BAB I
PENDAHULUAN

1.1 Latar Belakang

Program Magang Generasi Bertalenta Badan Usaha Milik Negara (Magenta BUMN) merupakan salah satu upaya strategis pemerintah dalam menyiapkan talenta muda yang kompeten dan siap bersaing di dunia kerja. Melalui program ini, mahasiswa dan lulusan perguruan tinggi memiliki kesempatan untuk mendapatkan pengalaman kerja langsung di lingkungan BUMN sekaligus berkontribusi dalam pembangunan nasional. Sejak diluncurkan, Magenta BUMN mulai dikenal luas dan menjadi topik yang banyak dibicarakan, termasuk di media sosial. Berdasarkan data resmi dari situs Magenta BUMN, hingga tahun 2025 tercatat 1.097.749 talenta telah terdaftar dalam program ini, dengan 406.679 talenta aktif melamar, 14.766 talenta telah lulus magang, dan 9.455 talenta sedang menjalani program magang. Angka ini menunjukkan besarnya partisipasi dan ketertarikan generasi muda terhadap Magenta BUMN. dengan cakupan peserta yang luas, opini publik yang beredar di media sosial menjadi faktor penting yang dapat memengaruhi citra dan keberlanjutan program.

Media sosial X (sebelumnya dikenal sebagai Twitter) menjadi salah satu ruang publik yang dinamis, tempat pengguna dapat dengan bebas menyampaikan pendapat, berbagi pengalaman, hingga memberikan kritik. Jumlah unggahan yang besar dan beragam ini menyimpan potensi informasi berharga untuk memahami pandangan publik terhadap Magenta BUMN. Namun, sifat data yang tidak terstruktur dan jumlahnya yang masif membuat analisis manual menjadi sulit dilakukan secara efektif.

Analisis sentimen menjadi salah satu solusi untuk mengolah opini publik menjadi informasi yang terukur. dengan memanfaatkan teknik pemrosesan bahasa alami (*Natural Language Processing*/NLP), teks dari media sosial dapat dianalisis untuk mengidentifikasi sentimen positif atau negatif. Data yang awalnya berupa kumpulan opini acak dapat diubah menjadi gambaran yang jelas, sehingga mendukung proses evaluasi dan pengambilan keputusan yang berbasis data.

Dalam proses analisis sentimen, diperlukan dua komponen utama yaitu representasi fitur teks dan algoritma klasifikasi. Pemilihan metode representasi fitur

sangat krusial karena menentukan seberapa baik makna semantik dan kontekstual dari teks dapat ditangkap oleh model. Metode tradisional seperti *Bag of Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) memiliki keterbatasan karena hanya menghitung frekuensi kemunculan kata tanpa mempertimbangkan hubungan semantik dan konteks antar kata (Anjani & Irmanda, 2024). Penelitian oleh Zhan (2025) menunjukkan bahwa TF-IDF mengalami *overfitting* parah dengan akurasi *Training* 99,16% namun *testing* hanya 73,9%, sementara Word2Vec menghasilkan performa yang lebih stabil dengan akurasi *Training* 68,4% dan *testing* 68,65%, menunjukkan kemampuan generalisasi yang jauh lebih baik.

Word2Vec hadir sebagai alternatif yang lebih unggul karena mampu memetakan kata ke dalam ruang vektor berdimensi tinggi sehingga hubungan semantik dan kontekstual antar kata dapat terdeteksi secara lebih baik. Penelitian komparatif oleh Anjani & Irmanda (2024) membandingkan tiga metode *Word Embedding* (Word2Vec, *GloVe*, dan *FastText*) untuk analisis sentimen ulasan Spotify dan menemukan bahwa ketiga metode ini jauh lebih unggul dibanding *feature engineering* tradisional karena kemampuannya menangkap makna semantik, sintaksis, dan kontekstual. Meskipun *GloVe* menghasilkan akurasi tertinggi (85%), Word2Vec tetap terbukti efektif dan konsisten dalam berbagai domain aplikasi. Sarwar et al., (2022) juga memvalidasi keunggulan Word2Vec dalam mengukur *semantic relatedness* untuk mencari artikel berita serupa, dengan kombinasi Word2Vec dan *Cosine similarity* menghasilkan NDCG *value* 0,97.

