pada tingkat mensimulasikan kehidupan sehari hari si pengguna.

1. Social networking sites

Ide utama dari *social networking sites* adalah sebagai *platform* yang menghubungkan pengguna untuk dapat saling terhubung satu sama lain dengan menyediakan fitur seperti membuat halaman profil pengguna dengan informasi yang cukup detail, mengajak pertemanan, dan mengirimkan pesan instan dan tautan publik. Contohnya seperti: Facebook dan Twitter. *Social networking site* masuk ke dalam kategori tinggi pada aspek *self-presentation / self-disclosure* dikarenakan platform ini memungkinkan penggunanya untuk menginformasikan profil penggunanya dengan sangat detail dan memungkinkan penggunanya untuk bisa membagikan kegiatan yang mereka lakukan setiap harinya. Sedangkan pada aspek *social presence / media richness* masuk ke dalam kategori sedang karena sama seperti *content communities,* walaupun cukup interaktif namun *social networking sites* tidak pada tingkat yang dapat mensimulasikan kehidupan sehari hari si pengguna.

1. Virtual game worlds

Ide dari *virtual game worlds* adalah sebagai platform yang dapat mensimulasikan interaksi sosial dari penggunanya dengan aturan ketat yang sudah ditentukan. *Virtual game worlds* biasanya mengusung tema *massively multiplayer online role-playing* game. Dimana masing masing pengguna bermain sesuai dengan karakter, tugas dan fungsi pokoknya, sehingga pada aspek *social presence / media richness* masuk ke dalam kategori tinggai dikarenakan dinilai cukup mensimulasikan kondisi interaksi sosialdi kehidupan sehari hari.

1. Virtual social worlds

Ide dari *virtual social worlds* adalah sebagai platform yang dapat mensimulasikan interaksi sosial dari penggunanya tanpa batas aturan yang ketat. Memungkinkan penggunanya untuk memanifestasikan profil personalnya ke dalam bentuk *avatar* pada platform *virtual social worlds.* Contoh dari *virtual social worlds* adalah *second live* dan *metaverse.*

## 2.2 Analisa Media Sosial

Menurut Khan (2015), Analisa media sosial adalah perpaduan dari seni dan ilmu pengetahuan dalam menggali wawasan 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* pendekatan dari analisa 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*. Aplikasi yang biasa dimanfaatkan dalam Analisa teks media sosial seperti penggalian *sentiment mining, intention mining,* dan *topic & idea mining*.

1. Network

Analisa media sosial yang dilakukan terhadap jejaring hubungan antar entitas di sosial media, seperti jejaring hubungan pertemanan Pertemanan, jejaring hubungan komunitas yang diikuti pengguna. Analisa *network layer* memungkinkan peneliti dan praktisi untuk dapat mengidentifikasi entitas yang berpengaruh, mengidentifikasi struktur jaringan pada media sosial, hal ini bisa bermanfaat dalam identifikasi 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 si 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 mengoptimalkan kepuasan pengguna terhadap *mobile application* yang mereka gunakan. [TBA]

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 terhadap area geografis. Contoh penerapannya adalah analisa dukungan terhadap calon presiden berdasarkan demografis lokasi.

1. Search Engines

Analisa *search engine layer* dilakukan terhadap histori kata kunci yang digunakan terhadap 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* dari data pada media sosial yang berbentuk *video,* gambar, *audio* dan bahkan animasi. Oleh karena itu dibutuhkan teknik khusus yang berbeda dengan analisa media sosial *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 *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* [TBA]

## 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 di ekspektasikan, selain itu kerangka kerja CRIPS-DM juga dapat dikombinasikan dengan metodologi *agile* (Grady et al., 2017). Hal ini tentunya dapat memecahkan masalah yang diturunkan 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*, seperti jenis *data mining* yang akan dilakukan, Apakah *classification* atau *regression* dan kriteria kesuksesan proyek, seperti nilai ambang batas akurasi yang diharapkan*.*  Selain itu penilaian akan 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 pengumpulan data dari sumber data, memeriksa kualitas data, mementukan atribute yang digunakan dan diakhiri dengan memberikan label / deskripsi pada data.

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 atau menormalisasi data agar layak untuk diolah pada tahapan selanjutnya.

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 bisnis yang teridentifikasi 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 diinginkan. 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 dinamika politik yang bersumber dari data media sosial yang berukuran besar.

## 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 opini publik, seperti manajemen, politik, sosial dan ekonomi. 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 mengukur tingkat elektabilitas calon presiden 2024 dengan mengidentifikasi sentimen positif atau negatif pada opini masyarakat di media sosial Twitter yang ditujukan terhadap para kandidat yang berkontestasi.

## 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. Sub proses 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 sub proses pada *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 memiliki signifikansi dalam memberikan arti pada teks yang dianalisis dengan tujuan penggalian wawasan dan klasifikasi dokumen (Perkins, 2014). Sehingga "Penghapusan *stopword*" merupakan proses menghapus kata dalam data teks yang tidak memiliki signifikansi dalam proses penggalian wawasan dan klasifikasi dokumen. Secara langsung penggapusan *stopword* akan berdampak pada berkurangnya kebutuhan kapasitas tempat penyimpanan. 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[[14]](#footnote-14),[[15]](#footnote-15).

2.5.2.3 Stemming

*Stemming*, secara harfiah, adalah proses memotong cabang-cabang pohon hingga hanya tersisa batangnya. 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 identifikasi jenis kata (*part-of-speech*) dan *Named Entity Recognition* (NER) (Hardeniya et al., 2016). Hal ini dikarenakan *stemming* akan mengubah token dan ini dapat memberikan hasil identifikasi NER dan jenis kata yang berbeda.

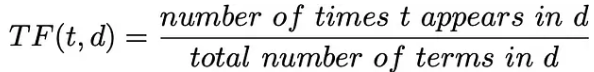
2.5.2.4 Lematisasi

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 kuat dan sistematis untuk mengubah variasi kata ke akar kata-nya. 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 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.

Table x menggambarkan klasifikasi *machine learning* berdasarkan pejelasan diatas.

|  |  |  |
| --- | --- | --- |
| Tipe variable dependen / Kesediaan label pada variable dependen | Dataset memiliki label / Supervised Machine Learning | Dataset tidak memilki label / Unsupervised machine learning |
| Kontinue | Regression | Dimentionality Reduction |
| Diskrit | Classification | 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 hanya menggunakan *supervised machine learning* hanya digunakan untuk tugas *classification,* yaitu untuk memprediksi sentiment dari *tweet* masyarakat Indonesia di *twitter* terhadap calon presiden 2024 yang berkontestasi. Berikut macam-macam algoritma *supervised machine learning* untuk tugas *classification* yang dapat digunakan:

2.5.3.2.1 Decission 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 X memberikan contoh *decision tree* yang terbentuk dari data latih yang tunjukkan pada tabel Z. *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).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

*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 kontinu 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 di pada garis *margin* yang disebut *support vector*, sedangkan *margin* (yang ditunjukkan sebagai garis putus-putus pada Gambar X 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 (memiliki *margin* terjauh ke *support vector*), bersama dengan optimalisasi *parameter* yang digunakan sebagai bobot perhitungan (Kuhn & Johnson, 2013). Gambar X 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 X sebelah kanan.

A comparison of colored lines and a graph

Description automatically generated with medium confidence

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

.

*Linear hyperplane* berupa titik untuk ruang 1 dimensi, garis untuk ruang 2 dimensi, *plane* untuk ruang 3 dimensi, dan *hyperplane* untuk ruang lebih dari 3 dimensi. Jumlah dimensi ditentukan berdasarkan fitur yang dimiliki oleh *instance*. Model pembelajaran SVM dapat dinotasikan dengan formula sebagai berikut (Kelleher et al., 2020):

A black text on a white background

Description automatically generated

Dimana M**a,***w*0 adalah model *predictive analytics* SVM dalam mengklasifikasikan **q** *query instance* adalah hasil perjumlahan *dot product* antara *ti* nilai batas atribut kelas, *a*[*i*] *parameter* bobot untuk masing-masing *support vector* dan *dot product* masing-masing fitur dari *support vector* **d**i dan *query instance* yang diiterasi sejumlah *s* atau jumlah *support vector* yang ada. Sebagai *quadratic optimization problem*, proses pembelajaran pada SVM dapat didefinisikan ke dalam 2 hal, yaitu: (1) Sekumpulan batasan dan (2) Optimalisasi kriteria.

* + 1. Sekumpulan batasan ditentukan berdasarkan nilai atribut kelas. sebagai contoh terdapat 2 atribut kelas dalam kasus sentiment analysis memiliki batasan nilai 1 atau lebih besar untuk sentimen positif dan nilai -1 atau lebih kecil untuk sentimen negatif. Dapat dinotasikan sebagai berikut:

A black and white rectangle with black text

Description automatically generated

A black and white rectangular sign with black text

Description automatically generated

Dimana *w0* dan vektor ***w*** adalah bobot terhadap vektor fitur ***d*** dan hasilnya akan berkorelasi dengan *ti* sebagai atribut kelas. pada tahap ini SVM akan menghasilkan banyak kemungkinan *hyperplane* seperti yang ditunjukkan pada Gambar X bagian kiri.

* + 1. Optimalisasi kriteria dilakukan untuk menemukan *parameter* terbaik ***w*** yang pada akhirnya dapat menghasilkan *hyperplane* paling optimal seperti yang ditunjukkan pada Gambar X bagian kanan. Optimasisasi kriteria didefinisikan dengan *perpendicular distance* dari setiap *instance* yang berada di dalam margin.

A black text on a white background

Description automatically generated



*Partial derivative* digunakan untuk mengoptimalisasikan nilai **w** yang direpresentasikan dengan *cost function***,** ini mengapa kemudiaan SVM juga dapat dianggap sebagai teknik *error-based machine learning* (Kelleher et al., 2020)*.*

Ketika *hyperplane* tidak ditemukan pada ruang dimensi asli, dimana sejumlah fitur pada *instance,* maka sebagai solusi pendekatan dengan menggunakan *bias function* dapat dilakukan(Kelleher et al., 2020)*.* *Bias function* mengubah dan meningkatkan dimensi data dengan tujuan dapat menemukan *hyperplane*. Selain itu teknik *kernel function* seperti *Linear kernel, Polynomial kernel,* dan *Gaussian radial basis kernel* juga dapat digunakan untuk mengoptimasilasi proses *dot product* fitur *support vector* dengan *query instance* yang membutuhkan beban komputasi yang tinggi (Kelleher et al., 2020).

2.5.3.2.3 Naïve Bayes

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

A mathematical equation with black text

Description automatically generatedA mathematical equation with black text

Description automatically generated

Dimana *Y* adalah nilai yang mungkin dimiliki oleh atribut kelas atau label, dimana masing-masing dapat dinotasikan dengan *yi*. variable *X* adalah fitur dari *instance,* sedangkan fungsi *P()* adalah probabilitas (Zafarani et al., 2014). Sehingga (a) *P(yi|X)* adalah probabilitas terjadinya label *Y* dengan fitur-fitur *X,* (b) *P(yi)* adalah probabilitas munculnya label *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 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 x 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 X. 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[[17]](#footnote-17)

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 yang memungkinkan untuk mempelajari data dengan pola yang tidak linear dan sangat cocok digunakan pada bidang dengan data input yang besar seperti pemprosesan gambar, suara dan text. 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 X 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* (Kelleher et al., 2020).

A diagram of a network

Description automatically generated

Gambar X Model *deep learning*

Proses *backpropagation* diawali dengan menentukan nilai inisial untuk setiap *weight.* Selanjutnya *backpropagation* menjalankan 2 tahapan yang disebut: *forward pass* dan *backword pass*. Berikut penjelasan langkah *forward pass*:

* 1. Proses dimulai ketika *sensing neuron* pada *input layer* memberikan nilainya untuk diproses pada masing-masing *neuron* di layer selanjutnya.
  2. Perjumlahan setiap nilai keluaran *neuron* dikalikan dengan bobotnya dilakukan, perjumlahan tersebut juga menyertakan nilai *bias parameter* yang dinotasikan sebagai w0.
  3. Hasil perjumlahaan dinormalisasi dengan *activation function*, seperti *logistic function* atau *ReLU function*. Gambar X menunjukan *domain* dan *range* dari *activation function* yang sering digunakan.

