BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam Pemrosesan Bahasa Alami atau *Natural Language Processing* (NLP), *Word Embedding* adalah teknik untuk mengenali makna teks secara matematis. Konsep dasar *Word Embedding* adalah mengubah setiap kata menjadi representasi numerik berupa vektor. Kata-kata yang memiliki makna atau konteks penggunaan yang serupa akan direpresentasikan oleh vektor yang saling berdekatan di ruang vektor. Model klasik seperti Word2Vec gagasan Mikolov dkk. (2013), sangat efisien mempelajari pola ini dari data teks yang besar. Namun, model tersebut memiliki kelemahan, ia menghasilkan representasi yang kaku. Word2Vec hanya memetakan satu vektor global untuk masing-masing kata. Padahal, makna kata di dunia nyata sering kali berubah makna mengikuti konteks. Misalnya, kata "*apple*" bisa bermakna buah di konteks pertanian, tetapi berubah drastis menjadi nama perusahaan di konteks teknologi [1].

Kekakuan representasi ini secara langsung mengurangi ketajaman analisis NLP saat dihadapkan pada kumpulan dokumen dari berbagai domain berbeda. Domain di sini merujuk pada sekelompok data dengan identitas khas yang seragam, bisa karena perbedaan topik, waktu, atau apapun yang bisa di kategorikan. Solusi paling primitif untuk mengatasi hal ini adalah memecah korpus dan melatih model *Word Embedding* secara terpisah pada masing-masing domain. Namun, *Word Embedding* sangat sensitif terhadap arah dan skala rotasi. Akibatnya, dua kata dengan arti identik dari dua domain berbeda bisa berada pada sudut koordinat yang berjauhan [2]. Peneliti umumnya menggunakan tahap penyelarasan (*alignment*) untuk menarik paksa semua vektor kembali ke ruang kordinat yang seragam. Meski lazim dilakukan, penyelarasan terpisah ini memiliki kelemahan bawaan (*error propagation*). Ketimpangan jumlah kemunculan kata atau distorsi distribusi antar domain dapat mengacaukan hasil penyelarasan [3].

Kebutuhan mendesak akan model yang bisa membaca pergeseran makna mendorong lahirnya metode-metode baru. Kemampuan mendeteksi perubahan nuansa kata ini berdampak krusial pada performa aplikasi NLP tingkat lanjut,

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

mulai dari analisis sentimen lintas waktu hingga pemrosesan dokumen multi-topik. Sebagai bentuk perbaikan terhadap rapuhnya teknik penyelarasan terpisah, Yao dkk. mengajukan *Dynamic Word2Vec* (DW2V) yang dirancang untuk melatih sekaligus menyelaraskan vektor kata secara sekaligus. Namun, DW2V dibangun dengan asumsi bahwa perubahan makna bergerak lurus secara kronologis, dan tidak mampu mengatasi pada struktur interaksi domain yang lebih acak [4].

Merespons permasalahan pada DW2V, Lassner dkk. mereformulasi ulang konsep tersebut menjadi *Word2Vec with Structure Prediction* (W2VPred). W2VPred menghasilkan *Word Embedding* untuk setiap domain sekaligus memprediksi bagaimana domain-domain tersebut saling memiliki ikatan kedekatan. Arsitektur ini menciptakan jembatan pertukaran informasi antar domain yang saling beririsan. Sehingga, *Word Embedding* yang dihasilkan menjadi lebih kaya, dinamis, dan dapat menyesuaikan dengan domain konteks asalnya [2].

Secara arsitektur, W2VPred menggabungkan tiga fungsi objektif sekaligus untuk membentuk arah vektor. Komponen pertama adalah *fidelity loss* ($\mathcal{L}_F$) untuk mengukur kecocokan vektor terhadap kemunculan kata pada korpus. Komponen kedua adalah *structure loss* ($\mathcal{L}_S$) untuk memastikan sebaran vektor mengikuti pola hubungan antar domain. Komponen terakhir adalah *regularization loss* ($\mathcal{L}_{RD}$) yang berfungsi mengendalikan jarak antar domain agar tidak terlalu jauh antar masing-masing domain. Ketiganya disatukan menjadi satu kesatuan persamaan objektif:

$$\mathcal{L}_{\text{W2VPred}} = \mathcal{L}_F + \lambda\mathcal{L}_S + \tau\mathcal{L}_{RD} \tag{1.1}$$

Koefisien pembobotan λ dan τ digunakan sebagai bobot penalti *loss*. Secara standar, nilai ini ditentukan lewat pencarian acak (*grid search*). Pendekatan ini menuntut model dilatih ulang dari awal hanya untuk mengevaluasi berbagai kombinasi skalar statis. Pencarian *grid search* sangat memakan daya komputasi dan menghasilkan bobot statis yang sama sekali tidak dapat merespons perubahan pola konvergensi di tengah proses optimasi [2].

Untuk mengatasi masalah penentuan koefisien pembobotan tersebut, Kendall dkk. (2018) mengusulkan metode *Uncertainty Weighting* (UW). Daripada menentukan besaran koefisien pembobotan secara acak, UW memodelkan komponen *loss* sebagai ketidakpastian atau *noise* (σ) yang bisa dipelajari langsung oleh model. Saat ada tugas yang menghasilkan *loss* yang tinggi, model secara otomatis mengurangi bobotnya untuk menghindari *overfitting*. Meskipun secara

konsep sangat baik, penerapan metode UW secara langsung pada kerangka W2VPred memiliki kekurangan matematis. UW mensyaratkan bahwa setiap komponen *loss* harus berasal dari *likelihood* data aktual. Pada model W2VPred, prasyarat ini hanya terpenuhi oleh $\mathcal{L}_F$. Dua komponen lainnya ($\mathcal{L}_S$ dan $\mathcal{L}_{RD}$) sama sekali tidak bergantung pada data observasi, melainkan berfungsi sebagai *prior* atau asumsi awal terhadap sebaran parameter yang tidak ada di data korpus. Menyamakan perlakuan antara observasi data (*likelihood*) dengan asumsi awal (*prior*) dalam satu optimasi akan merusak validitas model [5].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pemodelan probabilistik pada model W2VPred melalui pendekatan *Joint Maximum A Posteriori* (JMAP). Pada formulasi ini, $\mathcal{L}_F$ diposisikan sebagai *likelihood*. Di sisi lain, $\mathcal{L}_S$ dan $\mathcal{L}_{RD}$ dijadikan sebagai *prior*. Dengan memformulasikan setiap komponen *loss* ke dalam bentuk distribusi probabilitas, parameter ketidakpastian dari masing-masing distribusi tersebut ($\sigma_F, \sigma_S, \sigma_{RD}$) secara alamiah bertindak sebagai indikator ketidakpastian untuk mengatur bobot penalti secara dinamis. Melalui pemisahan ini, masalah pencarian bobot statis diubah menjadi estimasi ketidakpastian dinamis yang sesuai dengan aturan Bayesian.

Tahap akhir berfokus pada pelatihan model beserta nilai ketidakpastiannya. Penelitian ini menggunakan optimasi *Joint Maximum A Posteriori* (JMAP). JMAP menghindari kerumitan kalkulus Bayesian tradisional yang sulit dihitung pada matriks berskala besar. JMAP mampu memperbarui parameter *embedding* dan ketidakpastian secara bersamaan menggunakan *gradient descent* konvensional. Melalui JMAP, *grid search* manual pada W2VPred digantikan, sehingga model menyesuaikan bobot kerugiannya secara adaptif [6].

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, penelitian ini memformulasikan tiga rumusan masalah:

1. Bagaimana membangun kerangka probabilistik berbasis estimasi *Joint Maximum A Posteriori* (JMAP) pada fungsi objektif W2VPred yang memposisikan *fidelity loss* sebagai *likelihood* dan *structure loss* serta *regularization loss* sebagai *prior*?
2. Bagaimana mekanisme JMAP mengestimasi parameter *embedding* dan parameter ketidakpastian secara bersamaan sehingga menggantikan *grid search* manual pada W2VPred?

3. Bagaimana perbandingan efektivitas W2VPred-JMAP terhadap W2VPred *baseline* pada aspek konvergensi, akurasi analogi semantik, prediksi struktur antar-domain, dan analisis pergeseran makna kata?

1.3 Batasan Masalah

Batasan persoalan yang digunakan dalam penelitian ini mencakup:

1. Model yang digunakan adalah W2VPred dengan arsitektur faktorisasi matriks PPMI (*Positive Point-wise Mutual Information*).
2. Metode penyeimbangan yang dikaji difokuskan pada optimasi *Joint Maximum A Posteriori* (JMAP), yang diperbandingkan dengan model *baseline* W2VPred asli berbobot statis.
3. Pengujian model difokuskan pada efisiensi optimasi fungsi objektif, performa prediksi struktur, uji akurasi semantik, serta analisis pergeseran spasial makna kata.
4. korpus yang digunakan adalah korpus teks berbahasa Inggris WikiFoS (terbagi dalam 16 disiplin ilmu), New York Times (terbagi dalam 20 rentang waktu) dan korpus buatan untuk studi kasus evolusi makna kata ”*apple*” (domain teknologi vs *fruit*).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang ditetapkan, tujuan penelitian ini adalah:

1. Membangun kerangka probabilistik berbasis estimasi *Joint Maximum A Posteriori* (JMAP) pada fungsi objektif W2VPred yang memposisikan *fidelity loss* sebagai *likelihood* dan *structure loss* serta *regularization loss* sebagai *prior*, mengikuti prinsip Kendall et al. (2018) dengan penambahan formulasi *prior*.
2. Menganalisis mekanisme JMAP dalam mengestimasi parameter *embedding* dan parameter ketidakpastian secara bersamaan sebagai pengganti *grid search* manual pada W2VPred.
3. Mengevaluasi kualitas *word embedding* W2VPred-JMAP dibandingkan W2VPred *baseline* pada aspek konvergensi fungsi objektif, akurasi analogi semantik, prediksi struktur antar-domain, dan analisis pergeseran makna kata.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Adapun manfaat penelitian ini memberikan kontribusi signifikan baik dari aspek teoritis maupun praktis dalam pengembangan topik *word embedding* multi-domain. Secara teoritis, penelitian ini memperkokoh landasan matematis model W2VPred dengan menggantikan pendekatan heuristik deterministik menjadi kerangka kerja probabilistik berbasis *Joint Maximum A Posteriori* (JMAP), sehingga memberikan interpretasi yang jelas mengenai hubungan antara bobot fungsi objektif dan ketidakpastian (*uncertainty*) yang melekat pada masing-masing komponen *loss*. Sementara secara praktis, penerapan mekanisme pembobotan adaptif menawarkan solusi efisiensi komputasi yang krusial dengan mereduksi kebutuhan *grid search* kombinatorial bobot statis dari dua dimensi (λ, τ) menjadi satu parameter skalar (η_s), sekaligus menghasilkan model yang dinamis karena mampu menyesuaikan diri secara otomatis terhadap variasi karakteristik data dan struktur antar domain.

1.5 Metode Penelitian

Metode penelitian yang digunakan pada skripsi ini diantaranya:

1. Studi Literatur

   Mengkaji literatur terkait model W2VPred, inferensi Bayesian, dan estimasi *Joint Maximum A Posteriori* (JMAP) untuk mengidentifikasi kelemahan pendekatan pembobotan *loss* statis dan menetapkan landasan teoretis yang digunakan dalam penelitian ini.

2. Reformulasi Model W2VPred-JMAP

   Merumuskan ulang fungsi objektif W2VPred dari kerangka pembobotan skalar statis menjadi kerangka probabilistik. Setiap komponen *loss* dimodelkan sesuai perannya, yaitu sebagai *likelihood* atau *prior*, dengan parameter ketidakpastian adaptif yang diestimasi melalui *Joint Maximum A Posteriori* (JMAP).

3. Pengumpulan dan Pra-pemrosesan Data

   Mengumpulkan korpus WikiFoS (16 disiplin ilmu) dan arsip teks New York Times (20 rentang tahun). Data melalui tahap pra-pemrosesan standar sebelum dikonversi menjadi representasi matriks *Positive Point-wise Mutual Information* (PPMI) sebagai input pelatihan model.

4. Merancang Desain Eksperimen

   Merancang skenario eksperimen yang membandingkan W2VPred *baseline*

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI

(bobot statis) dengan W2VPred-JMAP untuk memvalidasi keunggulan model pada aspek efisiensi optimasi, akurasi semantik, prediksi struktur, dan ketajaman deteksi evolusi makna kata.

5. Implementasi dan Pelatihan

   Mengimplementasikan model dalam bahasa pemrograman Python. Kedua skenario model dilatih menggunakan korpus yang sama untuk memastikan hasil perbandingan yang adil dan konsisten.

6. Evaluasi dan Analisis Performa

   Mengevaluasi hasil pelatihan secara komprehensif, mencakup analisis konvergensi fungsi objektif, pengujian akurasi analogi semantik kata, evaluasi keandalan prediksi struktur afinitas antar domain, serta analisis pergeseran spasial makna kata lintas domain.

1.6 Sistematika Penulisan

Struktur penulisan skripsi ini terdiri dari lima bab, diantaranya:

1. **BAB I: PENDAHULUAN**
   Bab ini menguraikan latar belakang permasalahan, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

2. **BAB II: LANDASAN TEORI**
   Bab ini membahas teori pendukung penelitian, meliputi konsep *word embedding*, arsitektur W2VPred, metode optimasi, serta landasan matematika dan probabilistik yang relevan, seperti distribusi Gaussian dan Gibbs, *Maximum Likelihood Estimation* (MLE), inferensi Bayesian, dan *Joint Maximum A Posteriori* (JMAP).

3. **BAB III: PENGEMBANGAN MODEL W2VPRED DENGAN PEMBOBOTAN ADAPTIF BERBASIS JMAP**
   Bab ini menguraikan tahapan penelitian secara rinci, mulai dari reformulasi probabilistik fungsi objektif W2VPred ke dalam kerangka JMAP, strategi estimasi parameter ketidakpastian adaptif beserta justifikasi teoretisnya, prosedur pelatihan dan algoritma optimasi yang digunakan, deskripsi korpus, hingga desain eksperimen yang akan dilakukan.

4. **BAB IV: STUDI KASUS W2VPRED-JMAP PADA DATA SINTESIS DAN KORPUS NYATA**
   Bab ini menyajikan hasil eksperimen dan analisis komparatif antara model

W2VPred *baseline* dengan W2VPred-JMAP, mencakup validasi mekanisme adaptif pada data sintesis, analisis konvergensi fungsi objektif, dinamika estimasi parameter ketidakpastian, performa prediksi struktur, akurasi analogi kata, dan evolusi makna kata lintas domain pada korpus nyata.

5. **BAB V: KESIMPULAN DAN SARAN**

   Bab ini merangkum temuan utama penelitian, menarik kesimpulan berdasarkan rumusan masalah, dan memberikan saran untuk penelitian selanjutnya.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG