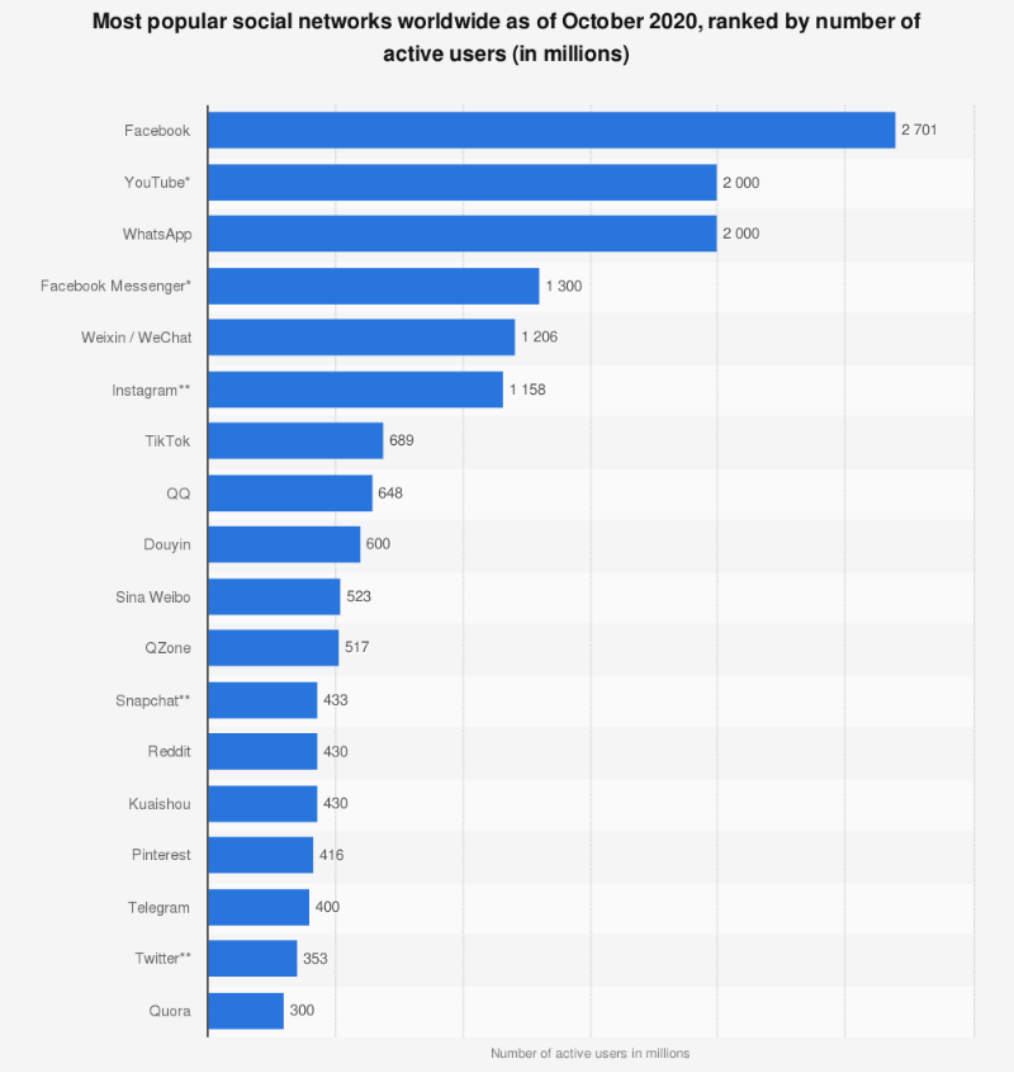
BAB II

LANDASAN TEORI

2.1. Media Sosial

Media Sosial atau juga dikenal sebagai *Social Network Sites* (SNSs) merupakan salah satu bentuk komunikasi yang dimediasi oleh sistem komputer yang memungkinkan seorang individu/kelompok untuk: (1) membangun profil dan informasi yang terkait dengan pribadinya sekaligus mengenali profil serta informasi pribadi dari individu/kelompok lain sehingga mereka bisa terkoneksi satu sama lain; (2) membagikan maupun menikmati konten seperti ide, kreasi, kesukaan, pendapat, informasi, dll dalam bentuk foto, video, tautan, teks, dll untuk/dari sesama individu atau kelompok; (3) berinteraksi dan saling memberikan reaksi terhadap konten masing-masing individu/kelompok – menciptakan jaringan sosial "virtual" yang melahirkan kehidupan sosial baru bagi manusia (Kietzmann et al., 2011; Obar & Wildman, 2015).

Berbagai *platform* media sosial seperti Facebook, Twitter, Instagram, Youtube, Snapchat, TikTok, dll memungkinkan pengguna untuk terhubung satu dengan yang lainnya dan saling berbagi informasi cukup dengan melakukan *click-and-tap* pada perangkat komputer atau *smartphone* mereka (Stewart, 2013). Gambar 2.1 menunjukkan peringkat *platform* media sosial yang paling sering digunakan oleh pengguna di dunia hingga Oktober 2020.



Gambar 2.1. Peringkat Platform Media Sosial Yang Paling Sering Digunakan Oleh Pengguna Di Dunia

Sumber: (Statista, 2020)

Media sosial ini di satu sisi memiliki keuntungan, yaitu meningkatkan kualitas konektivitas dari hubungan sosial yang sudah ada, bahkan bisa membentuk koneksi baru di antara orang-orang yang belum saling mengenal sebelumnya (Steijn & Schouten, 2013). Di sisi lain, kebebasan dan kemudahan dalam mengungkapkan sesuatu yang ditawarkan oleh media sosial ini menimbulkan berbagai isu, seperti

privasi pengguna, hak mengungkapkan pendapat, hak milik intelektual, hingga kekerasan atau agresi secara *online* (Obar & Wildman, 2015). Konten pelecehan, agresif, serta antisosial yang semakin meluas, merupakan masalah yang terus-menerus terjadi di media sosial (Liu et al., 2018), dan sebagian besar korbannya adalah generasi muda yang secara psikologis masih sangat rentan apabila mengalami serangan yang langsung menyentuh mental mereka – tidak jarang yang akhirnya sampai harus berakhir dengan nyawa (Livingstone et al., 2014; Wong-Lo & Bullock, 2011).

2.2. Kekerasan atau Pelecehan *Online*

Kekerasan atau pelecehan *online* atau dikenal dengan istilah-istilah lain, seperti *cyber*/*online/digital violence*, *cyber*/*online/digital abuse*, *cyber*/*online/digital harassment*, atau *cyber*/*online/digital aggression* merupakan tindakan penyalahgunaan dunia maya, khususnya pada media sosial untuk menciptakan pesan atau komentar yang mengusik, merendahkan, membenci, mengancam, hingga melecehkan secara seksual, baik berulang maupun tidak kepada korban (eSafety Commissioner, 2020). Hal ini merupakan salah satu bentuk bentuk pelecehan verbal atau emosional yang dilakukan secara daring (Love is Respect, 2014).

Kekerasan atau pelecahan *online* yang paling umum terjadi adalah intimidasi dunia maya atau lebih dikenal dengan *cyberbullying* dan ujaran kebencian atau lebih dikenal dengan *hate speech* (Ditch The Label, 2015). Masih terdapat beberapa tipe lain, misalnya Kekerasan Berbasis Gender Online (KBGO), yaitu salah bentuk tindakan yang memiliki maksud untuk melecehkan korban berdasarkan gender atau

seksual, khususnya sekarang-sekarang ini yang sangat mengkhawatirkan adalah bagi kaum perempuan (SAFEnet, 2019). Adapun bentuk-bentuk turunan dari tindakan *cyberbullying* (Kansara & Shekokar, 2015) juga perlu menjadi perhatian, di antaranya (1) *Flaming* (mengirim pesan kasar atau vulgar), (2) *Outing* (memposting hingga memanipulasi informasi pribadi/privasi seseorang tanpa persetujuan pemiliknya), (3) *Harassment* atau pelecehan (berulang kali mengirim pesan ofensif ke satu orang), (4) *Exclusion* (pengasingan seseorang oleh kelompok), (5) *Cyberstalking* (meneror akun seseorang dengan mengirimkan pesan yang mengancam dan mengintimidasi), (6) *Denigration/Defamation* atau fitnah, dan (7) *Impersonation* atau peniruan identitas. Bentuk-bentuk tindakan ini merupakan salah satu bagian dari kejahatan siber/dunia maya (Wikipedia, 2020).

Pada dasarnya, dampak dari kekerasan atau pelecehan *online* yang dirasakan oleh korban tidak berbeda jauh dengan korban yang mengalami kekerasan/pelecehan konvensional atau fisik. Korban akan mengalami hal-hal seperti depresi, kepercayaan diri menurun, merasa tidak berdaya, kecemasan sosial, serta perasaan asing yang begitu mendalam. Namun, mengingat karakteristik dari kekerasan/pelecehan *online* yang di mana pelaku dapat menyembunyikan identitasnya (anonim), dunia maya yang tak terbatas jumlah pengamatnya, penyebaran konten yang dapat dilakukan dengan begitu cepat, serta jejak digital yang tidak pernah terhapus, membuat kekerasan atau pelecehan secara digital memberikan luka yang lebih mendalam serta berkepanjangan bagi korbannya (Wong-Lo & Bullock, 2011).

2.3. Sistem Deteksi Otomatis Kekerasan atau Pelecehan *Online*

Mempertimbangkan dampak yang ada, upaya untuk mendeteksi kekerasan/pelecahan *online* sejak dini menjadi kunci penting bagi kesejahteraan mental khususnya generasi muda (Van Hee et al., 2018). Namun, dengan arus konten yang begitu besar dalam rentang waktu yang singkat, tidak memungkinkan bagi moderator selaku yang melakukan pengawasan terhadap aktivitas media sosial untuk memantau dan mengevaluasi setiap konten yang ada secara manual satu-persatu (Van Hee et al., 2018). Untuk mengatasi masalah ini, diperlukan sistem cerdas atau *Intelligent System* yang terautomisasi (*Automated System*). Sistem ini memproses informasi ini dengan cepat dan secara otomatis memberi sinyal potensi ancaman. Dengan cara ini, moderator dapat merespons dengan cepat dan mencegah meningkatnya situasi yang mengancam (Van Hee et al., 2018). Selain itu, sistem serupa juga dapat digunakan sebagai tindakan preventif sebelum pelaku melakukan aksinya, di mana sistem media sosial akan langsung melakukan pembatalan *posting* ketika sistem mendeteksi unsur kekerasan dari pesan atau komentar yang ingin dikirim oleh pelaku (Van Royen et al., 2017).

Automated Harassment Detection secara garis besar didefinisikan sebagai sistem cerdas untuk: (1) mengklasifikasikan suatu pesan yang mengandung konten kekerasan atau tidak berdasarkan fitur teks, atau mengambil kesimpulan berdasarkan fitur tingkat lanjut, seperti (2) fitur pengguna media sosial dan (3) fitur jaringan sosial (*social network*) (Emmery et al., 2019). Fitur teks adalah yang paling sering digunakan untuk mendeteksi aktivitas kekerasan di media sosial, tetapi beberapa penelitian juga melakukan analisis memanfaatkan fitur pengguna, seperti

usia dan jenis kelamin (Chatzakou et al., 2017, 2019; Dadvar et al., 2012, 2013; Hosseinmardi et al., 2016). Fitur jaringan sosial yang terbentuk dari pengguna juga digunakan untuk meningkatkan keabsahan dari detektor, seperti daftar pengguna lain yang terkoneksi dengannya dan interaksinya dengan pengguna lain (Al-Garadi et al., 2016; Chatzakou et al., 2017; Di Capua et al., 2016; Huang et al., 2014; V. K. Singh et al., 2016). Terdapat juga penelitian yang menggunakan fitur berbasis media, seperti gambar atau video untuk meningkatkan performa klasifikasi (Hosseinmardi et al., 2016).

Dalam melakukan deteksi kekerasan/pelecehan otomatis ini, setidaknya ada dua tugas umum yang biasa dilakukan, yaitu (1) *Binary classification* di mana model akan mengklasifikasikan objek menjadi dua kelas, yaitu apakah suatu konten mengandung kekerasan/pelecehan atau tidak – ini yang paling umum dilakukan; (2) *Fine-grained classification* yang mampu mengkategorikan konten lebih dari dua kelas, misalnya berdasarkan konteks atau peran dari pelaku yang terlibat (Van Hee et al., 2015).

Automated Harassment Detection ini dapat disematkan dalam *tool* digital yang nantinya bertujuan untuk mencegah, mengintervensi, atau menangani tindakan kekerasan/pelecehan *online* (Rosa et al., 2019). Sistem ini membutuhkan semacam penggolong atau *classifier* yang akurat sehingga dapat menentukan dengan baik mana aktivitas yang merupakan kekerasan/pelecehan dan mana yang bukan, sehingga insiden ini dapat diminimalisir (Rosa et al., 2019).

Berdasarkan penelitian yang sudah dilakukan sebelumnya, kaum remaja umumnya mendukung sistem ini, asalkan strategi tindak lanjut dari tindakan *cyberbullying* dirumuskan secara efektif, dan privasi korban serta otonomi moderator dan dijamin (Van Royen et al., 2015).

2.4. *Natural Language Processing*

Natural Language Processing (NLP) merupakan cabang ilmu dari Kecerdasan Artifisial, yang ditujukan untuk membuat mesin komputer memahami pernyataan atau kata-kata yang ditulis dalam bahasa natural manusia (Khurana et al., 2017). Pekerjaan ini tidak cukup hanya dilakukan dengan proses komputasi berbasis *rule* dikarenakan bahasa merupakan sesuatu yang sangat luas, serta manusia dalam percakapan sehari-hari jarang sekali menggunakan bahasa formal yang memiliki tata bahasa yang terstruktur dengan baik, melainkan menggunakan bahasa yang umum dan dimengerti semua orang (Bengfort et al., 2018).

Hal ini membutuhkan banyak pekerjaan yang terus berkembang hingga hari ini. Tetapi dua aktivitas besar yang sangat berpengaruh pada metode NLP sebelum pembangunan model NLP itu sendiri, yaitu prapemrosesan teks dan rekayasa/ekstraksi fitur dari data teks.

Tahap prapemrosesan atau *preprocessing* menjadi hal yang krusial pada kerangka kerja pembelajaran mesin apapun, salah satunya adalah untuk NLP. Hal ini bertujuan untuk mempersiapkan data teks agar lebih siap dan “layak guna” ketika dimasukkan sebagai input model. Teknik-teknik umum yang biasa digunakan pada tahap ini meliputi:

1) *Tokenization*: memecah bentuk teks menjadi satuan yang lebih kecil, misalnya dari paragraf menjadi satuan kalimat, atau dari kalimat menjadi satuan kata.
2) *Case Folding*: mengubah seluruh karakter huruf menjadi huruf kecil. Tahap ini bersifat alternatif tergantung kebutuhan analisis.
3) *Stop words removal*: menghilangkan kata-kata yang sering muncul, seperti dan, atau, ke, di, dari, dan lain-lain.
4) *Stemming*: proses mengubah setiap kata yang ada menjadi kata dasar.
5) *Lemmatization*: jika *Stemming* hanya mengubah setiap kata ke bentuk dasar dengan mengeliminasi awalan dan imbuhan kata, *Lemmatization* mampu mengubah variasi kata yang berbeda tetapi memiliki satu kata dasar standar yang sama.

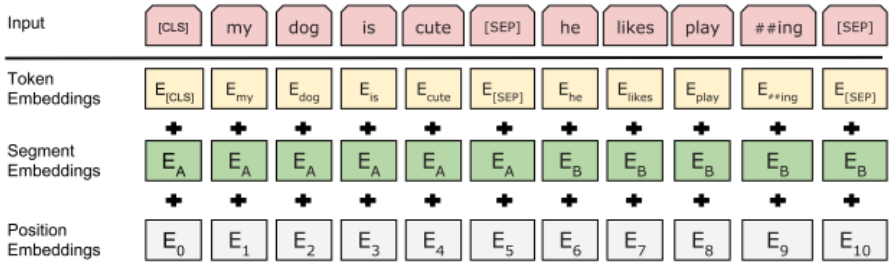
Jika pekerjaan prapemrosesan teks sudah menjadi pekerjaan umum dan fundamental bagi setiap tugas NLP, teknik rekayasa atau ekstraksi fitur teks merupakan pekerjaan yang terus berkembang hingga saat ini. Tahap ini bertujuan untuk mengubah teks yang tidak dimengerti oleh mesin menjadi suatu data yang dapat dipelajari dan dipahami. Pekerjaan ini dimulai dari mengekstrak fitur bersifat leksikal dan sintaksis, seperti teknik *Bag of Words* atau disingkat BoW, *N-grams*, TF-IDF, sampai kepada diperkenalkan teknik yang disebut *Word Embedding* yang bisa menangkap fitur semantik dari suatu teks, seperti Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017; Joulin, 2016), dan lain-lain.

