**ANALISIS SENTIMEN REVIEW HOTEL BERBAHASA INDONESIA MENGGUNAKAN BERT**



**Oleh:**

**Team KaSur**

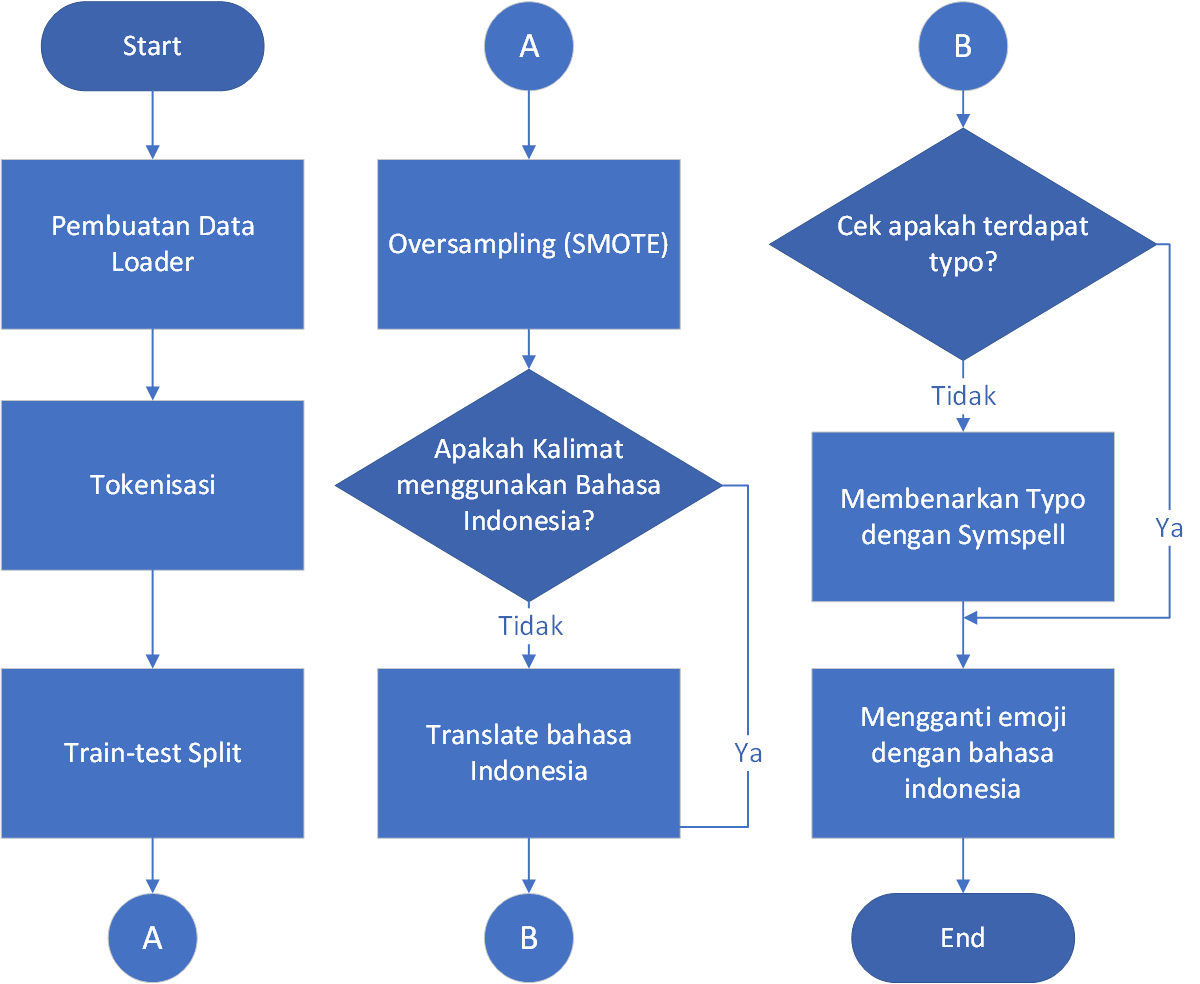
**(Syauqi Rahmat Perdana & Ahmad Hanif)**

**Latar Belakang**

Pada laporan kali ini akan dilakukan analisis sentimen dari review pelanggan. Objek yang dianalisis adalah hotel. Review berasal dari berbagai hotel namun tidak dijelaskan secara spesifik nama dari tiap hotelnya. Analisis sentimen yang akan dilakukan adalah memprediksi review mana yang memiliki sentimen baik dan review mana yang memiliki review buruk.