Word2Vec memiliki dua arsitektur utama: *Continuous Bag-of-Words* (CBOW) dan skip-gram. Berdasarkan kajian literatur yang ekstensif, skip-gram terbukti secara konsisten lebih unggul dibanding CBOW untuk analisis sentimen. Sahbuddin & Agustian (2022) menyatakan bahwa skip-gram dipilih karena sifat *analogical reasoning*-nya yang superior, didukung oleh temuan Imaduddin, A., (2019) yang menunjukkan skip-gram lebih baik dibanding CBOW, *GloVe*, dan Doc2*vec*. Rifai (2024) dalam penelitiannya terhadap 900 ulasan TikTok menemukan skip-gram (68%) mengalahkan CBOW (66%). Temuan paling signifikan datang dari Shyahrin et al.(2023) yang menggunakan LSTM dengan Word2Vec pada ulasan aplikasi Ferizy, dengan skip-gram (88,20%) mengungguli

CBOW (74,20%) dengan gap 14%. Widyastuti et al. (2024) pada *Dataset* besar *Amazon Customer Reviews* (500 ribu sampel) menggunakan skip-gram dan mencapai akurasi 91,3%, *Precision* 90,8%, *recall* 92,1%, dan *F1-score* 91,4%. Devi (2024) juga menemukan bahwa skip-gram lebih stabil pada *dataset* berukuran kecil dan lebih baik menangkap konteks kata. Konsistensi temuan ini di berbagai domain (Twitter, ulasan aplikasi, *e-commerce*) dengan berbagai algoritma klasifikasi (SVM, LSTM) menjadikan skip-gram sebagai pilihan yang solid untuk penelitian ini.

Proses klasifikasi dilakukan setelah representasi teks terbentuk menggunakan algoritma *Support Vector Machine* (SVM). SVM dikenal sebagai salah satu metode *supervised machine learning* yang memiliki performa tinggi dalam klasifikasi teks karena mampu menemukan *Hyperplane* optimal untuk memisahkan kelas data dan efektif menangani data berdimensi tinggi hasil *embedding* Word2Vec (Widyastuti et al., 2024). Terkait pemilihan kernel, penelitian Styawati et al. (2021) membandingkan kernel Linear dan *Radial Basis Function* pada analisis sentimen program Kartu Prakerja di Twitter, dengan hasil kernel Linear mencapai akurasi 98,67% (*Precision* 98%, *recall* 99%, *F1-score* 98%) sementara *Radial Basis Function* kernel 98,34% (*Precision* 97%, *recall* 98%, *F1-score* 98%). Meskipun perbedaannya tidak terlalu besar, kernel Linear tetap unggul. Dalam penelitian lain oleh Styawati et al. (2023) tentang vaksinasi Covid-19 yang membandingkan empat kernel (Linear*, Radial Basis Function, Polynomial, Sigmoid*), *Radial Basis Function* menghasilkan akurasi tertinggi 88,8%, namun Linear *kernel* urutan kedua dengan 88,3% (gap hanya 0,5%), menunjukkan bahwa Linear *kernel* tetap kompetitif dan lebih efisien secara komputasi. Meskipun penelitian Devi (2024) menunjukkan kernel *Radial Basis Function* mencapai akurasi tertinggi (81%) pada *Dataset* ACCESS by KAI, Linear kernel tetap menjadi pilihan yang efisien dan interpretatif untuk *text classification* karena kecepatan komputasi dan kemudahan interpretasi model. Keunggulan kernel Linear terletak pada efisiensi komputasi, kemudahan interpretasi, dan performa yang stabil pada data berdimensi tinggi seperti hasil *embedding* Word2Vec, sebagaimana dibuktikan oleh Widyastuti et al. (2024) yang mencapai akurasi 91,3% pada *Dataset* besar (500 ribu sampel) menggunakan kombinasi Word2Vec skip-gram dan SVM Linear.

Kombinasi Word2Vec skip-gram dengan SVM Linear telah terbukti efektif di berbagai penelitian. Sahbuddin & Agustian (2022) menggunakan kombinasi ini untuk klasifikasi sentimen vaksinasi Covid-19 di Twitter dengan *F1-score* 65% dan akurasi 69%. Rifai (2024) menerapkannya pada ulasan TikTok dengan akurasi 68%. Widyastuti et al. (2024) mencapai hasil terbaik dengan akurasi 91,3% pada *Dataset* Amazon yang sangat besar. Konsistensi hasil positif ini menunjukkan bahwa kombinasi Word2Vec skip-gram dan SVM Linear merupakan pilihan yang tepat untuk analisis sentimen di media sosial. Berkaitan dengan skema klasifikasi, kajian literatur secara konsisten menunjukkan keunggulan klasifikasi *binary* (dua kelas) dibandingkan *multiclass*. Shanto *et al.* (2023) melaporkan peningkatan akurasi sebesar 11,8% saat beralih dari klasifikasi empat kelas ke dua kelas. Averina *et al.* (2022) mencatat peningkatan dari 36% menjadi 83%. Putri *et al.* (2024) menemukan klasifikasi dua kelas secara konsisten 11-13% lebih tinggi di semua *kernel* SVM yang diuji. Untuk menangani data bersentimen netral, Arief & Samsudin (2023) membuktikan bahwa pendekatan *removal* (penghapusan data netral) menghasilkan AUC tertinggi sebesar 0,923, lebih baik dibandingkan strategi penggabungan kelas.

Berdasarkan konvergensi temuan literatur dan hasil eksperimen awal tersebut, penelitian ini menganalisis sentimen publik terhadap program Magenta BUMN di media sosial X menggunakan kombinasi Word2Vec arsitektur skip-gram sebagai metode representasi fitur dan algoritma SVM *kernel* Linear untuk klasifikasi sentimen *binary* (positif vs. negatif). Sejauh penelusuran yang dilakukan, belum ditemukan penelitian terdahulu yang secara khusus mengkaji program Magenta BUMN menggunakan kombinasi metode ini. Penelitian sebelumnya oleh Adam (2023) yang membahas program magang MSIB Kemendikbud hanya menggunakan metode Naïve Bayes yang relatif sederhana. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi akademis maupun praktis yang signifikan.

1.2 Rumusan Masalah

Penelitian ini dirancang untuk menjawab beberapa pertanyaan utama berikut:

1. Berapa nilai akurasi, presisi, *F1-score* dan *recall* yang dihasilkan oleh model Word2Vec skip-gram dan SVM kernel Linear dalam klasifikasi sentimen terhadap program Magenta BUMN?
2. Bagaimana hasil model Word2Vec skip-gram dan SVM kernel Linear efektif dalam mengklasifikasikan sentimen publik terhadap program Magenta BUMN di media sosial X berdasarkan hasil evaluasi metrik performa?
3. Apa saja aspek-aspek dominan yang menjadi fokus dalam sentimen positif dan negatif terhadap program Magenta BUMN berdasarkan distribusi sentimen dan frekuensi kemunculan kata kunci?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mengukur tingkat akurasi, presisi, *F1-score* dan *recall* yang dihasilkan oleh model Word2Vec skip-gram dan SVM kernel Linear dalam klasifikasi sentimen terhadap program Magenta BUMN.
2. Mengevaluasi hasil efektivitas model Word2Vec skip-gram dan SVM kernel Linear dalam mengklasifikasikan sentimen publik terhadap program Magenta BUMN di media sosial X berdasarkan metrik evaluasi.
3. Mengidentifikasi aspek-aspek dominan yang menjadi fokus dalam sentimen positif dan negatif terhadap program Magenta BUMN berdasarkan distribusi sentimen dan frekuensi kemunculan kata kunci untuk memberikan rekomendasi berbasis data.

1.4 Batasan Masalah

Penelitian ini difokuskan pada ruang lingkup tertentu untuk memastikan proses analisis berjalan secara efektif dan relevan dengan tujuan penelitian. Oleh karena itu, ditetapkan beberapa batasan masalah sebagai berikut:

1. Data yang digunakan hanya berasal dari media sosial X (Twitter) dalam periode waktu 1 Januari 2023 sampai 30 November 2025.
2. Analisis sentimen dibatasi pada dua kategori: positif dan negatif; data bersentimen netral dikeluarkan dari *dataset* pelatihan.

3. Representasi fitur teks dilakukan menggunakan model Word2Vec dengan arsitektur skip-gram, sedangkan klasifikasi sentimen dilakukan menggunakan algoritma *Support Vector Machine* (SVM) dengan kernel Linear.
4. Fokus penelitian adalah persepsi publik terhadap program Magang BUMN (Magenta) secara umum, bukan pada BUMN penyelenggara secara individual.

1.5 Manfaat Hasil Penelitian

1.5.1 Akademisi

Penelitian ini dapat menjadi referensi dalam pengembangan studi di bidang analisis sentimen, pemrosesan bahasa alami, serta penerapan *machine learning* untuk teks berbahasa Indonesia. Selain itu, penelitian ini berpotensi memperkaya sumber data dan studi perbandingan metode klasifikasi sentimen.

1.5.2 Masyarakat

Penelitian ini dapat membantu calon peserta magang memperoleh informasi yang lebih objektif mengenai pengalaman dan persepsi terhadap Magenta BUMN, sehingga dapat menjadi pertimbangan dalam mengambil keputusan.