A graph of a function

Description automatically generated

Gambar X Activation function, kiri-ke-kanan (1) fungsi berdasarkan threshold, (2) logistic and tanh function, dan (3) ReLU function.

* 1. Hasil *activation function*, yang merupakan output dari satu *neuron*, akan diberikan sebagai input kepada *neuron* di layer selanjutnya. (e) Proses pada butir c diulang untuk *neuron* pada lapisan selanjutnya, satu proses *pass forward* dikatakan selesai ketika komputasi telah menyentuh *neuron* pada *output layer*. Hasil prediksi ditentukan oleh *neuron* (merepresentasikan nilai atribut kelas) yang mendapatkan sinyal paling tinggi.

Berikit adalah penjelasan dari langkah *backward pass*:

* 1. Proses diawali dengan mengurangi *output activation* masing-masing *neuron* pada *output layer* terhadap *error cost*. Hal ini dilakukan untuk mengetahui tingkat perubahan *cost error* jika terjadi perubahan terhadap nilai *activation* dari *neuron* pada *output layer,* dinotasikan dengan δℰ/δ*a*.
  2. Masing-masing δℰ/δ*a* dari semua *output neuron* kemudian disebar ke belakang atau ke arah *neuron* pada *hidden layer* yang terkoneksi dengan masing masing *output neuron* tersebut.
  3. Nilai δℰ/δ*a* yang disebar kebelakang ini disesuaikan dengan porsi kontribusi *weight* pada *link* yang menghubungkan kedua *neuron* tersebut yang digunakan ketika tahapan *forward pass*. Penyesuaian ini menghasilkan tingkat perubahan *activation neuron* jika terjadi perubahan terhadap nilai *weight* dari *link* yang terkoneksi dengannya, ini mengapa sebabnya nilai *activation* dari setiap *neuron* akan dicatat selama tahapan *forward pass* berlangsung. Hal ini dilakukan terhadap setiap *neuron* yang terhubung dengan *output neuron*.
  4. Langkah selanjutnya adalah menemukan tingkat perubahan *activation neuron* jika terjadi perubahan terhadap nilai *z* pada *neuron* (Nilai *z* adalah perjumlahan dari pembobotan nilai *activation* yang berasal dari layer sebelumnya): ∂*ak* /∂*zk* atau ∂*logistic(z)* /∂*z*, dengan menggunakan *derivative* dari *activation function*. Gambar X menunjukkan *domain* dan *range* dari *derivative activation function*:

|  |  |
| --- | --- |
| A graph of a function  Description automatically generated |  |

Gambar X domain dan range dari derivative activation function.

langkah dari butir a ke d dapat dinotasikan sebagai berikut:

A math equations and formulas

Description automatically generated with medium confidence

* 1. Ketika δs untuk semua neurons pada hidden layer terakhir (yang terkoneksi dengan output layer) telah teridentifikasi, maka proses akan diulangi untuk menghitung δs neuron yang berada dari hidden layer sebelumnya sampai dengan input layer.
  2. Pada akhirnya ketika δ dari semua neuron telah teridentifikasi, penyesuaian nilai bobot weights yang ada pada jaringan saraf tiruan siap untuk dilakukan dengan menggunakan algoritma *gradient descent*.

#### 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 X menggambarkan contoh 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 X menjelaskan lebih detail tentang ketiga metrik tersebut.

|  |  |
| --- | --- |
| 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 pemodelan 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.

#### 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 merupakan hasil dari pemodelan topik adalah sekumpulan kata yang secara statistik kemunculannya signifikan dalam korpus (Jelodar et al., 2019). Daftar topik yang dihasilkan, bukan berasal dari analisa makna ataupun konsep linguistik dari teks pada korpus.

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 ingin 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 Blablabla 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 Blablabla 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 Blablabla 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 Blablablabla 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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Judul Penelitian | Compare | Contrast | Critisize | 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.* |
| ***Synthesize:***  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 X. Komponen pada kerangka penelitian yang disusun berdasarkan tahapan *predictive analytics* pada kerangka kerja *Text Mining*.

A group of white rectangular objects

Description automatically generated

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 pemodelan topik kasus XYZ” 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 memiliki sifat tidak terstruktur menjadi lebih terstruktur 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 kasus XYZ. 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,* dan *decisicion tree*. 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 *accuracy, precision, recall* dan *f1-score*.

# 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 kasus XYZ dengan mengukur sentimen masyarakat 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.

|  |  |  |
| --- | --- | --- |
| No | Atribut Penelitian | Deskripsi |
| 1 | Jenis Penelitian | Kuantitatif |
| 2 | Klasifikasi | Eksperimental |
| 3 | Tujuan Penelitian | Membentuk model untuk melakukan analisis sentimen dan menggali *concern* masyarakat terkait dengan kasus XYZ |
| 4 | Hasil Penelitian | Rekomendasi bagi lembaga ABC 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 kasus XYZ 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*, *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, dengan tujuan untuk dapat mengidentifikasi *concern* masyarakat baik yang berasal dari sentimen positif maupun yang negative, maka pemodelan topik juga dilakukan dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Table X 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 X.

|  |  |
| --- | --- |
| 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 kasus XYZ berdasarkan dokumen dan berita yang berkaitan dengannya. Observasi dan studi literatur yang dilakuakan terhadap dokumen terkait menghasilkan kesimpulan bahwa. Lembaga ABC diharapkan dapat dengan cepat menangkap aspirasi masyarakat terutama terkait dengan kasus ABC. Namun yang saat ini terjadi lembaga ABC dinilai kurang dapat menerima aspirasi yang disampaikan masyarakat, dari survey yang dilakukan oleh lembaga independen menunjukkan angka kepuasan masyarakat terkait penangganan kasus XYZ hanya 17%. Selain ini dari penelusuran yang dilakukan pada tahapan ini, penulis juga beranggapan bahwa Lembaga ABC dapat memanfaatkan opini masyarakat terkait kasus XYZ yang mereka sampaikan melalui media sosial *twitter.*

### Perumusan Masalah

Pada tahapan ini peneliti berupaya untuk dapat mengidentifikasi permasalahan umum yang muncul terkait dengan penangganan kasus XYZ. Maka dilakukan analisa kesenjangan berdasarkan harapan dan kenyataan yang diperoleh pada tahapan sebelumnya. Sehingga dapat disimpulkan bahwa permasalahan utama pada kasus XYZ adalah [TBA].

### Penentuan Pertanyaan Penelitian

Pada tahapan ini peneliti berupaya untuk dapat mengidentifikasi cakupan, tujuan, hasil dan pertanyaan penelitian. Berdasarkan permasalahan umum yang ada, peneliti membaginya kedalam 10 faktor yang berpotensi menjadi akar masalah yang dikelompokkan ke dalam 3 domain berdasarkan teori Blablabla. Lalu dilakukan analisa akar masalah dengan menggunakan *fishbone analysis diagram.* Hasilnya disimpulkan bawah kurangnya kemampunan Lembaga ABC dalam menangkap aspirasi masyarakat terkait dengan kasus XYZ sebagai akar masalahnya. 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 kasus XYZ dan analisis sentimen. Peneliti juga melakukan analisa 3C+2S (*criticize, compare, contrast, summarize, synthesize*) dari 5 penelitian terakhir yang memiliki kesamaan dengan topik penelitian ini baik dari sisi metode yang digunakan maupun tujuan penelitian yang disasar.

