BAB I

PENDAHULUAN

1.1 Latar Belakang

Upaya memodelkan bahasa secara matematis bukanlah gagasan baru dalam sejarah ilmu pengetahuan. Jauh sebelum era komputer modern, matematikawan Rusia Andrey Markov pada tahun 1913 telah menunjukkan bahwa kemunculan huruf dalam sebuah teks dapat didekati secara statistik melalui rantai probabilitas, di mana kemunculan suatu unit teks bergantung pada unit-unit sebelumnya [1]. Kerangka berpikir inilah yang kemudian diformalkan oleh Claude Shannon pada tahun 1948 dalam *A Mathematical Theory of Communication*, yang mengadopsi rantai Markov untuk memodelkan bahasa sebagai proses stokastik dan meletakkan fondasi matematis bagi seluruh perkembangan pemodelan bahasa selanjutnya [2]. Meskipun sempat diabaikan selama beberapa dekade akibat kritik Chomsky (1956–1957) yang menilai proses Markov berhingga tidak mampu merepresentasikan pengetahuan gramatikal manusia secara utuh [3], pendekatan statistik ini dihidupkan kembali oleh Fred Jelinek dan koleganya di IBM pada dekade 1970–1980-an, yang membuktikan secara empiris bahwa model statistik berbasis korpus besar mampu unggul dalam aplikasi nyata seperti pengenalan ucapan [4]. Sejak saat itu, model *n-gram* menjadi tulang punggung *Statistical Language Model* (SLM) yang mendominasi riset pemrosesan bahasa alami hingga awal tahun 2000-an.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Secara teknis, SLM berbasis *n-gram* bekerja dengan mengestimasi probabilitas kemunculan sebuah kata berdasarkan $n-1$ kata sebelumnya melalui penghitungan frekuensi statistik murni dari korpus data latih [5]. Pendekatan ini sederhana dan mudah diinterpretasikan, namun memiliki kelemahan mendasar yang dikenal sebagai *curse of dimensionality* (kutukan dimensionalitas): ukuran tabel probabilitas tumbuh secara eksponensial seiring bertambahnya n, sementara kombinasi kata yang valid secara gramatikal namun kebetulan tidak pernah muncul dalam data latih (*data sparsity*) akan selalu diberi probabilitas nol. Akibatnya, SLM gagal melakukan generalisasi semantik; model ini tidak "menyadari" bahwa dua istilah teologis tertentu memiliki kedekatan makna meskipun keduanya jarang atau tidak pernah muncul berdampingan secara literal dalam korpus.

Titik balik penting untuk mengatasi kelemahan tersebut hadir ketika Bengio et al. (2003) memperkenalkan *Neural Probabilistic Language Model* (NPLM) [6]. Alih-alih menghitung frekuensi kemunculan kata secara diskrit, NPLM memetakan setiap kata ke dalam ruang vektor kontinu berdimensi rendah (*word embedding*) dan melatih jaringan saraf *feed-forward* untuk mempelajari fungsi probabilitas bersyarat $P(w \mid \text{konteks})$ secara langsung dari data. Representasi vektor inilah yang memungkinkan NPLM melakukan generalisasi semantik yang tidak mampu dilakukan SLM: kata-kata dengan makna atau fungsi yang mirip akan memiliki representasi vektor yang berdekatan, sehingga model dapat memberikan probabilitas yang masuk akal bahkan untuk kombinasi kata yang belum pernah dilihat secara persis. Meski demikian, arsitektur NPLM orisinal tahun 2003 tetap memiliki keterbatasan struktural: jaringannya relatif dangkal (hanya satu *hidden layer*) dan hanya mengandalkan konteks jendela lokal berukuran k kata sebelumnya, sehingga kesulitan menangkap dependensi jarak jauh (*long-range dependencies*) yang justru sangat umum ditemukan dalam narasi penafsiran ayat yang panjang dan berlapis.

Untuk mengatasi keterbatasan dependensi jarak jauh ini, arus utama penelitian sejak pertengahan 2010-an bergeser ke arsitektur *recurrent* dan kemudian Transformer (misalnya BERT atau GPT), yang memang unggul dalam menangkap konteks panjang [5]. Namun demikian, model-model raksasa berbasis Transformer ini menuntut biaya komputasi dan memori yang sangat besar, serta bersifat *black box* sehingga sulit ditelusuri alur logikanya. Kondisi inilah yang melatarbelakangi mengapa Sun dan Iyyer (2021) merasa perlu untuk "meninjau ulang" (*revisit*) NPLM yang sederhana [7], alih-alih mengikuti tren memperbesar model, mereka mempertanyakan apakah NPLM klasik, yang sempat ditinggalkan, sebenarnya dapat dimodernisasi dengan teknik-teknik arsitektur kontemporer (koneksi residual, *layer normalization*, aktivasi ReLU, mekanisme agregasi konteks global, dan *adaptive softmax*) agar kompetitif tanpa harus menanggung beban komputasi Transformer. Hasilnya, NPLM modern ini terbukti mampu menyaingi performa Transformer pada konteks input pendek, dengan kompleksitas komputasi yang jauh lebih rendah dan alur logika yang jauh lebih mudah diinterpretasikan, karakteristik yang justru sangat dibutuhkan untuk riset pada domain teks khusus seperti tafsir Al-Qur'an.

Relevansi pemodelan bahasa terhadap teks keagamaan sendiri sejalan dengan semangat Al-Qur'an itu sendiri. Allah *Subhanahu wa Ta'ala* berfirman dalam QS. Al-'Alaq [96]: 1–5:

Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan. Dia telah menciptakan manusia dari segumpal darah. Bacalah, dan Tuhanmulah Yang Maha Mulia. Yang mengajar (manusia) dengan pena. Dia mengajarkan manusia apa yang tidak diketahuinya.

Ayat pertama yang diturunkan ini secara eksplisit menempatkan aktivitas membaca (*iqra'*) dan proses pembelajaran sebagai fondasi peradaban manusia, selaras dengan semangat *Islamic Digital Humanities* kontemporer yang mengintegrasikan alat komputasional untuk mengungkap pola tersembunyi dalam teks suci [8]. Seiring perkembangan ilmu tafsir selama berabad-abad, para ulama telah menghasilkan ribuan karya penjelas yang kini tersedia dalam berbagai bahasa, termasuk bahasa Indonesia, membentuk korpus data tekstual yang masif dan membuka peluang analisis struktur linguistik serta relasi semantik antaristilah teologis secara komputasional [9].

Namun demikian, penerapan NPLM pada korpus tafsir berbahasa Indonesia ini bukan tanpa tantangan. Teks tafsir memiliki karakteristik domain yang khas: kosakatanya mencakup istilah-istilah teologis dengan makna yang sangat spesifik (*tawakkal*, *istidraj*, *taubat*), struktur kalimatnya dipengaruhi oleh konvensi penulisan keislaman, dan relasi antarkata sering mencerminkan logika teologis yang tidak dapat ditangkap oleh model yang hanya dilatih pada teks umum [10]. Sementara itu, penelitian NLP pada bahasa Indonesia, khususnya subdomain tafsir, masih tertinggal dibanding bahasa dengan sumber daya komputasional yang lebih kaya; upaya awal seperti pengelompokan tematik dan pencarian semantik berbasis *word embedding* memang mulai menunjukkan potensi [11, 12], tetapi pengembangan model bahasa yang secara khusus dirancang untuk memprediksi kata dan menangkap nuansa semantik teks tafsir masih menjadi celah penelitian yang signifikan.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Berangkat dari kerangka NPLM modern oleh Sun dan Iyyer (2021) [7], penelitian ini memanfaatkan mekanisme agregasi konteks global untuk menyuntikkan informasi semantik dari seluruh kalimat atau paragraf secara terstruktur ke dalam tugas prediksi kata berikutnya. Hal ini krusial untuk teks tafsir, di mana makna sebuah istilah teologis sering ditentukan oleh gagasan utama (*global context*) dari keseluruhan ayat, bukan hanya oleh kata-kata yang mengapitnya secara langsung. Representasi yang dihasilkan oleh model ini kemudian dievaluasi dari dua sisi: pertama, analisis distribusi probabilitas bersyarat $P(w \mid \text{konteks})$ untuk memeriksa kata apa saja yang diprediksi model saat dihadapkan pada konteks teologis tertentu; kedua, kedekatan semantik berbasis *cosine similarity* pada

embedding yang dipelajari, untuk memverifikasi apakah relasi antaristilah teologis dalam ruang vektor buatan komputer benar-benar mencerminkan pemahaman konseptual yang sejalan dengan kajian keislaman.

1.2 Rumusan Masalah

Berdasarkan identifikasi masalah di atas, penelitian ini merumuskan tiga pertanyaan penelitian sebagai berikut:

1. Bagaimana rancangan dan penerapan arsitektur NPLM modern berdasarkan gagasan *revisiting* Sun dan Iyyer (2021), khususnya melalui pemanfaatan konteks global, dalam memodelkan distribusi probabilitas kata pada korpus terjemahan dan tafsir Al-Qur'an berbahasa Indonesia?
2. Sejauh mana pemanfaatan konteks global melalui *average pooling* dan *kernel aggregator* memberikan perbedaan kinerja dibandingkan NPLM tanpa konteks global?
3. Bagaimana karakteristik representasi kata yang dihasilkan oleh konfigurasi arsitektur NPLM terbaik berdasarkan analisis distribusi probabilitas bersyarat dan kedekatan antarkata dalam ruang *embedding*?

1.3 Batasan Masalah

Untuk menjaga fokus dan kelayakan penelitian, batasan masalah yang digunakan adalah sebagai berikut:

1. Penelitian ini berfokus pada penerapan dan evaluasi arsitektur NPLM modern berdasarkan gagasan *revisiting* yang dikemukakan oleh Sun dan Iyyer (2021). Penelitian ini tidak mengusulkan arsitektur baru yang sepenuhnya terpisah dari kerangka tersebut.
2. Data yang digunakan berupa korpus berbahasa Indonesia yang terdiri atas teks terjemahan Al-Qur'an, Tafsir Jalalayn, dan Tafsir Al-Muntakhab. Korpus tersebut terdiri atas 6.236 baris data yang merepresentasikan ayat dari 114 surah.
3. Pembaruan arsitektur yang digunakan mengacu pada kerangka NPLM modern, yang mencakup jaringan *feed-forward* residual, koneksi residual, *layer normalization*, fungsi aktivasi ReLU, agregasi konteks global, dan *adaptive softmax*. Penelitian ini membandingkan tiga konfigurasi, yaitu NPLM tanpa konteks global, NPLM dengan *average pooling*, dan NPLM

dengan *kernel aggregator*.

4. Penentuan konfigurasi pelatihan dilakukan melalui analisis sensitivitas hiperparameter menggunakan pendekatan *One-Factor-at-a-Time* (OFAT) pada subsampel data pelatihan sebesar 20%. Hasilnya kemudian diperiksa melalui eksperimen kombinasi naif untuk mengamati kemungkinan adanya interaksi antarhiperparameter.
5. Perbandingan tiga konfigurasi arsitektur dilakukan secara terkontrol dengan menggunakan hiperparameter pelatihan yang sama. Oleh karena itu, perbedaan kinerja antarmodel dianalisis terutama berdasarkan mekanisme konteks global yang digunakan.
6. Evaluasi kinerja model dibatasi pada *validation loss* dan *validation perplexity*. Penelitian ini tidak mencakup tugas turunan seperti klasifikasi tema ayat, penerjemahan otomatis, peringkasan teks, atau sistem tanya jawab.
7. Analisis representasi kata dilakukan melalui distribusi probabilitas bersyarat, *cosine similarity*, *nearest neighbor*, PCA, dan t-SNE. Analisis tersebut digunakan untuk mengamati pola distribusional dan kedekatan antarkata dalam korpus, bukan untuk membuktikan bahwa model memiliki pemahaman teologis secara utuh.
8. Analisis representasi difokuskan pada sejumlah istilah dan pasangan kata yang dipilih berdasarkan relevansinya terhadap wacana keagamaan dalam korpus. Hasil analisis bersifat eksploratif dan tidak dimaksudkan untuk menggantikan penafsiran atau validasi teologis oleh ahli.

