Seminar Nasional Rekayasa, Sains dan Teknologi
Vol 3 No 1Tahun 2023

Implementasi Metode *Bidirectional Encoder Representations from Transformers* (BERT) untuk Analisis Sentimen Komentar Pengguna Aplikasi Dana di Instagram

Firdaus Ihsan Septian*[1],Ivana Lucia Kharisma[2], Hermanto[3],Kamdan[4]
Program Studi Teknik Informatika, Universitas Nusa Putra[1,2,3,4]
firdaus.ihsan_ti19@nusaputra.ac.id*[1],ivana.lucia@nusaputra.ac.id[2],
hermanto@nusaputra.ac.id[3],kamdan@nusaputra.ac.id[4]

***Abstrak*— Kemajuan teknologi yang pesat saat ini mempengaruhi berbagai aspek kehidupan serta memberi kemudahan serta efisiensi pada berbagai aspek. Penerapan teknologi salah satunya di bidang finansial, yaitu dengan semakin banyak layanan keuangan digital yang memberi kemudahan bagi transaksi keuangan. Salah satu jenis keuangan digital yang banyak digunakan di masyarakat adalah aplikasi Dana. Dana menyediakan layanan yang dapat digunakan penggunanya serta sering memberikan informasi produk melalui akun media sosial Instagram. *Feedback* serta komentar tentang aplikasi didapatkan dari pengguna. Dengan menerapkan pemodelan *Bidirectional Encoder Representations From Transformers* (BERT) dari *IndoBert* pada proses analisa sentimen dari komentar pengguna Aplikasi DANA di Instagram pada penelitian ini , diharapkan dapat memberi informasi dan memudahkan dalam memahami opini dari pengguna, mendeteksi masalah dan keluhan, serta menjadikan wawasan bagi pengguna terhadap aplikasi Dana. Dari latar belakang tersebut, penelitian tentang analisa sentimen komentar dari pengguna aplikasi Dana dilakukan . Data yang digunakan didapat dari komentar akun Instagram Dana. Data tersebut terbagi menjadi 2 kategori yaitu positif dan negatif berdasarkan pelabelan otomatis oleh *transformer.* Dari hasil pemodelan dengan metode *pre-trained Bidirectional Encoder Representations From Transformers* (*BERT)* dari *IndoBert,* diperoleh hasil accuracy 98% dari data latih serta validasi *accuracy* sebesar 93% dengan *hyperparameter* yaitu *batch size* 32 dan *epoch* pelatihan 10 dengan proporsi data latih dan data uji 70:30. Pemodelan kemudian dilakukan proses *deployment* menggunakan *streamlit,* agar dapat diintegrasikan ke sistem atau aplikasi berbasis *web*.**

Keywords — Dana, pre-trained Bidirectional Encorder Representations From Transformers (BERT), IndoBert, Analisis Sentimen

Abstrak— Rapid chnological advances are currently affecting various aspects of life and providing convenience and efficiency in various aspects. One of the applications of technology is in the financial sector, namely with the increasing number of digital financial services that make financial transactions easier. One type of digital finance that is widely used in society is the Dana application. Dana provides services that its users can use and often provides product information via its Instagram social media account. Feedback and comments about the application are obtained from users. By applying Bidirectional Encorder Representations From Transformers (BERT) modeling from IndoBert to the sentiment analysis process from DANA Application user comments on Instagram in this research, it is hoped that it can provide information and make it easier to understand user opinions, detect problems and complaints, and provide insight for users. to the Fund application. From this background, research was conducted on sentiment analysis of comments from Dana application users. The data used was obtained from comments on Dana's Instagram account. The data is divided into 2 categories, namely positive and negative based on automatic labeling by the transformer. From the modeling results using the pre-trained Bidirectional Encorder Representations From Transformers (BERT) method from IndoBert, results obtained were 98% accuracy from the training data and validation accuracy was 93% with hyperparameters, namely batch size 32 and training epoch 10 with the proportion of training data and test data 70:30. The modeling process is then carried out using streamlit, so that it can be integrated into a web-based system or application.

Keywords — Dana, pre-trained Bidirectional Encorder Representations From Transformers (BERT), IndoBert, Sentimen Analytic

I. PENDAHULUAN

Kemajuan Teknologi yang pesat saat ini mempengaruhi berbagai aspek kehidupan menjadi serba mudah dan efisien, seiring dengan penetrasi teknologi digital yang sangat dalam dan digunakan secara luas, dampak teknologi digital akan semakin terasa, terutama di dunia bisnis. Salah satunya industri finansial yang berinovasi menyediakan layanan dompet digital atau *E-wallet*.

Dana adalah dompet digital Indonesia yang dirancang untuk menangani semua transaksi tunai dan kartu digital *online* dan *offline* dengan kecepatan, kenyamanan dan keamanan yang terjamin. Talenta terbaik Indonesia akan terus mengembangkan Dana sebagai dompet digital *open platform* yang dapat digunakan untuk mendukung segala aktivitas keuangan dan gaya hidup digital seluruh masyarakat Indonesia. Dengan Dana, masyarakat dapat menjadi lebih produktif, efisien dan kompeten. Dana juga dapat dioptimalkan untuk mendukung komitmen pemerintah dalam menekan biaya produksi dan distribusi uang fisik, serta meningkatkan literasi dan inklusi keuangan masyarakat Indonesia. Dana merupakan bukti kemampuan Indonesia dalam membangun dan mengembangkan teknologi dan infrastruktur ekonomi digital yang dapat dipercaya setiap saat [1].

