BAB I
PENDAHULUAN

1.1 Latar Belakang

Perubahan cara masyarakat berkomunikasi akibat kemajuan teknologi digital telah menjadikan informasi menyebar dengan sangat cepat, termasuk informasi mengenai kesehatan. Ketika wabah penyakit muncul, opini publik dapat terbentuk dalam hitungan jam melalui media sosial, yang sering kali memunculkan reaksi beragam seperti dukungan terhadap kebijakan pemerintah, kekhawatiran masyarakat, hingga kritik terhadap cara penanganan wabah. Dalam konteks inilah, memahami persepsi publik menjadi kebutuhan penting agar kebijakan kesehatan yang dibuat tidak hanya efektif secara medis, tetapi juga responsif terhadap dinamika sosial yang berkembang di masyarakat. Mekanisme komunikasi digital seperti memposting, membagikan, mengomentari, dan mengunggah ulang informasi ini menghasilkan data teks yang dapat dianalisis untuk mengetahui sentimen publik. Aktivitas-aktivitas inilah yang menghasilkan data teks yang kemudian dapat dianalisis untuk mengidentifikasi sentimen publik mengenai isu tertentu.

Di Kabupaten Sumenep, pada tahun 2025, wabah campak ditetapkan sebagai Kejadian Luar Biasa (KLB). Berdasarkan laporan dari Majelis Ulama Indonesia Jawa Timur (MUI Jatim), pemerintah daerah mencatat lebih dari 2.000 kasus suspek campak dengan 17 anak dilaporkan meninggal dunia(savhira20, 2025) Pemerintah Kabupaten Sumenep kemudian melaksanakan vaksinasi massal melalui program *Outbreak Response Immunization* (ORI) pada anak usia 9 bulan hingga 6 tahun, yang dimulai pada tanggal 25 Agustus, berlangsung hingga pertengahan September 2025 (Rahman Taufiqur, 2025) . Dalam jurnal "*analisis faktor risiko yang mempengaruhi kejadian campak di indonesia: literatur review*" menyatakan bahwa terdapat recana strategi global maupun regional tahun 2021-2030 yang direcanakan WHO bersama *Measles* dan *Rubella Initiative*, UNICEF, *American Red Cross, CDC*, dan *United Nations Foundation* pada tahun 2020 menyatakan bahwa pengendalian penyakit campak secara global ditujukan untuk mewujudkan dunia yang terbebas dari campak dan rubella melalui berbagai upaya pencegahan dan pengendalian penyakit.(Riantina & Sitorus, 2024)

Campak menyebar dengan sangat mudah melalui percikan cairan (seperti air liur) dari orang yang terinfeksi. Masa inkubasinya berkisar antara minimal tujuh jam hingga maksimal delapan belas jam. Gejala penyakit ini diawali dengan fase prodromal, yang ditandai dengan suhu tubuh 38°C atau lebih tinggi selama tiga hari. Gejala lainnya meliputi batuk, pilek, dan *konjungtivitis*.(Daniyah Zahrah et al., 2023). Sehingga wabah ini menimbulkan kekhawatiran masyarakat, tidak hanya karena jumlah kasus yang tinggi, tetapi juga karena adanya perdebatan publik mengenai efektivitas vaksin dan penanganan dari dinas kesehatan.

Media sosial, terutama platform X, menjadi wadah utama bagi masyarakat untuk menyuarakan pendapat mereka terkait situasi ini. Melalui unggahan, komentar, dan percakapan daring, masyarakat mengekspresikan dukungan terhadap vaksinasi, membagikan pengalaman pribadi, atau justru menyampaikan kekhawatiran dan kritik terhadap pelaksanaan program pemerintah. Dalam konteks komunikasi publik, fenomena ini menunjukkan bahwa sentimen masyarakat terhadap isu kesehatan sangat dipengaruhi oleh arus informasi di media sosial. Karena itu, analisis sentimen terhadap unggahan masyarakat dapat memberikan gambaran mengenai bagaimana persepsi publik terbentuk dan berubah selama masa wabah.

Sederhananya, analisis sentimen adalah proses komputasi yang bertujuan untuk mengidentifikasi opini atau emosi yang diungkapkan dalam bentuk teks dan mengelompokkannya ke dalam kategori seperti *positif* dan *negatif*. Analisis sentimen menjadi penting untuk mengevaluasi efektivitas kebijakan dan isu kesehatan. Teknik komputasi berguna untuk mengidentifikasi dan mengklasifikasikan opini yang diungkapkan dalam teks guna menentukan perspektif penulis mengenai topik yang sedang dibahas. Teknik tersebut dapat dimanfaatkan untuk memahami persepsi publik terhadap perlindungan data nasional, mengukur tingkat kepercayaan terhadap kebijakan pemerintah, serta mengidentifikasi potensi risiko dan kerentanan (N. Maulida, 2024). Meskipun sejumlah penelitian sebelumnya telah menggunakan analisis sentimen untuk menganalisis opini publik terkait wabah campak. Pendekatan berbasis leksikon dan machine learning telah digunakan, tetapi akurasi dan efektivitasnya masih perlu ditingkatkan.

Beberapa penelitian di Indonesia telah menerapkan analisis ini untuk berbagai isu, termasuk kesehatan, dengan menggunakan teknik representasi teks seperti *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW). Misalnya, penelitian oleh Alfiansyah dkk pada tahun 2023 dalam Jurnal PROKASDADIK menunjukkan bahwa pembobotan TF-IDF dengan algoritma *Naïve Bayes* efektif dalam mengklasifikasikan sentimen pengguna *Twitter* terhadap isu publik di Indonesia (Siti Sholihah, 2024) .

Namun demikian, sebagian besar penelitian analisis sentimen di bidang kesehatan di Indonesia masih berfokus pada isu-isu besar seperti COVID-19 atau vaksinasi secara umum. Studi yang secara khusus meneliti penyakit seperti campak, terlebih dalam konteks lokal seperti Kabupaten Sumenep, masih sangat terbatas. Selain itu, penelitian yang membandingkan dua pendekatan representasi teks *Bag of Words* dan TF-IDF dalam satu konteks data sosial masyarakat juga jarang dilakukan. Sebagai contoh, penelitian oleh Ramadhan dkk. (2022) Pada Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer menegaskan bahwa metode TF-IDF memiliki keunggulan dalam membedakan tingkat kepentingan antar-kata dibandingkan dengan Bag of Words, sehingga meningkatkan akurasi dalam klasifikasi teks (R. Ramadhan et al., 2021) .

Dalam penelitian ini, digunakan dua metode representasi teks yang umum dipakai dalam analisis sentimen, yaitu TF-IDF (*Term Frequency–Inverse Document Frequency*) dan *Bag of Words* (BoW)**.** TF-IDF bekerja dengan menimbang tingkat kepentingan suatu kata dalam dokumen berdasarkan frekuensi kemunculannya, sementara BoW merepresentasikan teks berdasarkan jumlah kemunculan kata tanpa mempertimbangkan bobot atau konteks. Dimana data diambil dari media sosial X dengan menggunakan kata kunci seperti “campak Sumenep”, “KLB campak”, dan “campak”. Dua model representasi teks, yaitu *Bag of Words* dan TF-IDF, akan dibandingkan pada kemampuan mereka dalam mempengaruhi hasil klasifikasi sentimen menggunakan algoritma *Logistic Regression*.

Logistic Regression merupakan algoritma klasifikasi statistik yang sederhana namun sangat efektif dalam memodelkan hubungan antara fitur dan label kelas biner, seperti sentimen positif dan negatif. Kelebihan utama Logistic Regression

adalah kemampuannya untuk memberikan interpretasi probabilistik, efisiensi komputasi, dan kinerja yang kompetitif dalam berbagai tugas klasifikasi teks. Untuk klasifikasi sentimen, algoritma *Logistic Regression* dipilih karena kemampuannya dalam menangani data teks dan memberikan *interpretasi* probabilistik yang jelas (Dedy, 2024)

Hipotesis dalam penelitian ini menyatakan bahwa representasi teks menggunakan TF-IDF akan menghasilkan performa klasifikasi yang lebih baik dibandingkan Bag of Words. Hal ini disebabkan oleh kemampuan TF-IDF dalam menurunkan bobot kata yang sering muncul tetapi kurang relevan serta memperkuat kata yang lebih spesifik terhadap konteks pembicaraan publik mengenai wabah campak. Selain itu, penerapan teknik penyeimbangan kelas diharapkan mampu meningkatkan *recall* dan *F1-score* pada kelas minoritas, sehingga model yang dihasilkan tidak hanya akurat tetapi juga adil dalam mendeteksi berbagai jenis sentimen.

Oleh karena itu, penelitian ini menyajikan kontribusi teoritis melalui pengayaan literatur analisis sentimen kesehatan pada konteks penyakit menular lokal seperti campak, sekaligus kontribusi praktis dalam memetakan persepsi publik di media sosial. Hasil analisis ini menjadi acuan berbasis data yang dapat diimplementasikan untuk mengoptimalkan edukasi kesehatan masyarakat dan mendukung efektivitas penanggulangan wabah secara menyeluruh.

1.2 Perumusan Masalah

Pada bagian perumusan masalah, diuraikan mengenai permasalahan, metode usulan serta pertanyaan penelitian yang diuraikan sebagai berikut:

1.2.1 Permasalahan

Permasalahan utama pada penelitian ini adalah bagaimana memahami dan menganalisis lebih lanjut mengenai opini dan sentimen publik berkaitan dengan wabah penyakit campak. Di Indonesia, penyakit campak masih satu dari sekian banyak masalah kesehatan yang mendapatkan perhatian karena tingkat penularannya yang relatif tinggi serta dampak yang ditimbulkannya bagi masyarakat.

Namun, persepsi publik terhadap wabah ini beragam, mulai dari kekhawatiran, ketidakpercayaan, hingga dukungan terhadap langkah pencegahan

yang dilakukan pemerintah. Informasi ini biasanya beredar di Twitter dan beberapa media sosial lainnya meski analisis terhadap informasi ini masih sangat terbatas. Maka dari itu, untuk memahami persepsi publik secara menyeluruh, diperlukan pendekatan yang lebih fleksibel dan akurat guna mendapatkan hasil yang lebih mendalam dan representatif.

1.2.2 Metode Usulan

Metode yang diusulkan dalam penelitian ini yaitu menggunakan dua metode representasi fitur teks, yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW), untuk mengubah data teks dari unggahan pengguna media sosial X atau *Twitter* menjadi bentuk numerik yang dapat diproses oleh algoritma *klasifikasi*. Kedua metode tersebut dibandingkan untuk mengetahui perbedaan pengaruh representasi fitur terhadap kinerja klasifikasi sentimen publik terkait wabah penyakit campak. Selanjutnya, algoritma *Logistic Regression* (LR) digunakan sebagai metode klasifikasi dalam pengelompokkan sentimen ke dalam dua kelas, yaitu positif dan negatif. *Logistic Regression* dipilih karena mampu menangani klasifikasi, memiliki proses komputasi yang relatif efisien, serta sesuai untuk pengolahan data teks dengan jumlah fitur yang tinggi. Kinerja model kemudian dievaluasi menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk membandingkan hasil penggunaan TF-IDF dan BoW.

1.2.3 Pertanyaan Penelitian

Berdasarkan latar belakang, dapat diambil pertanyaan penelitian sebagai berikut:

1. Bagaimana pengaruh perbandingan TF-IDF dan *Bag of Words* (BoW) terhadap kinerja algoritma *Logistic Regression* dalam mengklasifikasikan sentimen publik terhadap wabah penyakit campak pada media sosial Twitter?

1.3 Tujuan dan Manfaat

Pada bagian tujuan dan manfaat, diuraikan sasaran penelitian serta kontribusi yang diharapkan sebagai berikut:

1.3.1 Tujuan Penelitian

Berdasarkan rumusan masalah yang ada, dapat ditentukan tujuan dari penelitian ini diantaranya adalah sebagai berikut:

1. Menganalisis perbedaan penerapan dua metode representasi teks yaitu *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW) terhadap kinerja algoritma Logistic Regression dalam klasifikasi sentimen publik terhadap wabah penyakit campak pada media sosial *Twitter*.