2.5. *Bidirectional Encoder Representations from Transformers* (BERT)

Model *Word Embedding* pada NLP memiliki satu permasalahan, yaitu model tidak dapat menangkap makna bersifat polisemi, yaitu kata dalam bentuk yang sama tetapi memiliki arti yang berbeda, misalnya saja kata "bisa" yang berarti "dapat" tetapi juga bisa berarti "racun yang dihasilkan oleh binatang ular" (Young et al., 2018). Hal ini dikarenakan *Word Embedding* masih memperlakukan kata secara individual, walaupun sudah bisa melihat hubungan antara individual-individual kata lainnya. Tantangan selanjutnya pada NLP yaitu bagaimana model tidak hanya bisa memaknai suatu kata, tetapi juga bisa mengenali kata dalam konteksnya, misalnya saja ingin memprediksi kata selanjutnya dari suatu kalimat. Isu ini dapat diatasi dengan model seperti *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM) (Chung et al., 2014; Palangi et al., 2016; Wu et al., 2016; Young et al., 2018). Kemudian, sampai di mana kita diperkenalkan dengan *Transformers*, yaitu model atensi (*Attention Model*) yang melampaui model RNN atau LSTM karena model atensi ini memungkinkan lapisan *decoder* untuk mengenali langsung *hidden state* dari *encoder*-nya sendiri, sehingga setiap *decoder* yang ada bisa langsung mengerjakan bagiannya secara paralel dibandingkan arsitektur *neural network* yang bersifat sekuensial seperti RNN dan LSTM yang dimana setiap *decoder* harus menunggu *hidden state* dari *decoder* sebelumnya (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017). Model inilah yang kemudian diadopsi oleh **Google BERT** (Devlin et al., 2019).

Gambar 2.2 menunjukkan proses transformasi dari tokenisasi setiap kata menjadi *word embedding* dalam bentuk vektor yang akan menjadi input dari lapisan

Encoder dari *Transformers* milik BERT. Vektor input dari model *Transformer* merupakan penjumlahan dari vektor hasil *Token Embeddings*, *Segment Embeddings*, dan *Position Embeddings*. Vektor ini membuat model *Transformer* akan memahami posisi dari masing-masing kata yang walaupun memiliki bentuk yang sama, tetapi bisa jadi memiliki makna kontekstual yang berbeda (Devlin et al., 2019).

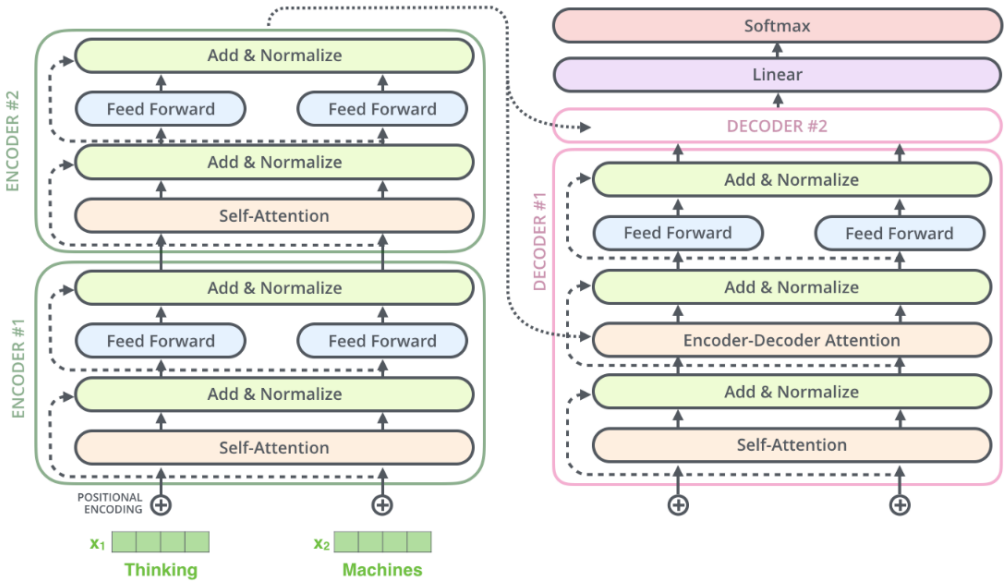


Gambar 2.2. Proses *Word Embedding* pada Arsitektur BERT

Sumber: (Devlin et al., 2019)

Lapisan *Encoder* dari *Transformers* dilatih oleh BERT secara dua arah (*bidirectional*), dalam arti tidak hanya membaca sebuah susunan teks berdasarkan arah (kiri-ke-kanan atau kanan-ke-kiri), tetapi model yang disebut *Masked Language Model* (MLM) ini dapat belajar konteks dari sebuah kata berdasarkan kata-kata di sekelilingnya, tidak hanya berdasarkan satu arah seperti pada model-model serumpun dengan BERT, seperti ELMo dan GPT (Devlin et al., 2019). Hasil output dari model BERT ini adalah distribusi probabilitas hasil perhitungan fungsi *Softmax*, yaitu pemodelan bahasa yang bersifat *unsupervised*, tanpa diberi label tetapi memahami memahami secara kontekstual dari kumpulan teks yang ada. Gambar 2.3 menunjukkan arsitektur *Transformers* secara keseluruhan yang terdiri

dari *Encoder* dan *Decoder*. BERT hanya akan memanfaatkan bagian *Encoder* dari *Transformers*.



Gambar 2.3. Arsitektur *Transformer*s

Sumber: (Alammar, 2018)

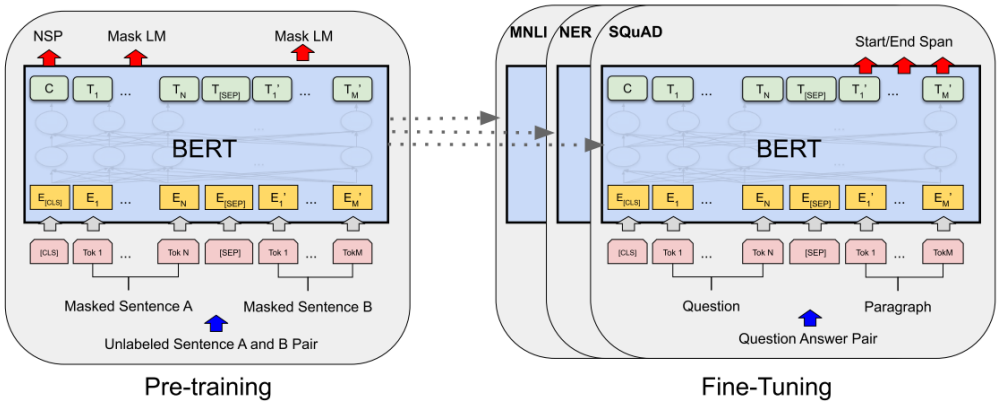
Hal yang menjadi keunggulan dari BERT adalah kemampuannya dalam melakukan *Transfer Learning*, di mana hasil output dari model BERT ini telah menyediakan model *pre-trained* yang bisa diadopsi untuk berbagai tugas NLP sekaligus, salah satu tugas khususnya adalah klasifikasi teks. Proses *Transfer Learning* ini diilustrasikan pada Gambar 2.4.

Dalam klasifikasi teks, BERT menggunakan satu token unik yang ditandai dengan `[CLS]` pada bagian awal dari lapisan *embedding* yang berisikan bobot perhitungan representasi teks (**h**) dikalikan bobot tersembunyi lainnya untuk melatih model BERT dalam melakukan tugas spesifik tertentu (*W*). Kemudian untuk mengklasifikasikan teks, BERT menambahkan perhitungan *Softmax* pada

lapisan *Encoder* terakhirnya untuk menghitung probabilitas label *c* (Devlin et al., 2019). Rumus fugsi *Softmax* secara umum dapat dilihat pada Rumus 2.1.

$$p(c|\boldsymbol{h}) = softmax(W\boldsymbol{h})$$

Rumus 2.1. Rumus Perhitungan Klasifikasi Teks pada BERT



Gambar 2.4. Arsitektur BERT

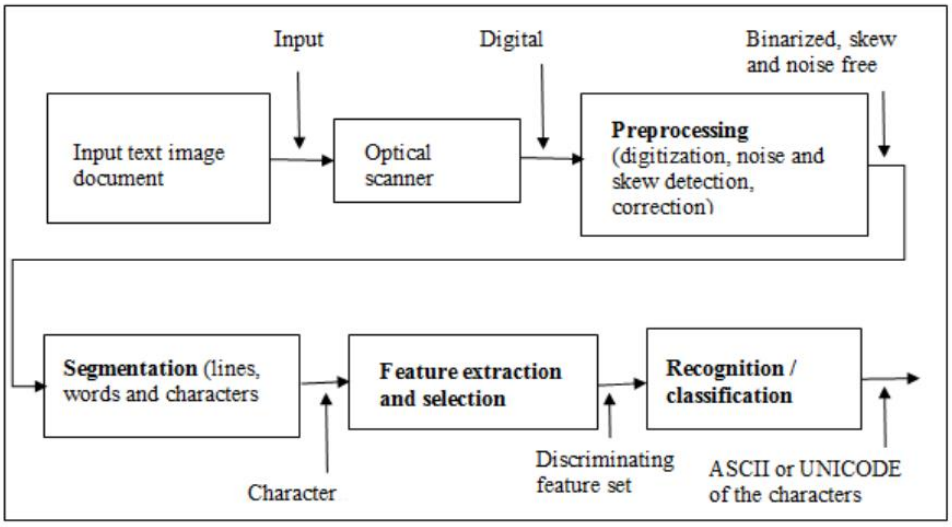
Sumber: (Devlin et al., 2019)

Gambar 2.4 menunjukkan arsitektur keseluruhan dari BERT. BERT sangat mahal dalam hal komputasi pada tahap *pre-training*, tetapi akan sangat optimal ketika sudah sampai pada tugas hilir (*downstream task*) pada tahap *fine tuning* (Salminen et al., 2020). Oleh karena itu Google sudah melakukan *pre-training* dengan sumber daya komputasi yang dimiliki dan para pengembang dapat memanfaatkannya cukup dengan melakukan *fine-tuning*.

2.6. *Optical Character Recogntion*

Optical Character Recognition (OCR) merupakan kemampuan proses komputasi yang memungkinkan mesin (komputer) untuk memproses, mengenali,

dan mengkonversi teks yang direpresentasikan dalam bentuk gambar digital menjadi teks yang kemudian ditampilkan kembali oleh komputer sebagai teks yang dapat disunting (S. Singh, 2013). Gambar 2.5 menunjukkan mekanisme OCR secara umum.



Gambar 2.5. Diagram Sistem OCR Secara Umum

Sumber: (Das et al., 2015)

2.7. Flask

Flask merupakan *micro framework* yang sekarang ini populer digunakan untuk pengembangan aplikasi web berbasis bahasa pemrograman Python. Disebut mikro bukan berarti aplikasi harus berada pada satu file Python, bukan juga Flask kekurangan fungsionalitas. Justru Flask menawarkan fleksibilitas dan ekstensibilitas yang menarik bagi pengembangnya. Misalnya saja Flask tidak memiliki lapisan abstraksi Database, tetapi dengan menggunakan Flask, kita bisa

menghubungkan aplikasi kita ke berbagai produk Database yang sudah ada sesuai dengan keinginan kita. (Flask, 2010)

Flask berintegrasi dengan berbagai *micro services* lainnya, dua di antaranya yang merupakan komponen utama dari Flask adalah Werkzeug dan Jinja. Werkzeug merupakan *toolkit* untuk *Web Server Gateway Interface* (WSGI) yang memungkinkan objek pemrograman Python untuk berinteraksi dengan *web server*, seperti melakukan *request*, *response*, dan utilitas lainnya. Jinja merupakan *template engine* yang digunakan Flask untuk memungkinkan interaksi antara halaman web dengan objek pemrograman Python.