2. LANDASAN TEORI

2.1. Tinjauan Pustaka

2.1.1. Hak Paten

Hak paten adalah salah satu bentuk perlindungan kekayaan intelektual yang memberikan hak eksklusif kepada penemu atas invensinya untuk jangka waktu tertentu. Hak ini bertujuan untuk mendorong inovasi dan memberikan insentif bagi penemu melalui perlindungan hukum. Menurut *World Intellectual Property Organization* (WIPO), paten mencakup hak untuk melarang pihak lain membuat, menggunakan, atau menjual invensi tanpa izin selama masa paten berlaku (WIPO, n.d.).

Sebuah dokumen paten umumnya terdiri dari beberapa bagian utama, yaitu:

1. Judul Invensi, memberikan gambaran singkat tentang invensi.
2. Abstrak, merupakan ringkasan dari invensi yang berisi tujuan, manfaat, dan esensi invensi secara ringkas.
3. Deskripsi Lengkap, menguraikan latar belakang masalah, solusi yang ditawarkan oleh invensi, dan penjelasan teknis.
4. Klaim, bagian yang mendefinisikan ruang lingkup perlindungan hukum dari paten.
5. Gambar Pendukung (jika ada), untuk memperjelas deskripsi teknis.

Abstrak paten memiliki peranan yang sangat penting sebagai ringkasan dari informasi yang terkandung dalam dokumen paten. Abstrak ini berfungsi untuk memberikan gambaran umum mengenai invensi yang diajukan, sehingga memudahkan para peneliti, pengembang, dan pihak-pihak terkait lainnya untuk memahami inti dari invensi tersebut tanpa perlu membaca keseluruhan dokumen. Menurut Undang-Undang Nomor 14 Tahun 2001 tentang Paten, abstrak harus mencakup pokok masalah, tujuan, metode, dan kesimpulan dari invensi, sehingga dapat membantu dalam proses penilaian dan pencarian informasi oleh pihak ketiga.

2.1.2. Prior Art Search

Prior art sering digunakan dalam dunia paten untuk merujuk pada paten yang sudah ada. Penemuan-penemuan lama yang sudah ada ini, tidak dapat dipatenkan lagi. *Prior art* tidak hanya terdiri dari paten yang sudah ada, tetapi juga terdiri dari dokumen, benda, dan proses yang sudah dijual dan digunakan di masa lalu (Lin, 2023).

Penelusuran paten adalah suatu upaya melakukan pencarian atau penelusuran teknologi-teknologi terdahulu dalam bidang yang berdekatan maupun sama yang nantinya dijadikan *prior art* ataupun dokumen pembanding maupun pendukung (IPIndo, n.d.). Dapat

disimpulkan bahwa *Prior art search* adalah proses pencarian atau penelusuran penemuan-penemuan terdahulu.

2.1.3. Vektorisasi Teks

Vektorisasi adalah suatu proses mengubah kata-kata atau teks menjadi representasi vektor numerik sehingga dapat diproses oleh suatu model machine learning (Gandhi et al., 2021). Menurut Nguyen, secara umum vektorisasi dibagi menjadi 3 jenis utama, yaitu:

1. Berbasis Frekuensi

   Metode yang mengonversi teks menjadi representasi numerik berdasarkan jumlah kemunculan kata dalam dokumen. Metode ini tidak mempertimbagnkan urutan kata atau hubungan semantik antar kata. Contoh: *Bag of Words (BoW)* dan *TF-IDF.*

2. Berbasis Distribusi

   Metode yang mewakili kata atau dokumen berdasarkan distribusi kata-kata di sekitarnya. Metode ini menangkap makna kata berdasarkan konteksnya dalam korpus teks. Contoh: *Word2Vec, GloVe*, *FastText.*

3. Berbasis Transformer

   Metode yang menggunakan model *deep learning* untuk menghasilkan vektor teks berdimensi tinggi yang bergantung pada konteks kalimat secara keseluruhan. Metode ini menangkap makna kata secara kontekstual dengan mempertimbangkan hubungan dalam keseluruhan teks. Contoh: *BERT*, *Sentence-BERT*, *GPT Embedding*.

2.1.4. BERT

BERT (Bidirectional Encoder Representations from Transformers) adalah model yang diperkenalkan oleh Google pada tahun 2018, yang telah merevolusi pemrosesan bahasa alami (NLP) dengan kemampuan untuk memahami konteks kata dalam kalimat secara lebih mendalam. Model ini melampaui pendekatan sebelumnya dalam berbagai tugas pemodelan bahasa, berkat kemampuannya untuk menangkap makna kata yang bervariasi tergantung pada konteks penggunaannya. Dalam konteks paten, *BERT* menawarkan potensi besar untuk meningkatkan efisiensi pencarian dan analisis dokumen paten, yang memiliki struktur teks unik dan kompleks.

Google pada tahun 2020 mengeluarkan BERT yang sudah disesuaikan khusus untuk klasifikasi paten. Model ini dilatih dengan >100.000 data teks paten terdiri dari abstrak, claims, dan deskripsi. BERT for Patents dikembangkan dengan tokenisasi khusus yang lebih sesuai dengan teks paten. Misalnya, kata "prosthesis" akan dipisahkan menjadi beberapa token dalam tokenisasi standar, sedangkan tokenisasi khusus menjaga "prosthesis" sebagai satu token utuh. Selain itu, penelitian ini juga menguji validitas untuk sinonim yang dihasilkan oleh *BERT for Patents.* Disimpulkan bahwa *BERT for Patents* mampu menghasilkan sinonim relevan berdasarkan konteks.

Berdasarkan dokumentasi *official BERT for Patents,* berikut cara penggunaan model yang disediakan secara *open-source*:

1. Unduh Model

   Disediakan dalam 2 bentuk format, yaitu:

   - SavedModel, untuk penggunaan langsung dalam *framework* seperti TensorFlow.
   - Checkpoint, untuk pembaruan model atau fine-tuning.

2. Input Data

   Format input yang diterima antara lain:

   - input_ids: Token ID hasil tokenisasi teks.
   - input_mask: Masking untuk membedakan token teks asli (1) dan padding (0).
   - segment_ids: Menandai segmen input (contoh: bagian klaim atau abstrak).
   - mlm_ids (opsional): Token yang dimasukkan untuk tugas *masked language modeling.*

3. Output Data

   Data output antara lain:

   - cls_token: Representasi seluruh teks.
   - encoder_layer: Embeddings kata dari lapisan terakhir.
   - mlm_logits: Prediksi untuk token yang dimasking.

2.1.5. Cosine Similarity

Cosine similarity adalah ukuran yang digunakan untuk menentukan tingkat kesamaan antara dua vektor dalam ruang produk dalam. Ukuran ini ditentukan oleh cosinus sudut antara dua vektor, yang menunjukkan seberapa besar kedua vektor tersebut mengarah ke arah yang sama. Dalam konteks analisis teks, *cosine similarity* sering digunakan untuk mengukur kesamaan antar dokumen (Han et al., 2012).

Cosine similarity dihitung dengan rumus yang melibatkan norma Euclidean dari kedua vektor yang dibandingkan. Jika kita memiliki dua vektor y dan x, *cosine similarity* dapat dinyatakan dengan:

$$Cosine\ Similarity(y,\ x)\ =\ \frac{y\ .\ x}{||y||\ ||x||} \qquad (2.1)$$

Di mana y · x adalah hasil kali dot dari kedua vektor dan // y // serta // x // adalah norma Euclidean dari masing-masing vektor. Nilai *cosine similarity* berkisar antara -1 hingga 1, di mana 1 menunjukkan kesamaan maksimum (vektor sejajar), 0 menunjukkan ketidak berhubungan (vektor ortogonal), dan -1 menunjukkan arah yang berlawanan

Cosine similarity memiliki beberapa keuntungan, antara lain sifatnya yang tidak terpengaruh oleh skala, sehingga tetap konsisten meskipun nilai dalam vektor bervariasi secara

signifikan. Selain itu, kemudahan dalam perhitungan menjadikannya pilihan populer dalam aplikasi real-time dan dataset besar. Metrik ini juga sering digunakan dalam sistem rekomendasi dan pengelompokan dokumen (Miesle, 2023).

2.1.6. Approximate Nearest Neighbor (ANN) Algorithm

Approximate Nearest Neighbor (ANN) merupakan teknik dalam komputasi yang digunakan untuk mencari titik data terdekat dalam ruang vektor berdimensi tinggi secara efisien. *Approximate Nearest Neighbor (ANN)* adalah operasi dasar dan penting dalam domain basis data, machine learning, maupun computer vision. Hal ini karena pencarian yang tepat seperti kata kunci tidak efisien untuk ruang berdimensi tinggi (Li, et al., 2019).

2.1.7. Milvus

Milvus adalah sistem manajemen data vektor yang dirancang khusus untuk menangani data skala besar dengan efisiensi tinggi, terutama dalam aplikasi AI dan data science (*What Is Milvus | Milvus Documentation*, n.d.). Sistem ini mendukung pencarian kesamaan berbasis vektor (*vector similarity search*) dan memiliki kemampuan pengolahan data dinamis (*insert, delete, update*) secara cepat.