1.3.2 Manfaat Penelitian

Berdasarkan rumusan masalah yang ada, dapat ditentukan manfaat dari penelitian ini diantaranya adalah sebagai berikut:

1. Manfaat Akademis:

   Penelitian dapat berkontribusi pada pengembangan ilmu di bidang *text mining* dan analisis kalimat, khususnya dalam analisis dan perbandingan dua metode representasi teks yaitu *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW) guna meningkatkan efisiensi algoritma Regresi Logistik. Selain itu, hasil penelitian ini dapat menjadi acuan bagi peneliti selanjutnya yang berkaitan dengan pemodelan opini publik mengenai isu kesehatan menggunakan pembelajaran mesin *machine learning*
2. Manfaat Praktis:

   Penelitian memberikan gambaran mengenai persepsi dan reaksi publik terhadap wabah penyakit campak berdasarkan data yang diperoleh dari media sosial *Twitter*. Hasil analisis sentimen ini diharapkan dapat menjadi bahan pertimbangan bagi instansi kesehatan, pemerintah, maupun lembaga komunikasi publik dalam merancang strategi penyebaran informasi, edukasi masyarakat, serta penanganan isu kesehatan secara lebih efektif sesuai dengan kecenderungan opini masyarakat.

1.4 Batasan Masalah

Berikut batasan-batasan yang akan diberikan agar permasalahan tidak terlalu luas dan tidak keluar dari topik sehingga sesuai dengan tujuan penulis berikut ini:

1. Penelitian ini fokus pada analisis sentimen publik terkait wabah penyakit campak.
2. Sumber data terbatas pada *platform* media sosial *Twitter,* yang dipilih karena banyak digunakan masyarakat untuk menyuarakan opini secara

terbuka.

3. Metode representasi teks yang digunakan pada penelitian adalah *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW), sedangkan algoritma klasifikasi sentimen yang digunakan adalah *Logistic Regression*.

1.5 Sistematika Penulisan

Sistematika penulisan tugas akhir ini disusun dengan mengikuti aturan penulisan proposal pada program studi S1 Sistem Informasi, Universitas Trunojoyo Madura. Berikut merupakan susunan tugas akhir pada penelitian ini:

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang yang mendasari penelitian, diikuti dengan perumusan masalah yang ingin diselesaikan. Selain itu, tujuan dan manfaat penelitian diuraikan untuk menunjukkan kontribusi penelitian baik dari segi akademis maupun praktis. Batasan masalah disertakan untuk memperjelas ruang lingkup penelitian, dan diakhiri dengan metodologi penelitian serta sistematika penulisan sebagai panduan pembaca untuk memahami alur penyusunan penelitian ini.

BAB II TINJAUAN PUSTAKA

Tinjauan pustaka memuat landasan teoritis, konsep-konsep penting, serta hasil penelitian terdahulu yang mendukung dan berkaitan dengan topik yang dikaji. Bagian ini berfungsi sebagai acuan untuk memperkuat dasar metodologis dan arah penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi yang digunakan dalam penelitian, mulai dari pengumpulan data di media sosial *Twitter*, *preprocessing* teks (pembersihan, *tokenisasi, stopword removal, stemming*), hingga representasi teks menggunakan TF-IDF dan *Bag of Words* (BoW). Proses klasifikasi sentimen dilakukan menggunakan algoritma *Logistic Regression*. Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, *precision, recall*, dan F1-*score*.

BAB IV HASIL & PEMBAHASAN

Bab ini menampilkan hasil eksperimen dan analisis mendalam mengenai perbandingan kinerja ekstraksi fitur TF-IDF dan *Bag of Words* (BoW) pada

klasifikasi sentimen masyarakat terkait wabah campak menggunakan algoritma *Logistic Regression.* Pembahasan mencakup visualisasi distribusi sentimen data, evaluasi performa kedua metode ekstraksi fitur berdasarkan metrik akurasi, *precision*, *recall*, dan *F1-score*, serta analisis *confusion matrix*.

BAB V PENUTUP

Bab V berisi kesimpulan dari seluruh rangkaian penelitian serta saran untuk pengembangan penelitian selanjutnya. Kesimpulan merangkum temuan utama mengenai efektivitas dan perbandingan akurasi antara metode ekstraksi fitur TF-IDF dan *Bag of Words* (BoW) ketika dipadukan dengan *Logistic Regression* dalam menganalisis sentimen wabah campak di media sosial *Twitter*. Sementara itu, saran menyajikan rekomendasi teknis dan akademis, seperti pengembangan dataset, penggunaan algoritma klasifikasi alternatif, serta penerapan teknik *feature engineering* yang lebih kompleks untuk penelitian mendatang.