Pada penelitian aplikasi Dana dipilih karena berdasarkan survei yang dilakukan oleh merdeka, Dana menjadi aplikasi popular setelah Gopay dan Ovo, dimana Dana memiliki tingkat pertumbuhan terpesat dalam jumlah penggunaanya [2]. Dana juga menyediakan banyak layanan yang dapat digunakan penggunanya, dari banyaknya layanan itu menimbulkan *feedback* dari pengguna seperti layanan *top up* yang gagal, tidak memperoleh *cashback* yang seharusnya, tidak sesuai dengan yang dikampanyekan dan lainnya, Pada akun Instagram Dana, Dana sering kali melakukan kampanye agar pengguna mengetahui informasi terbarunya. Kampanye tersebut menuai banyak komentar yang ditulis pengguna, dari banyaknya komentar yang diperoleh dari kampanye tersebut perlunya analisis sentiment agar perusahaan dapat terbantu dalam pemahaman opini pengguna, pemantauan respon pengguna untuk mengetahui bagaimana responnya, mendeteksi masalah dan keluhan, selain bagi perusahaan bagi pengguna juga mendapat wawasan terhadap aplikasi Dana, salah satunya reputasi perusahaan Dana .

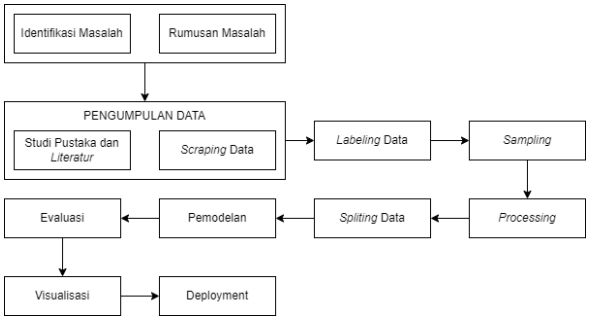
Media sosial saat ini sudah menjadi tempat bagi masyarakat untuk mengeluarkan ekpresi dan pendapat dari berbagai macam topik salah satunya yaitu Instagram. Instagram adalah sebuah *platfrom* media sosial untuk berbagi cerita dengan foto atau video ke sesama pengguna[3]. Indonesia merupakan negara pengguna Instagram terbanyak dengan 105 juta pengguna atau 37,8% dari jumlah populasi di Indonesia pada maret 2023 yang tercatat pada napoleoncat, jumlah pengguna tersebut merupakan terbesar keempat di dunia [4]. Pengguna Instagram dapat dengan bebas memberikan komentar untuk mengeluarkan pendapat terhadap postingan tersebut, tidak bisa dipungkiri pengguna sering kali berkomentar dengan kata-kata kasar dan tidak segan melontarkan ujaran kebencian. Pada akun Instagram @Dana.id yang postingannya selalu mendapatkan komentar, baik itu sebuah masukan yang bersifat membangun atau menjatuhkan. dari banyaknya komentar yang disampaikan pemilik akun, perlu mengidentifikasikan masalah yang muncul, namun untuk membaca dan mengklasifikan setiap komentar perlu waktu yang lama dan tidak efektif. Maka dari itu perlunya sebuah sistem yang dapat mengklasifikan komentar kedalam kelas sentimen secara otomatis serta analisis yang cocok.

Sentiment Analysis (analisis sentimen) atau sebuah *opinion mining* (penambangan opini) yang merupakan sebuah teknik pengolahan bahasa alami yang bertujuan untuk mengenali dan mengekspresikan opini, perasaan, evaluasi, sikap, subjektivitas, penilaian yang terkandung dalam suatu teks [5]. Penelitian pada bidang analisis sentimen sudah banyak diadakan dikarenakan persaingan pemasaran yang meningkat serta kebutuhan masyarakat yang berubah [6]. Analisis sentimen sangat berguna bagi pengembang *E-Wallet* untuk mengetahui pengalaman bertransaksi pengguna. Dengan membaca ulasan dari media sosial dapat memutuskan arah pengembangan dan peningkatan layanan dari *E-Wallet* [7].

Berdasarkan uraian diatas penulis memilih untuk melakukan Implementasi Metode *Bidirectional Encoder Representations from Transformers* (BERT) untuk Analisis Sentimen Komentar Pengguna Aplikasi Dana di Instagram.

II. METODE PENELITIAN

Penelitian ini memiliki beberapa tahapan, secara garis besar, alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart* Penelitian

Dari Gambar 1 diketahui alur dari penelitian dimana tahap pertama yang akan dilakukan adalah mengidentifikasi masalah dan membuat rumusan masalah, dilanjutkan oleh proses pengumpulan data yang diperlukan pada penelitian. Setelah data terkumpul, data akan diberikan label dan akan masuk ke tahap *processing*. Selanjutnya data akan masuk pada tahap *splitting* data, masuk pada tahap pemodelan, Tahap pemodelan akan di evaluasi, selanjutnya akan dilakukan visualisasi dan *deployment*.

Pada penelitian ini metode yang digunakan yaitu metode kuantitatif deskriptif yang dimana penelitian mempelajari populasi atau sampel tertentu dengan mengumpulkan data menggunakan alat tertentu [8]. Metode kuantitatif digunakan untuk mengukur penilaian dalam BERT dalam mengklasifikasin komentar secara numerik dengan skala atau nilai yang terukur.

A. Identifikasi Masalah

Aplikasi Dana merupakan dompet digital yang dari jumlah penggunaanya berkembang begitu pesat dibandingkan dompet digital popular lainnya, Dana memiliki banyak layanan yang dapat dipergunakaan oleh penggunanya, Dana juga sering melakukan kampanye lewat akun instagramnya agar infomasi terbaru tersampaikan ke penggunanya. Tentunya dari banyak layanan dan kampanye yang dilakukan menimbulkan *feedback* dari penggunanya yang dilontarkan lewat komentar pada postingan akun intagram Dana salah satunya, banyak komentar yang bersifat membangun atau menjatuhkan, namun untuk membaca dan mengklasifikasikan setiap komentar memerlukan waktu yang lama dan tidak efektif. Maka perlunya sebuah sistem untuk mengklasifikan komentar pada kelas sentimen secara otomatis serta analisis yang cocok. Setelah dilakukan analisis perlunya pengukuran perfoma model agar dapat diketahui sejauh mana model menyelesaikan tugasnya.

B. Rumusan Masalah

Rumusan masalah pada penelitian ini yaitu bagaimana implementasi metode Bidirectional Encoder Representations From Transformers (*BERT)* untuk analisis komentar Instagram pada akun Dana dan mengukur performa model analisis ini.

C. Pengumpulan Data

C.1. Studi Pustaka dan Literatur

Studi literatur merupakan sebuah proses pengumpulan data atau sumber yang berhubungan dengan judul penelitian yang digunakan untuk dipelajari. Studi literatur dapat diperoleh dari beberapa sumber yaitu jurnal, buku, internet, dan penelitian sejenis.

Pada penelitian ini studi pustaka dan literatur yang dilakukan yaitu bersumber dari buku yang berkaitan dengan analisis sentimen yang ada pada perpustakan kampus dan jurnal yang diperoleh menggunakan internet dari situs google scholar menggunakan kata kunci “analisis sentimen menggunakan BERT”, “implementasi algoritma BERT”, dan “analisis sentimen komentar Instagram menggunakan BERT” dari kata kunci tersebut jurnal jurnal yang diperlukan dan berkaitan akan muncul.

C.2. Scraping Data

Proses pengumpulan data dilakukan menggunakan bantuan alat yang ada pada *ekstensi google chrome* yaitu *Data Miner*. Data yang diambil merupakan sebuah komentar dari beberapa postingan pada akun dana dimana komentar tersebut harus digulir agar mendapatkan data yang lebih banyak. Data yang terkumpul sebanyak 1331, waktu yang diperlukan untuk pengumpulan data ini bergantung pada kecepatan internet yang digunakan untuk menggulir komentar pada setiap postingannya.

D. Pelabelan Dataset

Pada tahapan ini dataset yang sudah dikumpulkan diberikan label untuk setiap *record* data. Pelabelan ini bertujuan memberikan kategori pada setiap komentar, Label tersebut digunakan untuk mengidentifikasi atau mengklasifikasikan data ke dalam kelompok atau kategori tertentu. Pelabelan ini dilakukan secara otomatis menggunakan *Natural Language processing (NLP)* yang ada pada *library*

transformer, data yang sudah diberi label berguna sebagai data latih[9]. Dataset yang sudah dilabel secara otomatis akan dicek kembali apakah hasil label sudah memadai atau perlu diperbaiki secara manual.

E. Sampling

Pada tahapan ini dataset yang sudah memiliki label akan dihitung jumlah sampel yang diperoleh, jika pada jumlah sampel yang diperoleh dataset menunjukan ketidakseimbangan yang signifikan maka diperlukan *sampling* agar nantinya klasifikasi dilakukan dengan tepat dan tidak bias terhadap sampel mayoritas[10]. Pada penelitian ini *sampling* yang digunakan yaitu *Random Over Sampling* dimana data pada sampel minoritas akan diduplikat agar jumlahnya sama dengan sampel mayoritas.

F. Preprocessing

Preprocessing merupakan sebuah tahapan pembersihan data atau menyiapkan dataset agar dapat digunakan untuk melakukan proses training data. *Preprocessing* ini memudahkan data yang dimasukan dikenali oleh komputer [11]. Ada beberapa tahapan yang dilakukan saat *processing* ditunjukkan pada Gambar 2 berikut:

Gambar 2. Tahapan *Pre-Processing*

Pada *preprocessing* ini dataset akan dilakukan beberapa tahapan sesuai pada Gambar 2 dimana dataset akan melakukan *lowercase*, selanjutnya akan menghapus kata yang tidak memiliki arti penting, menghapus simbol serta tanda baca, menghapus angka, kemudian dilakukan *stemming*, serta *tokenisasi dan encorder*.

G. Spliting Data

Pada Tahapan ini dataset yang sudah melalui *preprocessing* dan sudah siap sebagai dataset untuk model dilakukan proses *splitting* data. *Spliting* data merupakan pembagian dataset menjadi 2 bagian yaitu data latih dan data uji. Data latih yaitu data yang akan dipergunakan untuk melatih model sedangkan data uji digunakan setelah proses *training* model selesai. Rasio yang digunakan pada pembagian data itu bergantung kepada data yang dimiliki hal ini karena tidak adanya panduan dalam data *split*[12]. Pada penelitian ini rasio yang digunakan 70:30 dimana 70% merupakan data latih dan sisanya 30% merupakan data uji.

H. Pemodelan

Pada penelitian ini model yang akan dipergunakan yaitu model *pretrained* model *indobenchmark/IndoBert-base-p2* dimana model ini sudah melakukan *pre-traning* dan akan disesuaikan kembali *(fine-tuning).* Model *IndoBert-base-p2* ini merupakan salah satu model dari *indobencmark* yang bertujuan untuk menyediakan dataset *benchmark*, model pra-pelatihan, dan metrik evaluasi untuk tugas pemrosesan bahasa alami (*Natural Language Processing*/NLP) dalam bahasa Indonesia [13].

Selanjutnya dilakukan *setup optimizer*. Pada penelitian ini *optimizer* yang digunakan yaitu Adam, hal ini memungkingkan untuk menemukan *learning rate* yang optimal pada proses iterasi. *Optimizer* Adam memiliki kemampuan untuk mengadaptasi *learning rate* secara otomatis untuk setiap parameter model, yang dapat membantu dalam mencapai konvergensi yang lebih cepat dan pelatihan yang lebih stabil.

I. Evaluasi

Evaluasi dilakukan untuk memahami sejauh mana model atau metode tersebut berhasil mencapai tujuan yang ditetapkan dan untuk menentukan seberapa baik model tersebut dapat memprediksi atau menggeneralisasi data yang tidak terlihat sebelumnya [9]. Perhitungan akurasi pada penelitian ini menggunakan *confusion matrix.* Hal ini meliputi :

1. Akurasi

Akurasi merupakan sebuah kinerja untuk menghitung seberapa akurat model yang digunakan dengan positif dan negatif yang diprediksi dengan benar terhadap total data. Berikut rumus dari akurasi:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

2. *Precision*

Precision adalah rasio observasi positif yang diprediksi dengan benar terhadap total observasi positif yang diprediksi. Berikut rumusnya

$$Precesion = \frac{TP}{TP+FP} \qquad (2)$$

3. *Recall*

Recall adalah rasio observasi positif yang diprediksi dengan benar terhadap semua observasi di kelas aktual.

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

4. *F1 Score*

F1 score adalah rata-rata tertimbang dari *precision* dan *recall.*

$$F1 = 2.\frac{Precesion\ .recall}{precesion+recall} \qquad (4)$$

J. Visualisasi

Pada Tahapan ini data akan divisualisasikan agar dapat mudah terbaca, salah satu visualisasi yang digunakan yaitu *Wordcloud, wordcloud* ini digunakan untuk menampilkan teks yang terdapat pada data dengan ukuran besar untuk kata yang lebih menonjol dalam data.

K. Deployment

Pada tahapan ini model yang sudah dilatih dan siap digunakan akan dilakukan *deployment* atau peluncuran aplikasi berbasis *web* yang menggunakan *framework Streamlit* agar dapat berfungsi dengan *User Interface* (UI), yang memungkinkan model dapat menerima *input* kalimat dan menghasilkan *output* analisis sentimen sesuai dengan prediksi model yang sudah dilatih sebelumnya.

III. HASIL DAN PEMBAHASAN

A. Dataset

Dataset yang berupa sebuah komentar dari beberapa postingan Instagram akun Dana, teknik Pengumpulan data (*Scraping*) dilakukan menggunakan bantuan alat yang ada pada *ektension google chrome* yaitu *Data Miner*. Total data yang berhasil dikumpulkan dapat dilihat pada tabel berikut :

Tabel 1. Hasil Pengumpulan Data

Sumber Data	Jumlah
Postingan Instagram Dana	1331

Dari hasil pengumpulan data tersebut penulis hanya akan menampilkan beberapa rincian kolom sebagai berikut :

Tabel 2. Rincian Data

NO	KOMENTAR
1	Viralka. Aja DANA ini..udah nggk jls pelayanannya...saldo tiba2 hilang nggk ada tanggung jawabnya...tutup aja DANA ini
2	Sejauh ini dana ku baik baik aja,kalian aja norak mungkin gapake premium dana, seharusnya ikuti peraturan dana,kalo uang kalian ilang terus tidak pake dana premium ya tanggung sendiri.
3	Ni gimana dah pelayanan dana kok gada yang bener dah apa minta di viral in dulu ya katanya sih #bebasdrama kocak
4	Selamanya saya bakalan black campaign layanan sampah kalian. Akun saya dibajak, uang saya dicuri, kalian tidak membantu saya sama sekali, bahkan sekedar memberitahu saya nomor rekening pembajak pun kalian tidak mau.
5	Dana gua premium setiap bulan limit abis mulu batas 40juta, tapi gak ada kendala, pada kenapa orang orang woy

Pada Tabel 2 menunjukan beberapa rincian data yang sudah terkumpul, dimana hanya komentar saja yang *discraping*.