### Pengumpulan Data

Pada tahapan ini peneliti berupaya mengumpulkan data opini masyarakat terkait dengan kasus XYZ yang mereka sampaikan melalui media sosial twitter. Dengan menggunakan pendekatan *web crawling* melalui Twitter API, proses penarikan data dilakukan pada periode Desember 2023 sampai dengan Januari 2024. Selain itu digunakan *keyword* ‘Kasus XYZ’, ‘XYZ Kasus’, ‘Lembaga ABC Kasus XYZ’, ‘Kasus XYZ lembaga ABC’ untuk dapat memperoleh data yang sesuai harapan yaitu *tweet* terkait dengan kasus XYZ.

### 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 X

|  |  |
| --- | --- |
| **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.

### 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 [TBA]

# 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

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

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France. *New Media and Society*, *16*(2), 340–358. https://doi.org/10.1177/1461444813480466

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

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

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5 SPEC. ISSUE), 861–871. https://doi.org/10.1093/poq/nfr057

Gunawan, W., Lestari, K., Purnomo, M., Silvya Sari Indonesian, D., & Silvya Sari, D. (2019). *Democracy without Conflict?*

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

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

Ottati, V., Wilson, C., & Lambert, A. (2016). Accessibility, priming, and political judgment. *Current Opinion in Psychology*, *12*, 1–5. https://doi.org/10.1016/j.copsyc.2016.04.010

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

Reveilhac, M., Steinmetz, S., & Morselli, D. (2022). A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia Tools and Applications*, *81*(7), 10107–10142. https://doi.org/10.1007/s11042-022-12101-0

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

Valentino, N. A., King, J. L., & Hill, W. W. (2017). Polling and Prediction in the 2016 Presidential Election. *Computer*, *50*(5), 110–115. https://doi.org/10.1109/MC.2017.139

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. https://www.kompas.id/baca/polhuk/2022/04/08/87-persen-milenial-dan-generasi-z-antusias-ikuti-pemilu-2024 [↑](#footnote-ref-1)
2. https://www.bbc.com/indonesia/indonesia-66531834 [↑](#footnote-ref-2)
3. https://www.cnbcindonesia.com/news/20230913051546-4-471930/32-hasil-survei-capres-2024-prabowo-vs-anies-vs-ganjar [↑](#footnote-ref-3)
4. https://www.youtube.com/watch?v=JlqdcHMY22g [↑](#footnote-ref-4)
5. https://apjii.or.id/berita/d/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang [↑](#footnote-ref-5)
6. https://dataindonesia.id/internet/detail/pengguna-media-sosial-di-indonesia-sebanyak-167-juta-pada-2023 [↑](#footnote-ref-6)
7. https://data.goodstats.id/statistic/agneszefanyayonatan/menilik-pengguna-media-sosial-indonesia-2017-2026-xUAlp [↑](#footnote-ref-7)
8. https://www.kompas.id/baca/riset/2023/02/20/ruang-media-sosial-terbuka-luas-siapa-partai-paling-diuntungkan [↑](#footnote-ref-8)
9. https://datareportal.com/reports/digital-2023-indonesia [↑](#footnote-ref-9)
10. https://databoks.katadata.co.id/datapublish/2022/03/28/survei-reuters-twitter-banyak-digunakan-untuk-mencari-berita [↑](#footnote-ref-10)
11. https://blog.twitter.com/en\_us/topics/product/2017/Giving-you-more-characters-to-express-yourself [↑](#footnote-ref-11)
12. https://www.kompasiana.com/virdaazzahra9944/63b592e94addee074147c912/media-sosial-sebagai-alat-sarana-komunikasi-politik [↑](#footnote-ref-12)
13. https://www.kompas.id/baca/polhuk/2019/04/14/jokowi-prabowo-dan-beda-strategi-cuitan-twitter/ [↑](#footnote-ref-13)
14. https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist [↑](#footnote-ref-14)
15. https://rahmadya.com/2019/04/24/stopword-berbahasa-indonesia/ [↑](#footnote-ref-15)
16. https://en.wikipedia.org/wiki/Conditional\_probability [↑](#footnote-ref-16)
17. https://en.wikipedia.org/wiki/Lazy\_learning [↑](#footnote-ref-17)