**Data Preprocessing & Feature Engineering**

Data yang diberikan berupa data teks sehingga preprocessing yang dilakukan khusus untuk data teks. Tahap data preprocessing tentu saja tergantung pada jenis data dan juga model yang nanti akan digunakan. Pada kasus ini akan digunakan transformer sebgai model yang digunakan. Oleh karena itu dilakukan tahap preprocessing data sebagai berikut:



Gambar 1 – Flowchart Preprocess Data

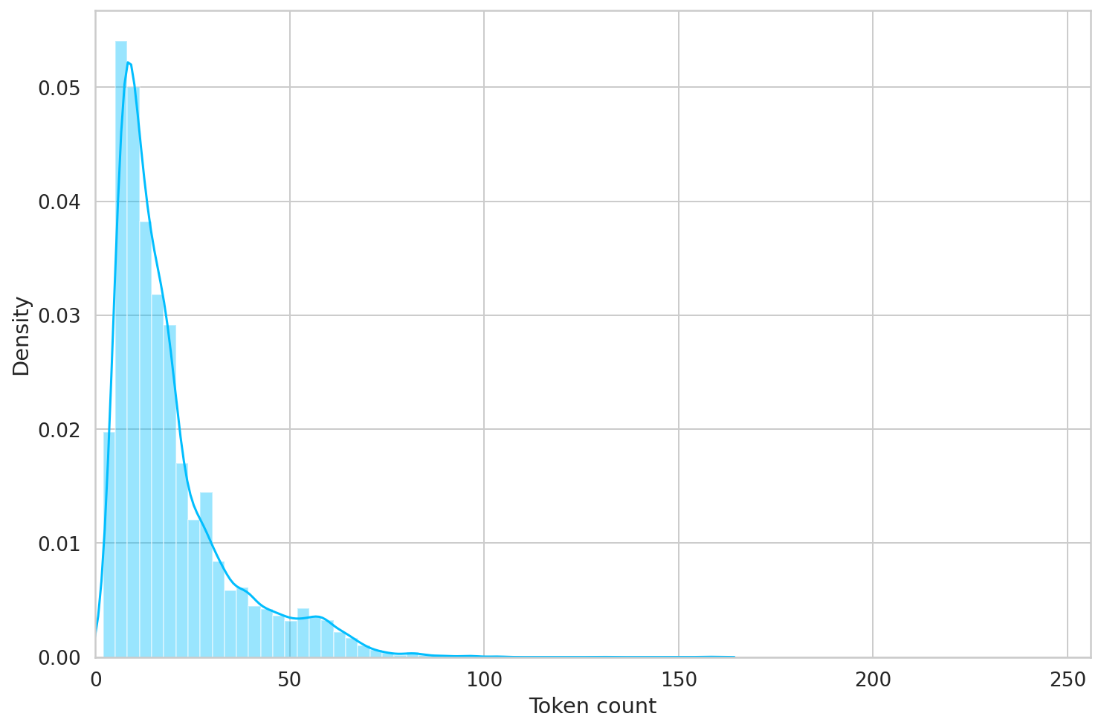
Untuk dapat melatih model transformer atau model lainnya yang berbasis deep learning maka model akan dilatih secara *batch*. Oleh karena itu, Langkah awal yang dilakukan adalah dengan membuat data loader yang dapat menginput data pada model secara *batch*. Langkah ini dilakukan dengan membuat *generator* pada python.

Langkah selanjutnya adalah melakukan tokenisasi sehingga tiap kata pada kalimat akan terpisah. Tokenisasi biasanya dilakukan dengan memisahkan tiap elemen pada kalimat bedasarkan spasi. Hanya saja, terkadang terdapat beberapa kata yang terpisah oleh spasi, misalkan saja “terima kasih”. Oleh karena itu langkah tokenisasi dilakukan dengan menggunakan BertTokenizer dengan pretrained model indobert-base-p1.

Kemudia dilakukan pemisahan data menjadi *training* dan *eval*. Data *training* merupakan data yang digunakan untuk melatih model sedangkan data *eval* adalah data yang digunakan untuk mengevaluasi performansi dari model. Setelah itu, dilakukan *oversampling* dengan menggunakan SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). *Oversampling* dilakukan karena data *imbalanced*, dalam kasus ini kebanyakan review merupakan review yang negatif. Untuk menghindari terbuatnya model yang *overfit* terhadap mayoritas maka dilakukan *oversampling*.

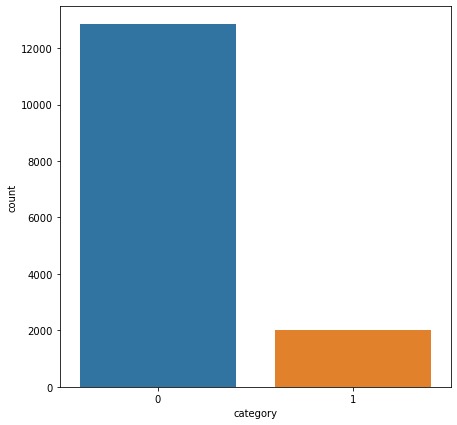
Beberapa review apabila kita cek secara manual maka akan ditemukan beberapa yang menggunakan Bahasa inggris. Berhubung model yang digunakan menggunakan *word embedding* berbahasa Indonesia maka Bahasa asing perlu di-*translate* ke Bahasa Indonesia. Pada kasus ini digunakan API google translate untuk mendeteksi sekaligus menerjemahkan Bahasa asing ke Bahasa Indonesia.

Typo sebenarnya sudah bisa di-*handle* oleh Bert karena memang sejatinya model ini dilatih dengan menggunakan data dengan Bahasa yang non-formal. Akan tetapi bisa saja beberapa typo tidak bisa di-*handle* oleh Bert karena tidak terdapat di data *training*. Oleh karena itu, agar dapat memastikan typo tidak ada, digunakan library symspell dengan kamus wordnet Bahasa Indonesia. Langkah terakhir adalah menerjemahkan berbagai emoji kedalam teks bukan kode. Hal ini mengingat kemungkinan besar *word embedding* yang digunakan tidak bisa menerjemahkan emoji secara otomatis.



**Explarotary Data Analysis**

Dapat dilihat bahwasannya kebanyakan sentiment pada data diklasifikasikan sebagai sentimen yang negatif. Lebih spesifik, sebanyak 86.42% dari data memiliki sentimen negatif sedangkan 13.58% sisanya diklasifikasikan memiliki sentiment yang positif.



Gambar 2 – Sebaran Sentimen

Jika dilihat dari sebaran kata, kebanyakan berbicara tentang kebersihan dari kamar terutama kebersihan dari kamar mandi. Kemudian, beberapa juga berbicara soal ketidak sesuaian dengan deskripsi serta mengenai air dan juga ruangan yang panas.



Gambar 3 – Wordcloud seluruh kata

Dari sebaran kata *review* yang positif dapat dilihat bahwasannya kebanyakan review positif diberikan karena kebersihan dari kamar, lokasi, serta pelayanannya. Hanya saja yang menarik adalah kata “tapi” juga cukup popular. Jika kita melihat review-review positif ini secara lebih lanjut ternyata pada beberapa review yang memberikan respon positif tapi kemudian mengkritik masih diberikan review yang positif. Misalnya “hotelnya bagus tapi sayang pelayanannya buruk”.



Gambar 4 – Wordcloud kata positif

Untuk sebaran kata *review* yang positif *insight* yang diberikan cukup mirip dengan sebaran seluruh kata mengingat kebanyakan dari *review* memang negatif sehingga wajar jika sebarannya mirip. Secara umum kebanyakan protes karena kebersihan dan ketidaksesuaian dengan yang dijanjikan atau diiklankan. Masalah yang berkaitan dengan *customer service* sepertinya tidak terlalu banyak. Hal ini menandakan kebanyakan hotel yang di *review* tidak memberikan fasilitas yang berkualitas.

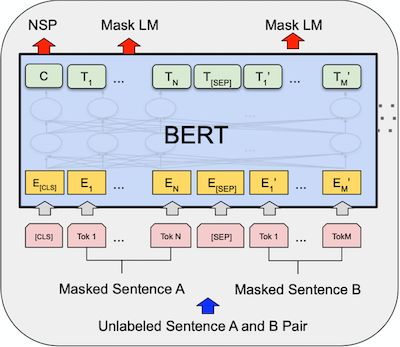


Gambar 5 – Wordcloud kata negatif

**Modelling**

Tahap terakhir akan dilakukan modelling untuk memprediksi *review* mana yang positif dan *review* mana yang negatif. Untuk melakukan hal ini dilakukan dua buah tahap. Tahap pertama adalah melakukan *word embedding* atau menerjemahkan kata-kata menjadi vektor yang terkait sedangkan tahap yang kedua adalah mencari fungsi matematis yang menghubungkan vektor dengan probabilitas vektor tersebut menjadi positif.

Pada kasus ini tahap embedding diakukan dengan menggunakan BERT. Secara sederhana BERT merupakan model yang sudah dilatih untuk memprediksi kata-kata yang di-masking contohnya “saya [mask] bola” maka bert akan memprediksi [mask] sebagai suka. Keunggulan dari BERT adalah model ini mengerti konteks dan juga dapat menghandle *word* yang tidak diketahui dengan cukup baik.



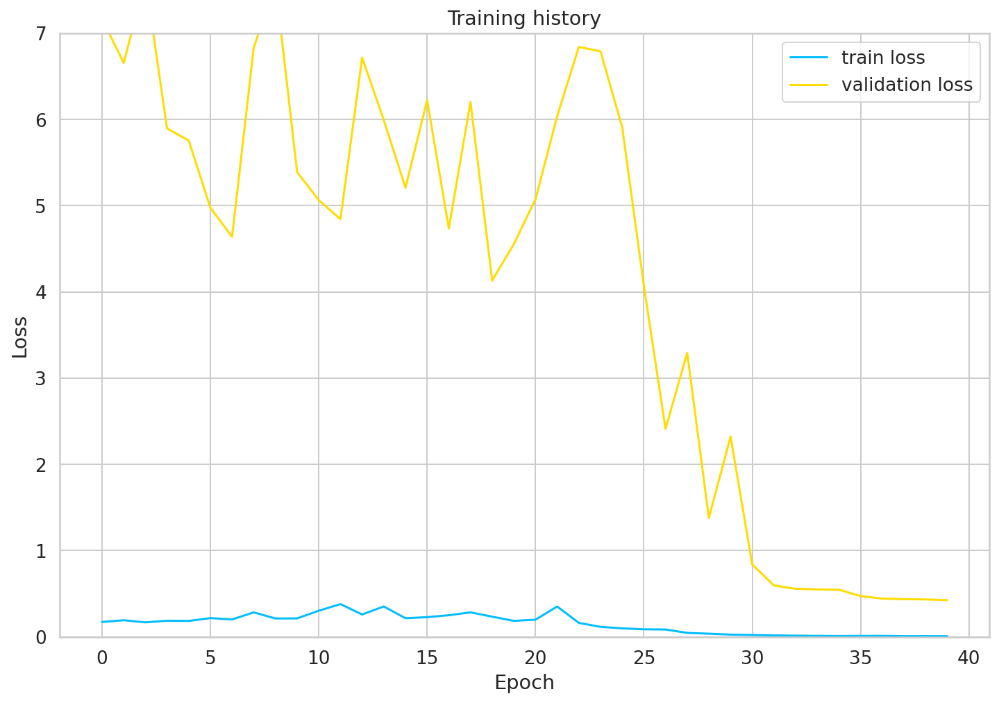
Gambar 6 – Ilustrasi BERT

Untuk tahap pembuatan fungsi matematis dilakukan *fine-tuning* terhadap Bert yaitu dengan menambahkan kalimat-kalimat yang terdapat pada *training* data beserta labelnya untuk mengkostumisasi *embedding* agar dapat menjadi vektor kalimat yang baik dalam menyelesaikan kasus yang dihadapi. Di depan embedding dari Bert ditambahkan layer *dropout* untuk menghindari *overfitting* serta ditambahkan juga satu layer softmax untuk mengubah output menjadi bentuk probability.

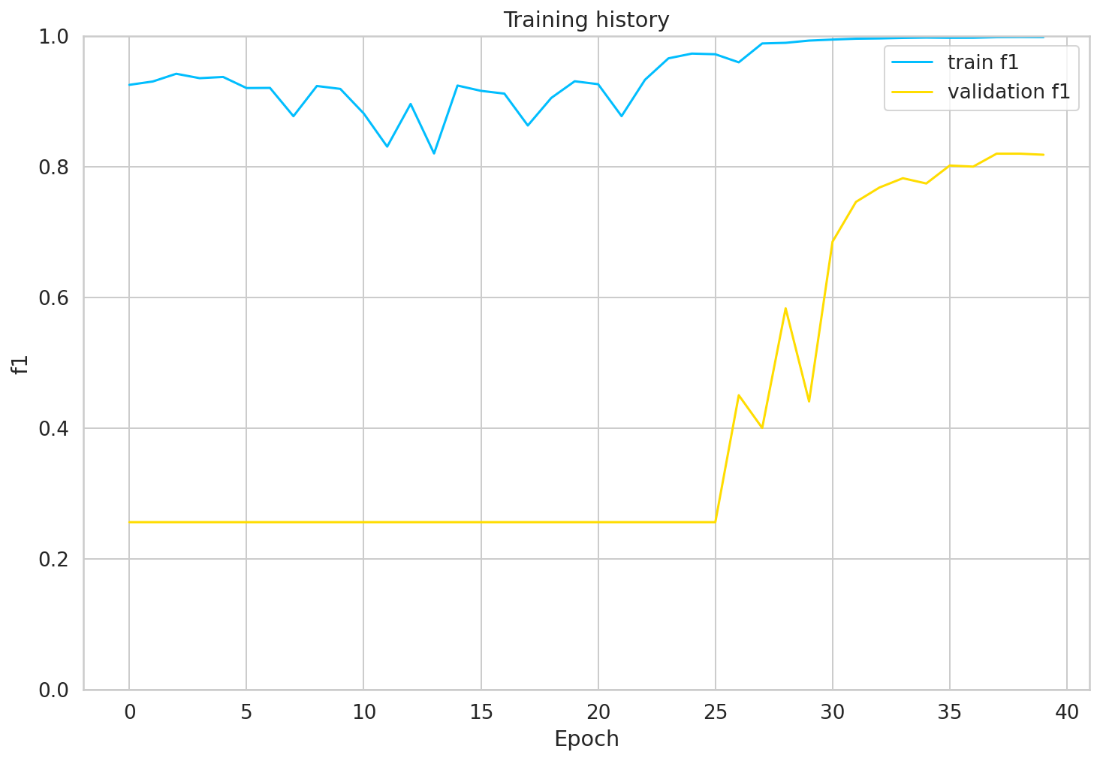
**Evaluation**

Evaluasi merupakan tahapan yang dilakukan untuk mengetahui seberapa baik performa dari model. Evaluasi pada penelitian ini terbagi menjadi 2 bagian yaitu validasi pada proses training model dan bagian evaluasi model akhir. Data training diambil sebagian dengan proporsi 0.1 secara acak yang akan dibagi ke dalam dua dataset untuk kebutuhan dua bagian evaluasi.

Proses evaluasi pada bagian pertama dijalankan setiap kali forward pass dan backward pass (setiap satu epoch). Proses evaluasi ini dilakukan untuk mengetahui performa model neural network dari tiap *epoch.* Hal ini berguna untuk melihat bagaimana perkembangan model seiring dengan bertambahnya *epoch* sehingga kita dapat tahu apakah masih perlu dilakukan train lagi pada model atau tidak. Evaluasi pada bagian ini juga menentukan state neural network terbaik yang akan dipakai sebagai model akhir. Berikut merupakan evaluasi tiap *epoch.* Terlihat training lebih lanjut dengan jumlah *epoch* lebih besar dapat meningkatkan performa lebih lanjut.

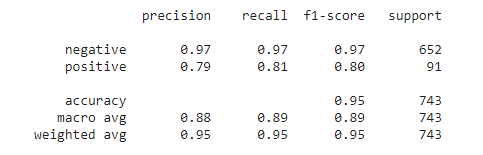


Gambar 7 Loss Training

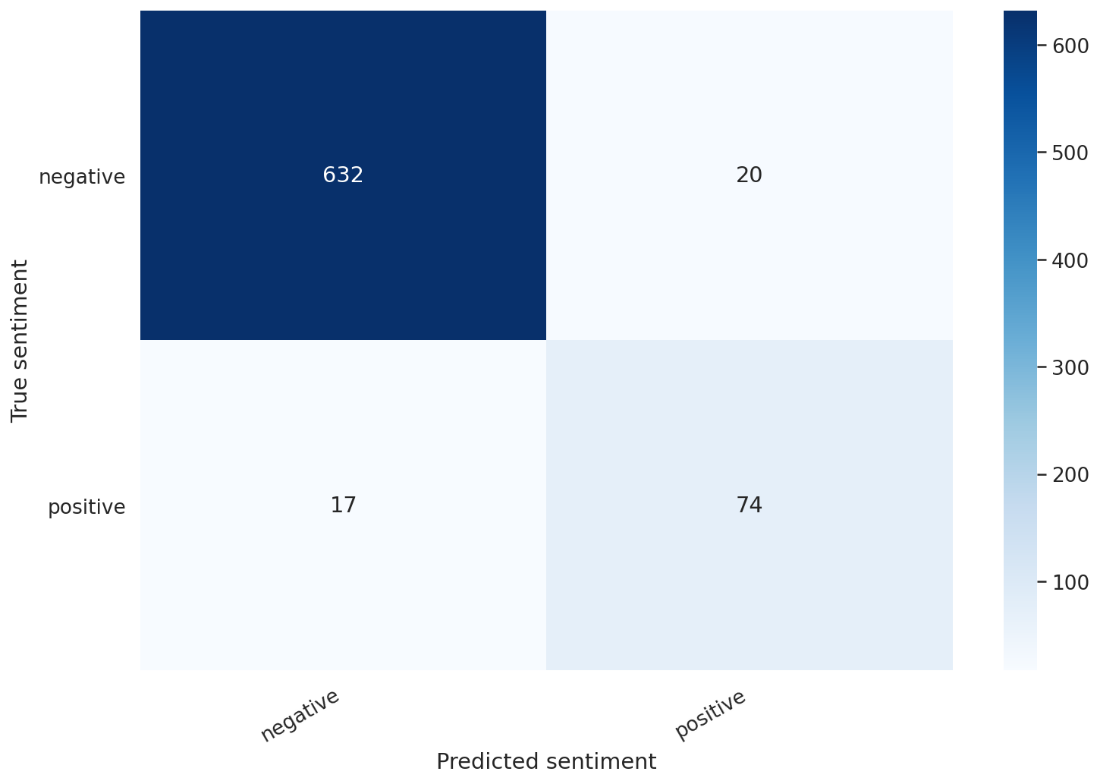


Gambar 8 F1 Training

Proses evaluasi kedua pada dasarnya merupakan pengecekan model akhir terhadap dataset baru yang tidak masuk dalam proses training. Proses ini berguna untuk melihat performa model jika nantinya digunakan pada dataset baru. Nilai evaluasi pada bagian kedua yang sama dengan pada bagian pertama menunjukan konsistensi model. Berikut hasil evaluasinya.



Gambar 9 Evaluasi Data Test



Gambar 10 Condussion Matrix Data Test

Nilai *loss function*, *accuracy*, dan F1 digunakan pada kedua bagian evaluasi. Namun pada penentuan state model neural network pada bagian pertama, hanya nilai F1 pada dataset validasi yang digunakan untuk memilih model. Hal ini mengacu pada penilaian kompetisi yang menggunakan nilai F1 pada leaderboard.