Menurut Wang et al. (2021), Milvus dioptimalkan untuk platform komputasi heterogen dengan CPU dan GPU modern serta mendukung berbagai jenis pencarian, termasuk attribute filtering dan *multi-vector query processing*. Salah satu pengoptimalan yang dilakukan Milvus adalah pencarian ANN. Dengan ANN, Milvus menyediakan pengalaman pencarian yang efisien (BaSic ANN Search | Milvus Documentation, n.d.).

2.1.8. Elasticsearch

Dalam penelitian Kathare et al. (2021), Elasticsearch dijelaskan sebagai mesin pencarian teks lengkap dan analitik yang bersifat open-source dan terdistribusi. Elasticsearch dibangun di atas Apache Lucene dan mampu menangani berbagai jenis data, termasuk data alfabetik, numerik, terstruktur, dan tidak terstruktur. Karena sifatnya yang terdistribusi, Elasticsearch menawarkan *high availability* dan *scalability*, sehingga dapat menangani beban kerja pencarian yang besar dan mengurangi latensi.

2.2. Landasan Teori

2.2.1. Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery (Srebrovic & Yonamine, 2020)

- Dalam dunia paten, *prior art search* yaitu proses untuk mengidentifikasi inovasi sebelumnya yang relevan merupakan tantangan besar. Solusi yang ada, adalah praktisi

menggunakan pencarian berbasis kata kunci. Metode ini memiliki keterbatasan dalam hal pemilihan istilah yang tepat. Hal ini semakin diperburuk dalam penggunaannya di paten, karena istilah baru sering kali digunakan dengan cara yang tidak konvensional, sehingga sulit untuk menemukan sinonim yang relevan. Oleh karena itu dibutuhkan algoritma yang dapat menghasilkan sinonim kontekstual yang lebih baik untuk mendukung pencarian terutama dalam hal kompleks seperti paten.

- Penelitian ini memperkenalkan model BERT yang dilatih khusus dengan teks paten. Beberapa langkah kunci metode yang digunakan adalah:
    1. Tokenisasi khusus teks paten, karena struktur bahasa dan istilah yang tidak umum.
    2. Pelatihan model dengan *hyperparameter* dengan fokus untuk memprediksi istilah yang disembunyikan dalam konteks paten.
    3. Generasi sinonim dengan langkah langkah sebagai berikut:
        a. Memilih kode CPC (*Cooperative Patent Classification).*
        b. Memilih istilah tertentu.
        c. Mengambil sejumlah dokumen paten yang mengandung istilah tersebut.
        d. Menghasilkan prediksi untuk istilah di tiap dokumen.
        e. Menghitung metrik agregat untuk menemukan istilah yang paling mungkin menjadi sinonim berdasarkan konteks.
- Penelitian ini menunjukkan bahwa model *BERT* dapat secara efektif menghasilkan sinonim kontekstual untuk istilah dalam dokumen paten. Hal ini dibuktikan dengan membandingkan prediksi sinonim untuk istilah yang sama di berbagai kode CPC, dan menemukan hasil model dapat membedakan penggunaan sinonim yang tepat berdasarkan konteks.
- Penelitian ini juga menyediakan versi *open-source* dari model *BERT* yang telah dilatih serta dokumentasinya.

2.2.2. The Use of PatentBERT for the Allocation of Patent in the UK IPO (McKenna et al., 2023)

- Kantor paten Inggris menghadapi tantangan dalam mengalokasikan aplikasi paten ke kelompok pemeriksa yang tepat. Proses ini sering kali dilakukan secara manual, hal ini menyebabkan sering kali ada kesalahan dan ketidakakuratan dalam pengklasifikasian. Dengan lebih dari 584.000 dokumen paten aplikasi paten yang harus dikelola, diperlukan sebuah alat otomasi yang dapat meningkatkan efisiensi dan akurasi alokasi paten berdasarkan *subclass* paten yang sesuai.

- Penelitian ini menggunakan model *PatentBERT*, yang merupakan varian dari *BERT* yang telah disesuaikan khusus untuk teks paten. Beberapa langkah metodologi yang digunakan adalah sebagai berikut:
    1. Pembersihan teks, teks paten dibersihkan dengan menghapus tanda baca dan mengubah huruf menjadi huruf kecil untuk mempertahankan konteks.
    2. *Embedding* teks, teks yang telah dibersihkan diubah menjadi bentuk vektor dengan menggunakan berbagai model termasuk *PatentBERT*.
    3. Klasifikasi, menggunakan klasifikasi *multilabel* dan *multiclass*, untuk menentukan *subclass* yang tepat untuk setiap aplikasi.
    4. Penggabungan hasil klasifikasi, hasil dari berbagai klasifikasi digabungkan untuk meningkatkan akurasi prediksi. Menggunakan algoritma *pruning* untuk menyederhanakan pohon perubahan simbol klasifikasi.
- Penelitian ini berhasil menunjukkan bahwa penggunaan *PatentBERT* secara signifikan meningkatkan akurasi alokasi aplikasi paten. Beberapa temuan kunci meliputi:
    1. Model berhasil mencapai probabilitas 96.1% untuk mengembalikan setidaknya satu simbol yang benar dan 93.6% untuk simbol utama.
    2. Menggabungkan hasil dari berbagai model meningkatkan akurasi prediksi dibandingkan dengan menggunakan model tunggal.
    3. Akurasi tertinggi diperoleh dari dokumen abstrak jika tunggal, dan terbaik adalah gabungan dari deskripsi dan abstrak.

2.2.3. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (Reimers & Gurevych, 2019)

- Model BERT dan RoBERTa menunjukkan performa yang sangat baik dalam tugas-tugas regresi pasangan kalimat, seperti pengukuran kesamaan teks semantik. Namun pendekatan ini membutuhkan kedua kalimat untuk diproses secara bersamaan, ini menghasilkan overhead komputasi yang besar. Hal ini membuat BERT kurang cocok untuk pencarian kesamaan semantik dan pengelompokan.
- Memperkenalkan model Sentence-BERT, sebuah modifikasi dari model BERT yang menggunakan struktur jaringan siamese dan triplet untuk menghasilkan embedding kalimat yang bermakna semantik. Langkah metode yang digunakan dalam penelitian ini antara lain:
    1. Modifikasi jaringan, mengadaptasi arsitektur BERT menjadi jaringan siamese dan triplet untuk menghasilkan embedding kalimat.

2. Penghitungan kesamaan, dengan menggunakan cosine similarity, memungkinkan perhitungan kesamaan yang cepat.
3. Evaluasi model, dengan pengujian di berbagai tugas STS umum untuk menilai kinerja.

- Penelitian ini berhasil menunjukkan bahwa *Sentence-BERT* secara signifikan mengurangi waktu yang diperlukan untuk menemukan pasangan kalimat yang serupa dari 65 jam menjadi sekitar 5 detik, sambil tetap mempertahankan akurasi yang sebanding dengan *BERT*. Beberapa temuan kunci lainnya meliputi:
    1. *Cosine Similarity* menjadi pengukuran jarak vektor yang cepat dan efektif.
    2. Penggunaan *Sentence-BERT* memungkinkan pencarian kesamaan semantik dilakukan dengan jauh lebih cepat tanpa mengorbankan akurasi.

2.2.4. *Prior Art Search and Reranking for Generated Patent Text* (Lee & Hsiang, 2021)

- Model bahasa generatif seperti GPT-2 telah menunjukkan hasil yang mengesankan dalam menghasilkan teks. Namun, tantangan utama yang dihadapi adalah bagaimana menentukan dari mana teks yang dihasilkan berasal. Penelitian ini bertujuan untuk mengidentifikasi teks paten paling mirip dalam data pelatihan GPT-2 menggunakan metode pencarian prior art dan reranking.
- Untuk meningkatkan akurasi pencarian prior art, penelitian ini mengembangkan pendekatan reranking yang terdiri dari beberapa langkah:
    1. Pencarian awal dengan BM25, yaitu metode berbasis bag-of-words untuk menemukan teks paten yang paling mirip dengan teks yang dihasilkan GPT-2.
    2. Konversi ke embedding BERT, di mana hasil pencarian dikonversi menjadi embedding menggunakan model BERT yang telah dilatih dengan teks paten.
    3. Peringkat ulang berdasarkan kesamaan embedding, dengan menggunakan cosine similarity untuk mengurutkan kembali hasil pencarian berdasarkan kesamaan semantik.
- Hasil penelitian menunjukkan bahwa pendekatan reranking dengan kombinasi BM25 dan embedding BERT memberikan hasil yang lebih baik dibandingkan hanya menggunakan salah satu metode. Beberapa temuan utama dari penelitian ini meliputi:
    1. BM25 memiliki cakupan luas, tetapi akurasinya lebih rendah dibandingkan embedding berbasis BERT.
    2. Embedding berbasis BERT memiliki akurasi tinggi, tetapi cakupannya lebih sempit.

3. Kombinasi BM25 dan reranking berbasis embedding meningkatkan relevansi hasil pencarian, dengan menyeimbangkan antara cakupan dan akurasi.