1.4 Tujuan Penelitian

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Berdasarkan rumusan masalah yang telah ditetapkan, penelitian ini bertujuan untuk:

1. Menerapkan dan mengevaluasi arsitektur NPLM modern berdasarkan gagasan *revisiting* Sun dan Iyyer (2021) pada korpus terjemahan dan tafsir Al-Qur'an berbahasa Indonesia;
2. Menganalisis pengaruh penggunaan konteks global melalui *average pooling* dan *kernel aggregator* terhadap kinerja arsitektur NPLM dibandingkan dengan konfigurasi tanpa konteks global berdasarkan *validation loss* dan *validation perplexity*; serta
3. Menganalisis karakteristik representasi kata yang dihasilkan oleh konfigurasi

arsitektur NPLM terbaik melalui distribusi probabilitas bersyarat, *cosine similarity*, *nearest neighbor*, dan visualisasi ruang *embedding*.

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam proses menyelesaikan penelitian ini meliputi:

1. **Studi Literatur**
   Metode utama dalam penelitian ini bersifat studi literatur. Pada tahap ini, peneliti mengumpulkan berbagai kajian teoritis mengenai Pemrosesan Bahasa Alami (NLP), *Neural Probabilistic Language Model* (NPLM) dan representasi vektor kata (*word embedding*). Sumber referensi kajian literatur ini mencakup buku, jurnal ilmiah, artikel prosiding, serta karya ilmiah lainnya yang relevan untuk merancang arsitektur model dan menguatkan pemahaman peneliti.

2. **Implementasi dan Analisis Komputasional**
   Tahap selanjutnya, peneliti melakukan implementasi dan analisis melalui pendekatan komputasional untuk membangun, melatih, dan mengevaluasi model NPLM dengan pembaruan arsitektur pada korpus teks keagamaan berbahasa Indonesia. Implementasi ini mencakup:

   - Analisis sensitivitas *hyperparameter* pada *subsample* data.
   - Validasi empiris terhadap interaksi antar-*hyperparameter* melalui eksperimen kombinasi naif.
   - Pelatihan penuh (*full training*) pada tiga konfigurasi arsitektur untuk *ablation study*.
   - Evaluasi multi-metrik meliputi *Validation Perplexity*, stabilitas pelatihan, dan kualitas representasi semantik.
   - Analisis distribusi probabilitas bersyarat pada konteks-konteks teologis.
   - Representasi kedekatan semantik berbasis *cosine similarity*.
   - Visualisasi *embedding* menggunakan t-SNE dan PCA.

   Seluruh implementasi dilakukan menggunakan kerangka kerja *deep learning* PyTorch dengan evaluasi yang rigor untuk memastikan validitas dan reliabilitas hasil penelitian.

SUNAN GUNUNG DJATI
BANDUNG

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini adalah:

BAB I PENDAHULUAN

Bab ini berisi mengenai latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini menguraikan teori dan konsep yang mendasari penelitian, meliputi pemrosesan bahasa alami, model bahasa statistik dan *N-Gram*, *Neural Probabilistic Language Model* (NPLM) beserta pembaruan arsitekturnya, serta metrik evaluasi dan representasi semantik yang digunakan pada bab-bab berikutnya.

BAB III METODE PENELITIAN

Bab ini menjelaskan desain penelitian secara menyeluruh, meliputi sumber dan karakteristik korpus data, prosedur prapemrosesan teks, rancangan arsitektur model untuk ketiga konfigurasi *ablation study*, serta prosedur pelatihan dan evaluasi model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi penerapan model dan membahas hasil eksperimen, meliputi perbandingan performa antarkonfigurasi model, analisis distribusi probabilitas dan visualisasi kedekatan semantik.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan dari pembahasan yang telah dianalisis serta saran untuk pengembangan penelitian selanjutnya.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG