BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam negara demokrasi, pemilihan legislatif 2024 adalah peristiwa penting yang memungkinkan penduduk secara langsung memilih pejabat publik (Amir, 2020). Persepsi masyarakat terhadap pemilihan legislatif (pileg) terdiri dari informasi tentang pileg dan dukungan untuk kemenangan. Persepsi yang negatif terdiri dari kata-kata buruk, caci maki, dan penghinaan terhadap partai politik atau calon pemimpin (Firdlous dan Andrian, 2022). Media sosial digunakan untuk mendukung atau meningkatkan popularitas calon pemimpin, mengurangi popularitas lawan mereka, dan mendapatkan massa (Putra, dkk., 2022).

Jumlah orang Indonesia yang menggunakan internet pada awal tahun 2021 sebesar 202,6 juta orang, dibandingkan dengan populasi total Indonesia yang mencapai 274,9 juta orang pada Januari 2020 (Hootsuite, 2022). *Twitter* telah berkembang menjadi salah satu *platform* media sosial di internet yang paling banyak digunakan untuk berbagi pendapat politik dan informasi. Badan Intelejen Negara (BIN) melakukan analisis media sosial dengan tema pemilu 2024 dari 1 Januari 2022 hingga 13 April 2023. Intelejen Sosio Analitik (ISA) menghasilkan hasil terbanyak sebaran isu, dengan 8.312.553 pos di *Twitter* (Nurrahman, 2023).

Klasifikasi teks digunakan dalam penelitian ini untuk mengklasifikasi teks ke dalam kategori yang telah ditentukan terkait pemilihan legislatif 2024 karena kemampuan untuk menganalisis sejumlah besar data teks dengan cepat dan efisien (Vindua dan Zailani, 2023). Klasifikasi teks dapat membantu melihat teks yang muncul secara alami di media sosial atau situs *web*. Penelitian Trusca (2019) menemukan model *Word2vec* sangat efisien untuk representasi teks dibandingkan *Doc2vec*. Dalam penelitian assidyk (2020) yang menggunakan klasifikasi *K-Nearest Neighbor*, ditemukan bahwa TF-IDF lebih baik dalam *confusion matrix* kinerja model, tetapi TF-RF masih kalah, dengan akurasi 63%. (Assidyk dkk., 2020) (Truşcă, 2019)

Penelitian perbandingan metode klasifikasi teks TF-IDF dan *Word2vec* dibutuhkan untuk mengklasifikasikan data penelitian menggunakan model yang tepat bagi peneliti (Cahyani dan Patasik, 2021). Metode TF-IDF dan *Word2vec*, yang menggunakan klasifikasi *Support Vector Machine* (SVM), digunakan untuk membandingkan klasifikasi teks pada data *Twitter*.

Peneliti membandingkan metode *Word2vec* dan TF-IDF untuk klasifikasi teks di *Twitter*. *Word2vec* adalah teknik *word embedding* populer yang merepresentasikan kata-kata sebagai vektor padat dalam ruang berdimensi tinggi, menangkap hubungan semantik antara kata-kata (Fauzi, 2022). TF-IDF, di sisi lain, adalah ukuran statistik yang mencerminkan pentingnya sebuah kata dalam sebuah dokumen berdasarkan frekuensi dan kelangkaannya dalam korpus (Kurniawan dan Maharani, 2020)

Penelitian ini merupakan pembaruan dari penelitian Dewi (2022) serta penelitian Cahyani, dkk. (2021). Penelitian Dewi (2022) mengevaluasi sentimen terhadap *tweet* yang berkaitan dengan vaksinasi *Covid-19* dengan menggunakan metode *word embedding* TF-IDF dan *Word2Vec* yang menggunakan *Recurrent Neural Network* (RNN), dengan 6490 data *tweet* dan perbandingan 7:3 untuk data pelatihan dan uji coba. Hasil penelitian menunjukkan bahwa model RNN-*Word2vec* mencapai akurasi 53%, presisi 56%, dan *recall* 78%, sementara model RNN-TF-IDF mencapai akurasi 51%, presisi 51%, dan *recall* 100%. Penelitian Cahyani, dkk. (2021) membandingkan kinerja model TF-IDF dan *Word2vec* untuk merepresentasikan fitur dalam klasifikasi teks emosi menggunakan metode *Support Vector Machine* (SVM) dan *Multinomial Naïve Bayes* (MNB) untuk klasifikasi teks emosional pada data tweet *commuter line* dan *transjakarta*. Metode SVM dengan TF-IDF menghasilkan akurasi tertinggi dibandingkan metode lain dalam klasifikasi, kemudian diikuti oleh MNB dengan TF-IDF, dan yang terakhir adalah SVM dengan *Word2vec*. Kemudian, evaluasi menggunakan hasil presisi, *recall*, dan *f1-score* menunjukkan bahwa SVM dengan TF-IDF memberikan metode terbaik secara keseluruhan. SVM merupakan algoritma yang paling cocok untuk mengklasifikasikan data, Penelitian masih bisa ditingkatkan dengan penambahan jumlah data (Pamungkas dan Kharisudin, 2021). Penelitian akan berbeda dari penelitian sebelumnya, seperti: *balancing* kelas data, fektor fitur dengan *n-gram*

diterapkan, adanya proses *cleaning* nilai duplikat dan nilai kosong untuk mengurangi *buzzer* (akun yang mendapat imbalan tertentu yang berperan untuk memperluas suatu informasi melalui aktivitas retweet terkait narasi, tautan dan hashtag harian hingga dapat dilihat oleh masyarakat dalam bentuk *trending* topik) (Felicia dan Loisa, 2018), normalisasi *slang* Bahasa Indonesia, dan penggunaan 2 label data yaitu positif dan negatif. Penelitian juga membandingkan kinerja berdasarkan pembagian data dengan model evaluasi dan model validasi pada *Word2vec* dan TF-IDF menggunakan klasifikasi *Support Vector Machine* (SVM) dalam klasifikasi teks pemilihan legislatif 2024 di Indonesia.

Dalam konteks ini, masalah yang ditemukan adalah jumlah *tweet* yang banyak di media sosial *Twitter* mengenai Pemilihan legislatif 2024 atau pileg 2024 dibutuhkan metode teknik pengolahan bahasa alami yang memiliki kinerja tinggi untuk klasifikasi teks publik dan untuk mengetahui kelebihan dan kekurangan masing-masing metode, perbandingan harus dilakukan antara *Word2vec* dan TF-IDF sebagai teknik pengolahan bahasa alami dengan klasifikasi SVM. Penelitian ini diharapkan dapat meningkatkan pemahaman dan menemukan model terbaik diantara *Word2vec* dan TF-IDF dalam klasifikasi teks terkait pemilihan legislatif 2024 di Indonesia.

Berdasarkan latar belakang yang telah diberikan, penulis menarik judul "PERBANDINGAN METODE *WORD2VEC* DAN TF-IDF DENGAN SVM UNTUK KLASIFIKASI TEKS PADA MEDIA SOSIAL *TWITTER* (STUDI KASUS PEMILIHAN LEGISLATIF 2024)" untuk menyelidiki masalah tersebut.

1.2 Rumusan Masalah Penelitian

Berdasarkan latar belakang yang telah dipaparkan mengenai fenomena yang terjadi, menghasilkan rumusan masalah sebagai berikut:

1) Bagaimana perbedaan hasil perbandingan uji evaluasi *confusion matrix* tingkat *accuracy*, *precision*, *recall* dan *f1-score* pada metode *Word2vec* dan TF-IDF dengan klasifikasi SVM untuk klasifikasi teks untuk pemilihan legislatif 2024 dari media sosial *Twitter*?

2) Bagaimana perbedaan hasil perbandingan uji validasi *K-fold cross validation* tingkat *accuracy, precision, recall* dan *f1-score* pada metode *Word2vec* dan TF-IDF dengan klasifikasi SVM untuk klasifikasi teks untuk pemilihan legislatif 2024 dari media sosial *Twitter*?

1.3 Tujuan Penelitian

Dalam penelitian ini menghasilkan tujuan penelitian, yaitu:

1) Mengetahui perbedaan hasil perbandingan uji evaluasi *confusion matrix* tingkat akurasi, *precision, recall* dan *f1-score* pada metode *Word2vec* dan TF-IDF dengan klasifikasi SVM
2) Mengetahui perbedaan hasil perbandingan uji validasi *K-fold cross validation* tingkat akurasi, *precision, recall* dan *f1-score* pada metode *Word2vec* dan TF-IDF dengan klasifikasi SVM

1.4 Manfaat Penelitian

Berdasarkan sub bab sebelumnya, penelitian menghasilkan manfaat penelitian, antara lain sebagai berikut:

1) Memberikan wawasan mengenai perbandingan antara *Word2vec* dan TF-IDF dalam melakukan klasifikasi teks pada pemilihan legislatif 2024.
2) Memberikan informasi tentang kinerja SVM dalam mengklasifikasikan teks pada pemilihan legislatif 2024 menggunakan *Word2vec* dan TF-IDF.
3) Memberikan kontribusi bagi pengembangan teknik klasifikasi teks dalam pengolahan bahasa alami.

1.5 Batasan Masalah

Pembatasan yang diterapkan dalam penelitian ini, sebagai berikut:

1) Studi ini akan membatasi teks pada pemilihan legislatif 2024 melalui pengumpulan data melalui media sosial *Twitter* dalam rentang waktu bulan April 2023 sampai Juni 2023, karena pada rentang waktu tersebut adanya pendaftaran

calon anggota DPR, DPRD provinsi dan DPRD kabupaten/kota (InfoPublik, 2023), adanya parpol dan caleg gencarkan sosialisasi (Rahayu, 2023), partai menetapkan calon-calon legislatif (Abduh, 2023) (Marjaya, 2023).

2) Studi ini akan membatasi teks pada pemilihan legislatif 2024 melalui pengumpulan data melalui media sosial *Twitter* dalam Bahasa Indonesia.
3) Fokus penelitian akan difokuskan pada komparasi antara metode *Word2vec* dan TF-IDF dengan klasifikasi *Support Vector Machine* (SVM) dalam klasifikasi teks terkait pemilihan legislatif 2024.

1.6 Sistematika Penulisan Skripsi

Studi ini terdapat lima bab dalam penelitian ini.

Bab 1 Pendahuluan, Membahas konteks dan alasan mengapa penelitian ini dilakukan. Rumusan masalah: Menjelaskan pertanyaan-pertanyaan penelitian yang ingin dijawab, Tujuan penelitian: Mengungkapkan tujuan utama dari penelitian ini, Manfaat penelitian: Menjelaskan kontribusi penelitian ini terhadap ilmu pengetahuan atau praktik, Hipotesis penelitian: Merumuskan dugaan awal yang akan diuji dalam penelitian, Sistematika penulisan: Menjelaskan struktur dan urutan penulisan dalam penelitian ini, Hipotesis Penelitian: Bagian ini berisi hipotesis penelitian yang dirumuskan berdasarkan teori-teori yang relevan. Hipotesis merupakan dugaan awal yang akan diuji dalam penelitian

Bab 2 Studi Pustaka menjelaskan berbagai teori dasar penelitian. Kajian Pustaka: Bagian ini membahas teori-teori yang menjadi dasar penelitian ini dan membahas penelitian sebelumnya yang relevan atau terkait langsung dengan masalah yang dibahas dalam penelitian ini. Penulis dapat menambah penelitian tambahan untuk mendukung penelitian saat ini. Teori-teori dan Konsep-konsep: Dalam bagian ini, penulis dapat mengutip teori-teori dari para ahli yang relevan dengan judul skripsi mereka dan menjelaskan konsep-konsep yang digunakan dalam penelitian.

Bab 3 membahas subjek dan metodologi penelitian yang digunakan penulis. *Design Science Research Method* (DSRM): Bagian ini menjelaskan metode penelitian DSRM. Tahapan Penelitian: Bagian ini membahas proses perencanaan, pengumpulan

data, analisis data, dan evaluasi. Objek dan Subjek Penelitian: Bagian ini memberikan penjelasan tentang subjek dan objek penelitian. Populasi dan Sampel Penelitian: Bagian ini memberikan penjelasan tentang populasi dan sampel yang digunakan dalam penelitian. Metode Pengumpulan Data: Metode dalam penelitian untuk pengumpulan data dibahas seperti observasi dan lain lain dibahas di bagian ini. Jenis dan Sumber Data: Jenis data yang digunakan dalam penelitian dijelaskan dalam bagian ini. Metode Analisis Data: Metode seperti analisis deskriptif dibahas dalam bagian ini.

Bab 4 adalah hasil dan pembahasan, yang membahas diskusi yang didasarkan pada temuan penelitian. Hasil penelitian: Menyajikan temuan-temuan yang diperoleh dari penelitian ini. Pembahasan: Menganalisis dan mendiskusikan temuan penelitian berdasarkan teori-teori yang relevan.

Bab 5 Kesimpulan dan Saran: Bab ini menjelaskan temuan penelitian dan membuat saran yang relevan. Kesimpulan: Merangkum temuan penelitian dan menjawab pertanyaan penelitian. Saran: Memberikan saran yang relevan berdasarkan temuan penelitian

1.7 Hipotesis Penelitian

Penelitian ini menggunakan dua hipotesis, yaitu:

1) Hipotesis nol (H0): perbedaan signifikan tidak ada diantara performa metode *Word2vec* dan TF-IDF dalam klasifikasi teks pada pemilihan legislatif 2024 dari media sosial *Twitter*.
2) Hipotesis alternatif (H1): terdapat perbedaan signifikan antara performa metode *Word2vec* dan TF-IDF dalam klasifikasi teks pada pemilihan legislatif 2024 dari media sosial *Twitter*.