B. Labeling

Pada tahapan ini dataset akan diberikan label agar model dapat mengenali pola atau karakteristik yang terkait setiap label, label ini dilakukan secara otomatis menggunakan *library Transformer* berikut merupakan hasil label yang dilakukan:

Tabel 3. Hasil *Labeling* Otomatis

Komentar	**Label**
Sejauh ini dana ku baik baik aja,kalian aja norak mungkin gapake premium dana, seharusnya ikuti peraturan dana,kalo uang kalian ilang terus tidak pake dana premium ya tanggung sendiri.	Positif
Ni gimana dah pelayanan dana kok gada yang bener dah apa minta di viral in dulu ya katanya sih #bebasdrama kocak	Negatif

Dari hasil label otomatis ini dataset mendapatkan dua kategori yaitu positif dan negatif kemudian data akan dicek kembali untuk memastikan bahwa data yang sudah diberi label otomatis sudah sesuai, jika tidak sesuai akan dilakukan perbaikan secara manual, berikut

jumlah positif dan negatif dari hasil *labeling* otomatis.

Tabel 4. Jumlah Hasil Label Otomatis

Label	Jumlah
Positif	147
Negatif	1184

Setelah dilakukan pengecekan kembali dari dataset yang sudah diberi label otomatis ada beberapa data yang diberi label tidak sesuai sehingga dilakukan label manual pada dataset sehingga jumlah data positif dan negatif mengalami perubahan sebagai berikut :

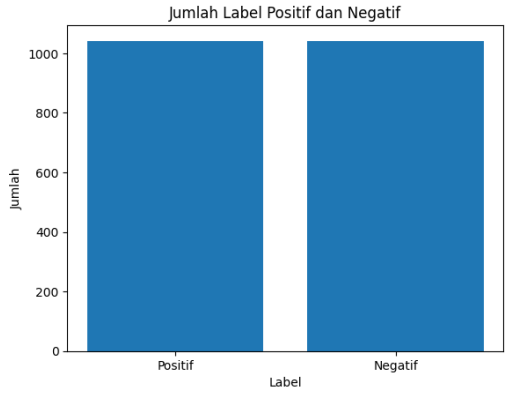
Tabel 5. Jumlah Setelah Perbaikan Label

Label	Jumlah
Positif	289
Negatif	1042

Dari jumlah label positif dan negatif menunjukan adanya ketidak seimbangan data dimana data negatif lebih dominan dibandingkan positif.

C. Sampling

Pada dataset yang sudah diberikan label diketahui jumlah label positif dan negatif tidak seimbang, maka dataset akan dilakukan *random oversampling* agar jumlah positif dan negatif seimbang dapat dilihat sebagai berikut :



Gambar 3. Hasil *Random Oversampling*

Dari Gambar 3 menunjukan jumlah label positif pada dataset sekarang seimbang dengan jumlah label negatif, dari hasil *random oversampling* dataset mengalami jumlah dataset sebagai berikut :

Tabel 6. Jumlah Dataset setelah *Random Oversampling*

	Data sebelum sampling	Data sesudah sampling
Jumlah Data	1331	2084
Positif	289	1042
Negatif	1042	1042

Diketahui dataset saat ini memiliki jumlah data sebanyak 2084 dengan label positif dan negatif yang jumlanya sama.

D. Preprocessing

D.1. Lowercase

Lowercase ini mengubah semua teks dalam dataset menjadi huruf kecil agar membantu dataset memiliki konsisten.

Tabel 7. *Lowercase*

Sebelum	Sesudah
BAYAR TAGIHAN AIR. SALDO SUDAH BERKURANG. TAPI UDH 1 BULAN STATUS SEDANG DI PROSES SAMPE TAGIHAN NUNGGAK !. UDH SAYA CHAT CS GADA RESPON SAMPE SEKARANG. KNP DANA SEKARANG SULIT BANGET DAN JARANG DI TANGAN	bayar tagihan air. saldo sudah berkurang. tapi udh 1 bulan status sedang di proses sampe tagihan nunggak !. udh saya chat cs gada respon sampe sekarang. knp dana sekarang sulit banget dan jarang di tangan
SALDO NYANGKUT DI TAMBAH KENA TIPU LGI 500RB ! ANYINGG EMNG GRGR APK DANA GAK JELAS INI BANGSAAAT!	saldo nyangkut di tambah kena tipu lgi 500rb ! anyingg emng grgr apk dana gak jelas ini bangsaaat!

D.2. Menghapus Kata yang Tidak Memiliki Arti Penting

Menghapus kata yang tidak penting ini merupakan sebuah istilah yang dikenal dengan *stopword*. *Stopword* ini adalah kata-kata umum yang tidak memiliki arti penting akan dihilangkan atau dihapus agar mengurangi komplek dan memfokuskan pada kata yang lebih informatif atau berarti [14].

Tabel 8. *Stopword*

Sebelum	Sesudah

ni gimana dah pelayanan dana kok gada yang bener dah apa minta di viral in dulu ya katanya sih #bebasdrama kocak	gimana pelayanan dana gada bener apa minta viral dulu katanya #bebasdrama kocak
saldo nyangkut di tambah kena tipu lgi 500rb ! anyingg emng grgr apk dana gak jelas ini bangsaaat!	saldo nyangkut tambah kena tipu lgi 500rb ! anyingg emng grgr apk dana gak jelas bangsaaat!

D.3. Menghapus Simbol dan Tanda Baca

Pada Tahapan ini kalimat pada dataset akan dibersihkan dari simbol dan tanda baca agar tidak mempengaruhi hasil analisis yang dapat menyebabkan klasifikasi kurang optimal[14].

Tabel 9. Hapus simbol dan tanda baca

Sebelum	Sesudah
Sejauh dana baik baik ,kalian aja norak mungkin gapake premium dana, seharusnya ikuti peraturan dana,kalo uang kalian ilang terus tidak pake dana premium tanggung sendiri.	Sejauh dana baik baik kalian norak mungkin gapake premium dana seharusnya ikuti peraturan dana kalo uang kalian ilang terus tidak pake dana premium tanggung sendiri
saldo nyangkut di tambah kena tipu lgi 500rb ! anyingg emng grgr apk dana gak jelas ini bangsaaat!	saldo nyangkut tambah kena tipu 500rb anyingg emng grgr apk dana gak jelas bangsaaat

D.4. Menghapus Angka

Proses penghapusan angka ini dilakukan karena pada penelitian kali ini berfokus pada teks, jika mencocokan kata, menghapus angka membantu menyamakan kata-kata yang sebenarnya memiliki makna yang sama.

Tabel 10. Menghapus Angka

Sebelum	Sesudah
Kalo 500 emng salah, udah ikhlasin nyangkut permasalahanya , gk percaya sama apk dana	Kalo emng salah, udah ikhlasin nyangkut permasalahanya , gk percaya sama apk dana
saldo nyangkut tambah kena tipu 500rb anyingg emng grgr apk dana gak jelas bangsaaat	saldo nyangkut tambah kena tipu rb anyingg emng grgr apk dana gak jelas bangsaaat

D.5. Stemming

Stemming merupakan sebuah proses pengolahan bahasa alami *Natural Language Processing* (NLP). Pada penelitian ini *stemming* yang digunakan yaitu sastrawi, sastrawi ini adalah library untuk mengubah kata menjadi kata dasar [3].

Table 11. *Stemming*

Sebelum	Sesudah
Sejauh dana baik baik kalian norak mungkin gapake premium dana seharusnya ikuti peraturan dana kalo uang kalian ilang tidak pake dana premium tanggung sendiri	jauh dana baik baik kalian aja norak mungkin gapake premium dana harus ikut atur dana kalo uang kalian ilang tidak pake dana premium tanggung sendiri

D.6. Tokenisasasi dan Data Encoder

Pada proses *tokenisasi* ini akan ditambahkan token khusus yang dimiliki kosa kata model yaitu menambahkan [CLS] pada bagian depan untuk memberitahu model bahwa kita sedang akan melakukan klasifikasi, yang kedua menambahkan [SEP] yang menandakan akhir kalimat dan yang terakhir [PAD] yang merupakan *padding* digunakan untuk menyamakan panjang data yang ada [15]. Selanjutnya token tersebut *diencode* menjadi angka yang sesuai dengan daftar kosa kata yang dimiliki model. Terakhir akan dibuat sebuah *attention mask*, *attention mask* ini menghasilkan token asli dan *padding* yang akan memberitahu model agar mengabaikan token *padding*, *attention mask* ini ditandai dengan 1 sebagai token asli dan 0 sebagi token *padding*.

Tabel 12. *Tokenisasi dan Encoder*

Proses	Sebelum	Sesudah
Tokenisasi	selamat malamsaya tunggu kebijakannya ya pihak dana tolong untuk tidak mengambil hak orang lain	['[CLS]', 'selamat', 'malam', '##saya', 'tunggu', 'kebijakan', '##nya', 'pihak', 'dana', 'tolong', 'tidak', 'mengambil',

		'hak', 'orang', '[SEP]', '[PAD]']
Encode	['[CLS]', 'selamat', 'malam', '##saya', 'tunggu', 'kebijakan', '##nya', 'pihak', 'dana', 'tolong', 'tidak', 'mengambil', 'hak', 'orang', '[SEP]', '[PAD]']	[2, 2368, 1217, 4660, 4034, 2315, 57, 1241, 1869, 3854, 119, 1632, 1319, 232, 3, 0]
Attention Mask	[2, 2368, 1217, 4660, 4034, 2315, 57, 1241, 1869, 3854, 119, 1632, 1319, 232, 3, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0]

E. Spliting Data

Pada Tahapan ini data akan dibagi menjadi data latih dan data uji dengan proporsi 70:30 Hal ini karena dataset yang dimiliki tidak begitu banyak selain itu dengan memberikan 70% data untuk dilatih membantu dalam pembelajaran pola yang lebih baik. Dari data uji akan dibagi kembali menjadi data test dan data validasi dengan proporsi 30:70 sehingga menghasilkan jumlah sebagai berikut :

Tabel 13. Jumlah Pembagian data

Nama Data	Jumlah
Data Latih	912
Data Test	275
Data Validasi	117

F. Pemodelan

Pada Tahapan ini data yang sudah siap akan dimasukan kedalam model *Pretrained* BERT dari *IndoBert* dengan parameter sebagai berikut :

- *Epochs : 5*
- *BATCH_SIZE : 32*
- *LEARNING_RATE : 5e-5*

Kemudian dilakukan *setup optimizer* menggunakan *optimizer* ADAM. Pada Gambar 4 merupakan hasil pelatihan data menggunakan model.

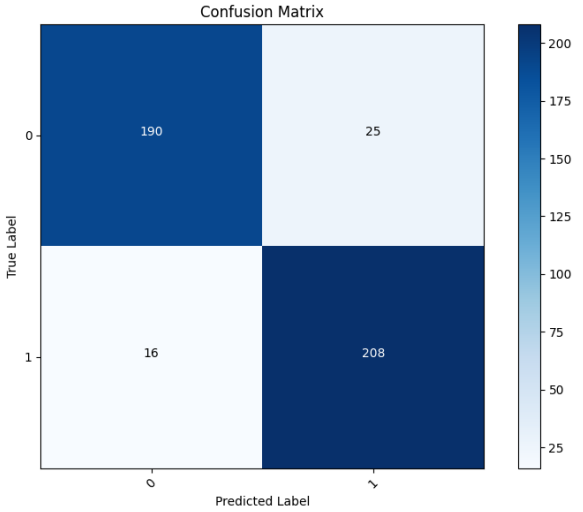
Epoch No.	Train Accuracy	Train Loss	Val Accuracy	Val Loss
1:	0.7188	0.5692	0.8717	0.3464
2:	0.8759	0.3095	0.8449	0.5195
3:	0.9499	0.1381	0.9198	0.2474
4:	0.9678	0.1020	0.9358	0.1846
5:	0.9753	0.0677	0.8663	0.5460
6:	0.9815	0.0575	0.9198	0.3024
7:	0.9870	0.0396	0.8770	0.4668
8:	0.9877	0.0391	0.8984	0.3830
9:	0.9856	0.0329	0.8610	0.6338
10:	0.9815	0.0573	0.9198	0.3251

Gambar 4 Hasil Pelatihan Model.

Pada saat Pelatihan model, model mendapatkan akurasi sebesar 98% dengan validasi akurasi sebesar 93%.

G. Evaluasi

Setelah pelatihan model yang dilakukan model akan dievaluasi untuk mengukur sejauh mana perfoma yang dimiliki oleh model terhadap data yang tidak terlihat sebelumnya. Dengan pengujian menggunakan *confusion matrix* sebagai berikut:



Gambar 5. *Confussion Matrix*

Dari confussion matrix tersebut didapatkan hasil laporannya sebagai berikut:

	precision	recall	f1-score	support
0	0.92	0.88	0.90	215
1	0.89	0.93	0.91	224
accuracy			0.91	439
macro avg	0.91	0.91	0.91	439
weighted avg	0.91	0.91	0.91	439

Gambar 6. *Classification Report*

Pada *classification report* yang dibuat menghasilkan presisi sebesar 0,92, recall 0.88, f1-score 0.90 untuk negatif yang ditandai dengan 0 sedangkan positif yang ditandai 1 mendapatkan presisi sebesar 0.89, recall 0.93, f1-score 0.91 dan berdasarkan *classification report* tersebut mendapatkan akurasi sebesar 0,91

H. Visualisasi

Dengan *visualisasi* ini memungkinkan data yang sebelumnya sulit dibaca dan tidak terlihat jelas dapat dibaca dengan jelas dan mudah, pada penelitian ini memanfaatkan *wordcloud* dengan menonjolkan data positif dan negatif yang dibedakaan berdasarkan warnanya.

Gambar 7. *Wordcloud* Sentiment positif

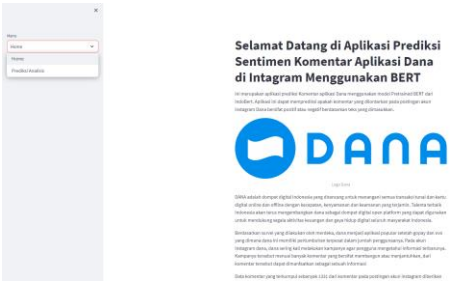
Dari Gambar 7 dapat diketahui kata yang menonjol pada sentimen positif yaitu ‘dana’, ‘masuk’, ‘ada’, ‘top’, ‘saldo. Selain dari sentimen positif terdapat sentimen negatif yang ditampilkan oleh *wordcloud* dapat dilihat pada Gambar 8.

Gambar 8. *Wordcloud* Sentimen Negatif

Diketahui pada Gambar 8 kata yang menonjol pada sentiment negatif yaitu ‘dana’, ‘uang’, ‘tolong’, ‘tolong, ‘transaksi

I. Deployment

Hasil modeling yang sudah dilakukan sebelumnya menggunakan BERT akan dibangun kedalam sebuah sistem menggunakan *framework streamlit*, agar model ini memiliki *user interface* dan dapat melakukan *input*, ditunjukkan pada Gambar 9 berikut :



Gambar 9. Tampilan UI *Streamlit*

Merupakan tampilan *user interface* hasil *modeling* yang dibuat menggunakan *framework streamlit* dan dapat melakukan prediksi terhadap komentar yang dimasukan.

Gambar 10. Tampilan prediksi komentar positif

Terlihat pada Gambar 10 jika dimasukkan komentar ‘aplikasi ini sangat membantu’, hasil prediksi yang dilakukan oleh model berupa positif dengan akurasi 88.39%. hasil lain yang berupa masukan komentar negatif dapat dilihat pada Gambar 11.

Gambar 11. Tampilan Prediksi Komentar Negatif

Dari Gambar 11 diketahui masukan komentar ‘baru kali ini dana mengecewakan’ menghasilkan prediksi negatif dengan akurasi 95.92%.

IV. KESIMPULAN

Berdasarkan pengujian dan penelitian Implementasi Metode *Bidirectional Encoder Representations from Transformers* (BERT) untuk Analisis Sentimen Komentar Pengguna Aplikasi Dana di Instagram dapat disimpulkan :

1. Dari data yang sudah terkumpul sebanyak 1331 menggunakan *Data Miner* yang merupakan data komentar di Instagram akun resmi Dana yang kemudian diberikan label memiliki data positif sebanyak 147 dan negatif sebanyak 1184, dari data yang yang sudah diberi label memiliki kesenjangan data yang sangat jauh antara positif dan negatif sehingga dilakukan *oversampling* agar data seimbang.

2. Dengan menggunakan model BERT dari *IndoBert* penulis dapat merancang dan membangun sebuah analisis sentiment berdasarkan kata yang menghasilkan sebuah prediksi positif maupun negatif.
3. Analisis Sentimen menggunakan BERT dari *IndoBert* menghasilkan akurasi sebesar 98% dan validasi akurasi sebesar 93% pada pelatihan selama 10 epoch dengan pembagian proporsi data 70:30, dapat disimpulkan bahwa model BERT dari *IndoBert* ini memiliki performa yang baik dalam menganalisis kata.
4. Berdasarkan pengujian model yang dilakukan menggunakan *confusion matrix*, model mendapatkan akurasi sebesar 91% serta presisi sebesar 0,92, *recall* 0.88, *f1-score* 0.90 untuk negatif sedangkan positif mendapatkan presisi sebesar 0.89, *recall* 0.93, *f1-score* 0.91.

DAFTAR PUSTAKA

[1] "About Dana." https://www.dana.id/about (accessed Feb. 05, 2023).

[2] Merdeka, "survei aplikasi dana," 2022. https://www.merdeka.com/teknologi/riset-sebut-dana-jadi-dompet-digital-dengan-pertumbuhan-pengguna-tercepat.html (accessed Oct. 21, 2023).

[3] M. K. Maulidina, "Analisis Sentimen Komentar Warganet Terhadap Postingan Instagram Menggunakan Metode Naive Bayes Classifier dan TF-IDF," *Naskah Publ. Univ. Teknol. Yogyakarta*, pp. 1–15, 2020.

[4] Napoleoncat, "pengguna instagram," *Instagram users in Indonesia*, 2023.

[5] N. Jindal and B. Liu, "Opinion spam and analysis," *WSDM'08 - Proc. 2008 Int. Conf. Web Search Data Min.*, pp. 219–229, 2008, doi: 10.1145/1341531.1341560.

[6] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Comput. Sci.*, vol. 87, no. December, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[7] A. A. Muhammad, Ermatita, and D. S. Prasvita, "ANALISIS SENTIMEN PENGGUNA APLIKASI DANA BERDASARKAN ULASAN PADA GOOGLE PLAY MENGGUNAKAN METODE SUPPORT VECTOR MACHINE Prodi S1 Informatika / Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 194–204, 2022.

[8] D. Pradipta, Kusrini, and H. Al Fatta, "Sentiment Analysis Comments Covid-19 Variant Omicron on Social Media Instagram with Bidirectional Encoder from Transformers (BERT) Sentimen Analisis Komentar Covid-19 Varian Omicron pada Media Sosial Instagram dengan Bidirectional Encoder from Transformer," *J. Sist. Telekomun. Elektron. Sist. Kontrol Power Sist. Komput.*, vol. 632, 2023, [Online]. Available: https://doi.org/10.32503/jtecs.v3i1.3219

[9] C. A. Putri, "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 2, pp. 181–193, 2020, doi: 10.35957/jatisi.v6i2.206.

[10] S. S. Utomo, T. A. Cahyanto, and B. H. Prakoso, "Penggunaan Algoritma Random Over Sampling Untuk Mengatasi Masalah Imbalance Data Pada Klasifikasi Gizi Balita," pp. 1–9, 2020.

[11] R. Mas, R. W. Panca, K. Atmaja1, and W. Yustanti2, "Analisis Sentimen Customer Review Aplikasi Ruang Guru dengan Metode BERT (Bidirectional Encoder Representations from Transformers)," *Jeisbi*, vol. 02, no. 03, p. 2021, 2021.

[12] P. Studi *et al.*, *PERBANDINGAN RASIO SPLIT DATA TRAINING DAN DATA TESTING MENGGUNAKAN METODE LSTM dalam memprediksi harga indeks saham asia*. 2022.

[13] Bryan Wilie and Karissa Vincentio and Genta Indra Winata and Samuel Cahyawijaya and X. Li and Zhi Yuan Lim and S. Soleman and R. Mahendra and Pascale Fung and Syafri Bahar and A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *wilie2020indonlu*, 2020. [Online]. Available: https://huggingface.co/indobenchmark/indobert-base-p2

[14] B. Kurniawan, A. Ari Aldino, and A. Rahman Isnain, "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (Pse) Menggunakan Algoritma Bidirectional Encoder Representations From Transformers (Bert)," *J. Teknol. dan Sist. Inf.*, vol. 3, no. 4, pp. 98–106, 2022, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[15] R. Kusnadi, Y. Yusuf, A. Andriantony, R. Ardian Yaputra, and M. Caintan, "Analisis Sentimen Terhadap Game Genshin Impact Menggunakan Bert," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 6, no. 2, pp. 122–129, 2021, doi: 10.36341/rabit.v6i2.1765.