**PENGUKURAN INDEKS KEBAHAGIAAN MASYARAKAT**

**KOTA SEMARANG PADA MEDIA SOSIAL TWITTER**

**MENGGUNAKAN PENDEKATAN BERT DENGAN**

***K-FOLD CROSS-VALIDATION***

**Logo

Description automatically generated**

# **PROPOSAL PENELITIAN**

**Disusun untuk Memenuhi Tugas Besar  
pada Mata Kuliah Metodologi dan Penulisan Ilmiah  
yang diampu oleh Dr. Retno Kusumaningrum, S.Si., M.Kom.**

**Disusun oleh:**

**Dyan Azka Ingkafi**

**24060118130139**

**DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA**

**FAKULTAS SAINS DAN MATEMATIKA**

**UNIVERSITAS DIPONEGORO**

**2020**

# **ABSTRAK**

Dewasa ini pengukuran tingkat kemakmuran suatu negara dari pendapatan per kapita *Gross National Product,* telah bergeser ke *happiness index* atau *Gross National Happiness* (GNH). Semarang sebagai ibukota dan kota terbesar di Provinsi Jawa Tengah memiliki *track record* indeks kebahagiaan yang cukup baik. Indeks kebahagiaan di Kota Semarang meningkat dua poin dalam 3 tahun. Tak hanya itu, peningkatan indeks kebahagian tersebut juga linier dengan peningkatan seluruh indikator, pada 17 Tujuan Pembangunan Berkelanjutan (TPB) atau *Sustainable Development Goals* (SDG's) di Kota Semarang. Pengukuran indeks kebahagiaan suatu wilayah menjadi hal penting karena dapat menjadi tolak ukur perkembangan dan keberhasilan pembangunan suatu wilayah. Sebelumnya pengukuran indeks kebahagiaan dilakukan dengan cara *survey* yang dapat memakan biaya besar dan waktu yang lama sehingga dibutuhkan suatu metode yang lebih cepat dan lebih praktis untuk diterapkan. Penelitian ini akan menganalisis tweet masyarakat Semarang dari Twitter secara berkala, ke dalam parameter *happiness* menggunakan pendekatan BERT *atau Bidirectional Encoder Representations from Transformers*. Terdapat empat tahap dalam penelitian ini yaitu, tahap pertama adalah *crawling* (pengambilan) data Twitter berdasarkan wilayah kota Semarang menggunakan API Twitter. Tahap kedua yaitu *preprocessing* *data* yang mencakup *data cleaning*, *case folding*, *tokenizing*, *stopword* *removal*, *stemming,* dan normalisasi kata. Tahap ketiga yaitu sentimen analisis untuk mengklasifikasikan tweet (opini) ke dalam opini positif dan negatif menggunakan BERT. Tahap terakhir yaitu melakukan evaluasi hasil klasifikasi dan melakukan validasi menggunakan *k-fold cross-validation*.

***Kata Kunci****:* Indeks kebahagiaan*,* sentimen analisis*,* BERT*, k-fold cross-validation,* Kota Semarang*,* Twitter

# **DAFTAR ISI**

[ABSTRAK i](#_Toc69464953)

[DAFTAR ISI ii](#_Toc69464954)

[DAFTAR GAMBAR iv](#_Toc69464955)

[DAFTAR TABEL v](#_Toc69464956)

[BAB I PENDAHULUAN 1](#_Toc69464957)

[1.1. Latar Belakang 1](#_Toc69464959)

[1.2. Rumusan Masalah 3](#_Toc69464960)

[1.3. Tujuan dan Manfaat 3](#_Toc69464961)

[1.4. Ruang Lingkup 3](#_Toc69464962)

[BAB II LANDASAN TEORI 4](#_Toc69464963)

[2.1. State of The Art 4](#_Toc69464965)

[2.2. Landasan Teori 5](#_Toc69464966)

[2.2.1. Pengukuran Happiness Index 5](#_Toc69464967)

[2.2.2. Analisis Sentimen 5](#_Toc69464968)

[2.2.3. *Natural Language Processing* (NLP) 6](#_Toc69464969)

[2.2.4. *Machine Learning* 6](#_Toc69464970)

[2.2.5. *Neural Network* 7](#_Toc69464971)

[2.2.6. *Deep Learning* 7](#_Toc69464972)

[2.2.7. *Bidirectional Encoder Representations from Tranformers* (BERT) 8](#_Toc69464973)

[2.2.8. *K-Fold Cross-Validation* 10](#_Toc69464974)

[BAB III METODOLOGI PENELITIAN 12](#_Toc69464975)

[3.1. Arsitektur Umum 12](#_Toc69464977)

[3.2. Pengumpulan Data (scraping) 13](#_Toc69464978)

[3.3. Labelisasi Dataset 13](#_Toc69464979)

[3.4. Pre-processing Dataset 13](#_Toc69464980)

[3.5. Implementasi BERT 14](#_Toc69464981)

[3.6. Evaluasi 15](#_Toc69464982)

[3.7. Validasi k-fold cross-validation 16](#_Toc69464983)

[BAB IV RENCANA JADWAL PENELITIAN 17](#_Toc69464984)

[4.1. Jadwal Penelitian 17](#_Toc69464986)

[DAFTAR PUSTAKA 18](#_Toc69464987)

[LAMPIRAN 20](#_Toc69464988)

# **DAFTAR GAMBAR**

[Gambar 2.1 Hubungan antara Kecerdasan Buatan, *Machine Learning*, dan *Deep Learning* (Chollet, 2017). 8](#_Toc69464645)

[Gambar 2.2 *Encoder* (kiri) dan *Decoder* (kanan) (Vaswani et al., 2017). 10](#_Toc69464646)

Gambar 3.1 Arsitektur Umum Sistem.........................................................................12

# **DAFTAR TABEL**

[Tabel 2.1 Penelitian Terdahulu Terkait Analisis Sentimen 4](#_Toc69464851)

[Tabel 3.2 Confusion Matrix 15](#_Toc69464852)

# **BAB I**

# **PENDAHULUAN**

## Latar Belakang

Dalam konteks masyarakat sebagai objek pembangunan, maka diperlukan suatu indikator untuk mengukur perkembangan kehidupan atau tingkat kesejahteraan masyarakat itu sendiri. Indikator keberhasilan suatu negara atau daerah bisa dilihat dari laju pertumbuhan ekonominya. Untuk melihat tingkat kesejahteraan dari segi pertumbuhan ekonomi secara umum indikator yang digunakan adalah Produk Domestik Regional Bruto (PDRB). PDRB perkapita merujuk pada pertumbuhan output perkapita, apabila output perkapita meningkat akan terjadi perubahan pada pola konsumsi. Tingkat kesejahteraan masyarakat dapat diukur dengan menggunakan standar yang tidak sama (indikator subyektif). Salah satu indikator kesejahteraan yang mengukur capaian berdasarkan standar yang tidak sama untuk masing-masing individu adalah indeks kebahagiaan. Indeks kebahagiaan adalah tingkat kebahagiaan atau kepuasan hidup penduduk Indonesia dalam skala 0-100. Menurut BPS Indeks kebahagiaan metode 2014 merupakan indeks komposit yang disusun oleh tingkat kepuasan terhadap 10 aspek kehidupan yang esensial. Kesepuluh aspek tersebut secara substansi dan bersama-sama merefleksikan tingkat kebahagiaan yang meliputi kepuasan terhadap: pendidikan, kesehatan, pekerjaan, pendapatan, keamanan, hubungan sosial, ketersediaan waktu luang, kondisi rumah, kondisi lingkungan dan keharmonisan keluarga.

Pengukuran indeks kebahagiaan saat ini masih dilakukan dengan cara survey, metode survey ini memiliki kekurangan dalam hal biaya yang besar, waktu yang lama dan umumnya dibutuhkan usaha keras dalam pelaksanaannya. Pengambilan opini tidak hanya dapat dilakukan melalui survey secara langsung, tapi dapat dilakukan melalui media sosial mengingat pertumbuhan media sosial sangat pesat di masyarakat. Didukung dengan pernyataan kominfo bahwa Indonesia menepati peringkat 5 sebagai pengguna media sosial Twitter terbanyak di dunia dengan 19,5 juta pengguna. Opini yang diberikan masyarakat dalam Twitter sangat beragam dan memungkinkan antar opini terdapat kesamaan istilah, struktur dan makna yang sebenarnya mengekspresikan domain pengetahuan yang sama. Oleh karena itu pada penelitian ini dilakukan sentimen analisis dengan pendekatan *deep learning* untuk mengelompokkan dan menganalisis opini masyarakat Kota Semarang di Twitter ke dalam parameter *happiness index*.

Hasil klasifikasi oleh BERT akan dilakukan validasi menggunakan *k-fold cross-validation*. *Cross-validation* adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis *cross-validation*dapat didasarkan pada ukuran dataset. Biasanya *k-fold cross-validation* digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi.

Permasalahan dengan topik analisis sentimen Twitter sudah banyak dilakukan sebelumnya, penelitian tersebut membahas tentang bagaimana menganalisis sentimen mahasiswa di Twitter menggunakan metode *backpropagation* (Habibi et al., 2016). Hasil penelitian ini menunjukkan rata-rata nilai presentase kebenaran yang didapatkan lebih dari 85%. Rata- rata nilai evaluasi performa adalah nilai precision sebesar 0,930, nilai recall sebesar 0,899, dan nilai F-measure sebesar 0,901. Penelitian lainnya yaitu oleh (Putri, 2020) yang melakukan analisis sentimen *review* film menggunakan pendekatan BERT. Hasil penelitian tersebut yaitu didapatkan hasil akurasi yang dihitung dengan menggunakan confusion matrix sebesar 73%. Penelitian yang berkaitan dengan pengukuran indeks kebahagiaan yaitu penelitian oleh (Rahayu Ponilan et al., 2016) yang menggunakan pendekatan Ontologi *Top-Down* *Hierarchy.* Hasil penelitian tersebut yaitu nilai f1 yang dihasilkan dari pengklasifikasian sentimen menggunakan SVM di Weka sebesar 84%, sedangkan nilai f1 yang dihasilkan dari pengklasifikasian parameter happiness menggunakan ontologi sebesar 67%. Hingga saat ini belum ada penelitian yang mengukur indeks kebahagiaan pada twitter dengan pendekatan BERT.

Berdasarkan latar belakang tersebut, maka pada penelitian yang akan dilakukan penulis yaitu menggunakan metode *deep learning* dengan *language model* yaitu BERT atau *Bidirectional Encoder Representations from Transformers* untuk mengukur indeks kebahagiaan Kota Semarang dengan menganalisis sentimen masyarakat Kota Semarang pada twitter.

## Rumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut.

1. Bagaimana implementasi BERT dan *k-fold cross-validation* untuk mengukur indeks kebahagiaan Kota Semarang dengan menganalisis sentimen masyarakat Kota Semarang pada twitter?
2. Bagaimana hasil analisis sentimen masyarakat Kota Semarang pada twitter menggunakan pendekatan BERT dan *k-fold cross-validation*?

## Tujuan dan Manfaat

Tujuan dalam penelitian ini adalah sebagai berikut.

1. Mengetahui bagaimana implementasi BERT dan *k-fold cross-validation* untuk mengukur indeks kebahagiaan Kota Semarang dengan menganalisis sentimen masyarakat Kota Semarang pada twitter.
2. Mengetahui bagaimana hasil analisis sentimen masyarakat Kota Semarang pada twitter menggunakan pendekatan BERT dan *k-fold cross-validation*.

Manfaat dalam penelitian ini adalah sebagai berikut.

1. Menambah pengetahuan mengenai penerapan BERT dan *k-fold cross-validation* pada proses sentimen analisis.
2. Menjadi bahan pertimbangan bagi pemerintah dalam hal pengembangan Kota Semarang.
3. Penelitian ini bermanfaat untuk dijadikan informasi dan bahan referensi untuk penelitian selanjutnya mengenai pendeteksian dan klasifikasi *hate speech* pada Twitter.

## Ruang Lingkup

Ruang lingkup dalam penelitian ini adalah.

1. Data berupa cuitan masyarakat Semarang pada twitter.
2. Data hanya menggunakan bahasa Indonesia.
3. Hanya membagi data ke dalam 3 jenis sentimen yaitu positif, netral, dan negatif*.*

# **BAB II**

# **LANDASAN TEORI**

## *State of The Art*

Tabel 2.1. Penelitian Terdahulu Terkait Analisis Sentimen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NO | PENELITIAN | DOMAIN MASALAH (BAHASAN) | PENDEKATAN | METODE | HASIL |
| 1 | (Habibi et al., 2016) | Analisis sentimen Twitter | *Backpropagation* | *Backpropagation* | Rata-rata nilai prosentase kebenaran yang didapatkan adalah lebih dari 85%. Nilai precision sebesar 0,930, nilai recall sebesar 0,899, dan nilai F-measure sebesar 0,901. |
| 2 | (Rahayu Ponilan et al., 2016) | Analisis sentimen Twitter | SVM | Weka | Nilai f1 sebesar 84%, sedangkan nilai f1 yang dihasilkan dari pengklasifikasian parameter happiness menggunakan ontologi sebesar 67%. |
| 3 | (Putri, 2020) | Analisis sentimen *Review Film* | BERT | BERT | Hasil akurasi yang dihitung dengan menggunakan confusion matrix sebesar 73% |
| 4 | (Dwi, 2021) | Analisis sentimen *Review Film* | BERT | BERT | Hasil rata-rata yang diperoleh sebesar 66.7% |
| 5 | (Mozafari et al., 2020) | Analisis sentimen Twitter | BERT | CNN | Nilai F1 sebesar 88% untuk dataset Waseem dan 92% untuk dataset Davidson |
| 6 | (Alfina et al., 2018) | Analisis sentimen Twitter | RFDT | *N-gram Features* | Menghasilkan *f-measure* sebesar 93.5% |
| 7 | (Liu & Sari, 2019) | Analisis sentimen Twitter | *Naive Bayes* | *Information Gain Features* dan Normalisasi Kata | Hasil akurasi sebesar 98%, nilai *precision* sebesar 100%, nilai *recall* sebesar 96,15%, dan nilai *f*-*measure* sebesar 98,03%. |
| 8 | (Sidiq et al., 2019) | Analisis sentimen Twitter | SVM | *Crowdsourced Labeling* | Nilai akurasi sebesar 97% dan validitas data sebesar 98% data valid |
| 9 | (Zulfa & Winarko, 2017) | Analisis sentimen Twitter | DBN | DBN | Hasil akurasi sebesar 93,91% |

## Landasan Teori

### **Pengukuran Happiness Index**

Pengukuran Happiness merupakan usaha yang telah diupayakan selama bertahun-tahun di dunia. Akan tetapi baru berhasil dilaporkan sejak tahun 2012 dengan diterbitkannya *World Happiness Report* pada 2 April 2012 oleh Perserikatan Bangsa-Bangsa. Seiring dengan perkembangan media sosial yang sangat pesat, terdapat pendekatan lain sebagi alternatif untuk pengukuran *happiness*, yaitu yang sudah dilakukan oleh beberapa penelitian diantaranya oleh (Kramer, 2010), dengan mengukur *behaviour model* masyarakat sebuah negara yang tergambar pada media sosial. Pendekatannya dilakukan dengan memodelkan sentimen positif dan negatif masyarakat di suatu negara dan menghitung score-nya. Apabila lebih besar score sentimen positif maka dapat disimpulkan masyarakat pada negara tersebut bahagia dan sebaliknya.

### **Analisis Sentimen**

Menurut (Lin et al., 2016) tentang analisis sentimen, yang disebut juga dengan *opinion mining*, merupakan salah satu cabang ilmu dari data mining yang bertujuan untuk menganalisis, memahami, mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas seperti produk, servis, organisasi, individu, dan topik tertentu. Analisis ini digunakan untuk mendapatkan suatu informasi tertentu dari suatu kumpulan data yang ada*.* Analisis sentimen berfokus pada pengolahan opini yang mengandung polaritas, yaitu memiliki nilai sentimen positif, netral, ataupun negatif.

### ***Natural Language Processing* (NLP)**

*Natural Language Processing* atau Pengolahan Bahasa Alami adalah salah satu cabang ilmu kecerdasan buatan yang mempelajari dan mengembangkan bagaimana komputer dapat mengerti, memahami, dan memproses bahasa alami dalam bentuk teks atau tuturan kata. NLP menganalisa bahasa manusia sedemikian

rupa sehingga komputer dapat memahami bahasa alami seperti halnya manusia (Ghosh et al., 2012). NLP adalah salah satu bidang antar disiplin yang menggabungkan komputasi linguistik, ilmu komputasi, ilmu kognitif, dan kecerdasan buatan. Pada umumnya, NLP banyak diaplikasikan di berbagai hal seperti *speech recognition*, pemahaman bahasa lisan, sistem dialog, analisis leksikal, mesin penerjemah, *knowledge graph*, analisis sentimen, sistem pintar dan peringkasan bahasa alami.

Sebuah sistem NLP dapat dimulai dari tingkat kata untuk menentukan struktur dan sifat morfologis (seperti *part-of-speech* atau makna) dari kata; kemudian dapat beralih ke tingkat kalimat untuk menentukan urutan kata, tata bahasa, dan arti dari seluruh kalimat. Kemudian ke konteks dan keseluruhan domain. Kata atau kalimat yang diberikan mungkin memiliki makna atau konotasi yang berbeda dalam konteks tertentu, yang terkait dengan banyak kata atau kalimat lain dalam konteks yang diberikan.

### ***Machine Learning***

*Machine learning* atau pembelajaran mesin adalah salah satu bidang ilmu di

kecerdasan buatan. *Machine learning*, sesuai dengan namanya, bertujuan untuk membuat mesin dilatih dengan banyak contoh atau *dataset* yang berhubungan dengan tugas yang dibutuhkan. Mesin mempelajari pola-pola yang diberikan berdasarkan *dataset* dan menghasilkan sebuah rule sendiri. Sehingga ketika suatu data dimasukkan ke dalam mesin, mesin sudah dapat mengenali data tersebut. Secara umum, *machine learning* terbagi menjadi empat kategori besar yaitu *supervised learning*, *unsupervised learning*, *self-supervised learning*, dan *reinforcement learning* (Chollet, 2017). *Supervised learning* adalah pendekatan yang paling sering digunakan*. Supervised learning* membuat mesin belajar dari *dataset* yang sudah diberi label atau anotasi. Sedangkan *unsupervised learning* merupakan kebalikannya, dengan memberikan dataset yang tidak diberi label *Selfsupervised learning* adalah sebuah *supervised learning* tetapi tanpa *dataset* yang dilabeli oleh *annotator*. *Dataset* yang digunakan tetap menggunakan label akan tetapi label diperoleh dari input data yang menggunakan algoritma *heuristic* (Chollet, 2017).Algoritma yang sering digunakan pada machine learning antara lain *K-Nearest Neighbor, Naïve-Bayes, Support Vector Machine, K-Means*, dan lain-lain.

### ***Neural Network***

*Neural Network* atau jaringan syaraf tiruan adalah salah satu teknik *machine* *learning* yang populer dengan mensimulasikan mekanisme pembelajaran yang terinspirasi dari bagaimana cara sistem syaraf manusia atau makhluk biologis lainnya. Sistem saraf terdiri dari sel yang disebut dengan *neuron*. *Neuron*-*neuron* tersebut saling terhubung dengan satu sama lain menggunakan *axon* dan *dendrites*. Sinapsis adalah penghubung antara *axon* dan *dendrites* (Aggarwal, 2018).

*Network* sendiri disebut sebagai arsitektur, di mana berbagai layers saling terhubung dengan satu sama lain. *Layer* yang ada di antara *input* *layer* dan *output* *layer* disebut dengan *hidden layers* dan *output* dari *hidden layer* disebut dengan *hidden units* (Osinga, 2018). Istilah *hidden* diberikan karena unit-unit tersebut tidak dapat langsung terlihat dari luar sebagai *input* atau *output*. Inti dari *neural network* adalah *hidden layer* yang dibentuk dari *hidden units*, yang masing-masing merupakan *neural units,* mengambil *weighted sum* dari *input* dan kemudian menerapkan *non-linearity*. Setiap unit pada tiap *layer* mengambil *input* dan *output* dari semua unit di *layer* sebelumnya dan hubungan antara setiap pasangan unit dari dua *layer* yang saling berdekatan sehingga setiap *layer* saling terhubung. Tiap *hidden units* akan menjumlahkan semua *input* unit (Jurafsky & Martin, 2008).

### ***Deep Learning***

*Deep Learning* adalah cabang dari *machine learning* yang merupakan bagian dari Kecerdasan Buatan. *Deep learning* merupakan *neural network* yang lebih modern dan bersifat *deep* atau mendalam karena memiliki jauh lebih banyak layer dibandingkan dengan *neural network* pada biasanya (Jurafsky & Martin, 2008) (Chollet, 2017). Kata “*deep*” mengacu pada jumlah hidden layer yang ada, semakin banyak layernya, maka semakin “*deep*” pembelajaran yang dilakukan oleh jaringan. *Deep learning* bekerja untuk mempelajari sehingga tidak hanya dapat memprediksi tetapi juga merepresentasikan data dengan benar, sehingga cocok untuk melakukan prediksi (Goldberg, 2017). *Deep learning* dapat dibagi ke dalam tiga metode pendekatan yaitu *supervised*, *semi-supervised*, dan *unsupervised learning*. *Deep learning* didukung oleh banyak framework seperti Torch, Theano, TensorFlow, dan lain-lain. Hubungan antara kecerdasan buatan, *machine learning,* dan *deep learning* dapat dilihat pada Gambar 2.1.

Diagram, venn diagram

Description automatically generated

Gambar 2.1 Hubungan antara Kecerdasan Buatan, Machine Learning, dan Deep Learning (Chollet, 2017).

### ***Bidirectional Encoder Representations from Tranformers* (BERT)**

*Bidirectional Encoder Representations from Transformers* (Devlin et al., 2019) atau disingkat BERT adalah model representasi bahasa terlatih yang dikembangkan oleh para peneliti di Google AI Language pada tahun 2018. BERT dikembangkan berdasarkan teknik-teknik deep learning dan berbagai metode seperti *semi-supervised* *learning*, ELMo, ULMFiT, OpenAI Transformers, dan Transformers. Sesuai dengan namanya, BERT menggunakan Transformer. Transformer adalah sebuah mekanisme yang mempelajari hubungan kontekstual antara kata-kata dalam teks (Vaswani et al., 2017). Transformer dapat memahami dan mengkonversi pemahaman yang diperoleh dengan mekanisme yang bernama *self-attention mechanism*. *Self-attention mechanism* adalah cara Transformer untuk mengubah “pemahaman” kata terkait lainnya menjadi kata-kata yang akan diproses dengan mekanismenya. Pada Transformer terdapat dua mekanisme, yaitu:

* 1. *Encoder*

*Encoder* berfungsi untuk membaca seluruh *input* teks sekaligus. *Encoder* terdiri dari *stack* (tumpukan) dari N = 6 *layers* yang identik. Setiap *layer* memiliki dua *sub-layer* yaitu *self-attention layer* dan *feed-forward neural* *network*. Dengan *self-attention layer*, *encoder* dapat membantu *node* untuk tidak hanya fokus kepada kata yang sedang dilihat tetapi juga untuk mendapatkan konteks semantik dari kata tersebut. Setiap posisi di *encoder* dapat menangani semua posisi di *layer* sebelumnya di *encoder*.

* 1. *Decoder*

*Decoder* berfungsi untuk menghasikan urutan *output* yang berupa prediksi. *Decoder* juga terdiri dari *stack* (tumpukan) dari N = 6 *layers* yang identif. Setiap layer terdiri dari dua sub-layerseperti yang ada pada *encoder*, dengan tambahan *attention* *layer* di antara dua *layers* tersebut untuk membantu *node* saat ini mendapatkan *key* *content* yang membutuhkan *attention* (Vaswani et al., 2017) dengan melakukan *multi-head attention* pada *output* dari *encoder*. Sama dengan di *encoder*, *self-attention layer* di *decoder* membuat setiap posisi di *decoder* dapat menangani semua posisi sebelumnya dan posisi saat itu. Diagram mengenai cara kerja *encoder* dan *decoder* pada Transformer dapat dilihat pada Gambar 2.2.

Diagram

Description automatically generatedGambar 2.2 *Encoder* (kiri) dan *Decoder* (kanan) (Vaswani et al., 2017).

Arsitektur model BERT berupa *multi-layer bidirectional* Transformer seperti yang dilakukan pada implementasi asli Transformer tetapi hanya menggunakan proses sampai encoder saja. Pada implementasinya, terdapat dua ukuran model yang ada pada BERT, yaitu BERTBASE dan BERTLARGE. Kedua ukuran model BERT ini memiliki banyak lapisan *encoder* atau *Transformer* *Blocks*. BERTBASE memiliki *encoder* dengan 12 *layers*, 12 *self-attentions heads*, *hidden size* sebesar 768, dan 110M *parameters*. Sedangkan BERTLARGE terdapat 24 *layers*, 16 *self-attention heads, hidden size* sebesar 1024, dan 340M *parameters*. BERTBASE dilatih selama 4 hari menggunakan 4 *cloud* TPUs sedangkan BERTLARGE membutuhkan 4 hari menggunakan 16 TPUs.

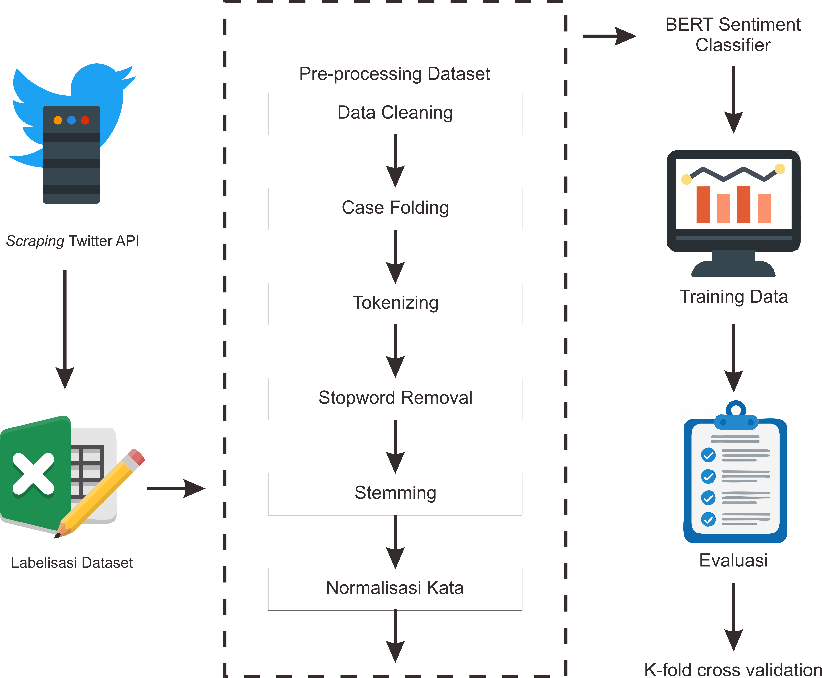
### ***K-Fold Cross-Validation***

*K-fold cross-validation* adalah teknik yang dapat digunakan apabila memiliki jumlah data yang terbatas (jumlah instance tidak banyak). *K-fold cross validation* merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. *K-fold cross validation* diawali dengan membagi data sejumlah *n-fold* yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan ukuran yang sama D1, D2, D3.. Dn selanjutnya proses testing dan training dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data testing dan sisanya akan menjadi data training. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan *10-fold cross validation* dalam model (Kohavi, 2001).

# **BAB III**

# **METODOLOGI PENELITIAN**

## Arsitektur Umum

Metode yang diajukan dalam Pengukuran Indeks Kebahagiaan Masyarakat Kota Semarang Pada Media Sosial Twitter Menggunakan Pendekatan BERT Dengan *K-Fold Cross-Validation* terdiri dari beberapa langkah seperti yang tertera pada Gambar 3.1.

Gambar 3.1 Arsitektur Umum Sistem

Arsitektur umum sistem di atas menunjukkan bahwa analisis sentimen dimulai dengan melakukan scraping atau ekstraksi data twitter masyarakat Semarang berdasarkan wilayah per kecamatan kota Semarang menggunakan twitter API. Hasil dari scraping kemudian dikumpulkan dan menjadi sebuah dataset. Dataset tersebut dipisah menjadi per kalimat. Setelah itu, dataset kemudian dianotasi dengan label negatif atau positif. Kemudian, dataset yang sudah dianotasi masuk ke dalam tahap *preprocessing*. *Preprocessing* dataset adalah tahap untuk mempersiapkan data yang awalnya tidak terstruktur menjadi data yang lebih terstruktur dengan melakukan beberapa tahapan, yaitu *data cleaning*, *case folding*, *tokenizing*, *stopword* *removal*, *stemming,* dan normalisasi kata. Dataset yang sudah melalui proses proses tersebut dilatih untuk diklasifikasi menjadi tiga kategori, yaitu negatif, netral, dan positif menggunakan BERT. Hasil klasifikasi kemudian dievaluasi melihat hasilnya dan dilakukan validasi menggunakan *k-fold cross-validation***.**

## Pengumpulan Data (*scraping)*

Penelitian ini menggunakan data yang diperoleh dari Twitter API yang mengambil data twitter tiap kecamatan di kota Semarang menggunakan *geocode*. Dataset kemudian disimpan lalu melalui proses *sentence splitting* untuk memisahkan data dalam teks menjadi per kalimat. Kalimat dipisah menggunakan salah satu fungsi pada nltk, yaitu sent\_tokenize. *Sentence splitting* dilakukan karena seringkali pada suatu tweet terdapat lebih dari dua kalimat dan memiliki sentimen yang berbeda. Dataset yang sudah melalui proses tersebut kemudian disimpan dalam bentuk *excel worksheet* (.xlsx).

## Labelisasi Dataset

Dalam analisis sentimen dengan metode *supervised learning*, diperlukan dataset yang sudah memiliki label atau dianotasi oleh *annotator*. Labelisasi ini perlu dilakukan karena metode *supervised learning* membutuhkan contoh. Esensi dari *supervised learning* adalah membuat suatu mekanisme di mana model dapat melihat contoh dan menghasilkan generalisasi sehingga keluaran dari model adalah prediksi yang sesuai dengan label yang diinginkan (Goldberg, 2017). Model pun dapat melihat, memahami, dan mengerti bagaimana tweet yang memiliki sentimen negatif, netral, dan positif. Labelisasi yang dilakukan bertujuan untuk menentukan komentar ke dalam kategorinya yaitu negatif, netral, atau positif dengan memberi nilai sebagai penandanya. Komentar dengan sentimen positif diberi nilai 2. Sedangkan komentar dengan sentimen netral diberi nilai 1 dan sentimen negatif diberi nilai 0. Anotasi ini dilakukan oleh tim annotator sebanyak 5 orang.

## *Pre-processing Dataset*

Pada penelitian ini, *preprocessing* dilakukan untuk mengubah dataset yang tidak terstruktur menjadi terstruktur sehingga mempermudah data untuk diproses dengan melakukan beberapa tahapan yaitu *data cleaning*, *case folding*, *tokenizing*, *stopword* *removal*, *stemming,* dan normalisasi kata. Selain itu, dengan preprocessing maka hasil dari analisis sentimen akan lebih baik. Tahap-tahap yang dilakukan meliputi :

* + 1. ***Data cleaning***

Pada tahap ini, kalimat-kalimat pada dataset dibersihkan dari segala sesuatu yang dapat mempengaruhi hasil dari analisis seperti kata dengan karakter yang berulang dua atau lebih, *link*, *username* (@*username*), *hashtag* (#), angka, simbol-simbol, spasi berlebih, tanda baca, dan angka. Untuk melakukan data cleaning, penulis menggunakan regular expression untuk dicocokkan dengan yang akan dihapus.

* + 1. ***Case folding***

Case folding dilakukan dengan membuat semua huruf besar (*uppercase*) pada dataset menjadi huruf kecil (*lowercase*). Tahap ini dilakukan agar semua karakter pada dataset menjadi sama, yaitu menggunakan huruf kecil. Dengan membuat semua kata menjadi huruf kecil akan sangat membantu untuk melakukan generalisasi (Jurafsky & Martin, 2008). Sehingga kata “Saya” dan “saya” akan dianggap sama.

* + 1. ***Tokenizing***

Tokenisasi adalah suatu proses yang dilakukan untuk memecah kalimat-kalimat menjadi potongan kata-kata, tanda baca, dan ekspresi bermakna lainnya sesuai dengan ketentuan Bahasa yang digunakan.

* + 1. ***Stopword* *removal***

*Stopwords Removal* adalah proses yang dilakukan untuk menghapus kata-kata yang tidak memiki arti. Tahap ini akan menggunakan library stopwords Bahasa Indonesia yang disediakan oleh NLTK.

* + 1. ***Stemming***

*Stemming* adalah proses yang dilakukan untuk mengubah kata yang memiliki imbuhan menjadi kata dasarnya (*root form*) dengan menghapus imbuhan seperti prefiks, sufiks, dan konfiks. Pada tahap ini, stemming dilakukan dengan menggunakan *library* Sastrawi.

* + 1. **Normalisasi kata**

Tahap normalisasi adalah tahap di mana dataset yang memiliki kata-kata tidak baku diubah menjadi kata yang baku atau sesuai dengan ejaan. Hal ini dilakukan karena pada umumnya banyak sekali yang menggunakan kata-kata gaul seperti: bgt, trus, slalu, pngn, aq, kereeen, kereenn, dan lain-lain. Jika kata-kata tersebut tidak melewati proses normalisasi, maka sistem akan menganggap kata kereeeen, kereenn, kereennn, kerenn adalah kata yang berbeda. Padahal seharusnya kata tersebut memiliki makna yang sama yaitu keren. Oleh karena itu, normalisasi dilakukan untuk mengubah kata tidak baku menjadi kata baku.

## Implementasi BERT

Pada penelitian ini, penulis menggunakan BERTBASE dan model *bert-multilingual-base-cased.* Model ini dipilih karena model tersebut mendukung 104 bahasa termasuk Bahasa Indonesia. Pada penelitian ini, penulis menggunakan *library* Transformers yang disediakan oleh HuggingFace. *Library* ini menyediakan ribuan *pre-trained* *model* yang dapat digunakan untuk melakukan tugas-tugas klasifikasi, ekstraksi informasi, tanya jawab, *summarization*, translasi, *text generation* dan lain-lain dalam 100 bahasa. *Transformers* didukung oleh dua *library* *deep learning* yang terkemuka yaitu PyTorch dan TensorFlow.

## Evaluasi

Tahapan evaluasi ditujukan untuk melihat hasil analisis sentimen terhadap kalimat yang ada pada dataset. Nilai akurasi tertinggi yang didapatkan dari proses *training* sebelumnya akan menjadi nilai akurasi model. Untuk memperoleh prediksi dari model, *confusion matrix* digunakan sebagaimana terlihat pada Tabel 3.1.

Tabel 3.2 Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | True Class | | |
| Positive | Neutral | Negative |
| Predicted Class | Positive | True Positive (TP) | False Positive (FNt) | False Positive (FP) |
| Neutral | False Neutral (FNt) | True Neutral (TNt) | False Neutral (FNt) |
| Negative | False Negative  (FN) | False Negative  (FNt) | True Negative  (TN) |

Kategori pada *confusion matrix* terdiri dari empat kategori, yaitu *True Positive* (TP), *False Positive* (FP), *True Neutral* (TNt), *False Neutral* (FNt), *True Negative* (TN), dan *False Negative* (FN). *True Positive* adalah kalimat memiliki sentimen positif dan hasil prediksinya juga menunjukkan sentimen positif. *False Positive* (FP) adalah kalimat yang memiliki sentimen positif tetapi hasil prediksinya menunjukkan sentimen netral atau negatif. *True Neutral* (TNt) adalah kalimat memiliki sentimen netral dan hasil prediksinya juga menunjukkan sentimen netral. *False Neutral* (FNt) adalah kalimat yang memiliki sentimen netral tetapi hasil prediksinya menunjukkan sentimen positif atau negatif. *True Negative* (TN) adalah kalimat yang memiliki sentimen negatif dan hasil prediksinya juga menunjukkan hasil negatif. *False Negative* (FN) adalah kalimat yang memiliki sentimen negatif tetapi hasil prediksinya menunjukkan sentimen netral atau positif. Setelah mendapatkan nilai untuk *confusion matrix*, nilai *accuracy*, *precision*, *recall*, dan *F-measure* juga dapat diperoleh. Accuracy bertujuan untuk menunjukkan persentasi dari input yang berhasil diprediksi oleh neural network dengan benar. Nilai akurasi akan semakin baik ketika nilai *loss* menurun. *Precision* bertujuan untuk menghitung persentase dari input yang dideteksi oleh sistem misalnya, sistem memberi label input sebagai positif yang pada aslinya adalah positif juga. *Recall* bertujuan untuk menghitung persentase dari input yang diidentifikasi *True* secara benar oleh sistem. Sedangkan *F-measure* adalah rata-rata yang diperoleh dari *precision* dan *recall*. Rumus perhitungan untuk mendapatkan *accuracy* dan *F-score* ditunjukkan pada persamaan 3.1 dan persamaan 3.4. Sedangkan rumus untuk perhitungan precision dan recall dilakukan untuk tiap sentimen dengan contoh seperti pada persamaan 3.2 dan 3.3.

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 = ...........................................................................(3.1)

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒 = …............................................................................................(3.2)

𝑅𝑒𝑐𝑎𝑙𝑙𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒 = .........................................................................................(3.3)

𝐹 − 𝑚𝑒𝑎𝑠𝑢𝑟𝑒 = 2 ∗ . ................................................................................(3.4)

## Validasi *k-fold cross-validation*

Tahap terakhir yaitu pengujian, pengujian dilakukan menggunakan metode *k-cross validation* dengan nilai k sebanyak 10 *fold*, pengujian ini bertujuan untuk mengetahui akurasi metode BERT yang diterapkan pada analisis sentimen jika diuji dengan *data training* dan *data testing* yang berbeda. Penggunaan 10 fold ini dianjurkan karena merupakan jumlah fold terbaik untuk uji validitas.

# **BAB IV**

# **RENCANA JADWAL PENELITIAN**

## Jadwal Penelitian

Tabel 4.1. Jadwal Penelitian

| NO | JENIS  KEGIATAN | Juli 2021 | | | | Agustus 2021 | | | | September 2021 | | | | | Oktober 2021 | | | | | November 2021 | | | | | Desember 2021 | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | | 2 | 3 | 4 | 1 | | 2 | 3 | 4 | 1 | | 2 | 3 | 4 |
| 1 | Mengumpulkan Data |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 2 | Labelisasi Dataset |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 3 | *Pre-processing* Dataset |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 4 | Implementasi BERT |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 5 | Evaluasi Hasil Klasifikasi |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 6 | Validasi Menggunakan *Cross-validation* |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 7 | Tahapan Selesai |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 8 | Evaluasi Bulanan |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 9 | Seminar Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 10 | Revisi Perbaikan (setelah Seminar Proposal) |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 11 | Fixsasi |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 12 | Sidang Skripsi |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |
| 13 | Selesai |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |

# **DAFTAR PUSTAKA**

Aggarwal, C. (2018). *Neural Networks and Deep Learning: A Textbook*. https://doi.org/10.1007/978-3-319-94463-0

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2018). Hate speech detection in the Indonesian language: A dataset and preliminary study. *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, *2018*-*Janua*(October), 233–237. https://doi.org/10.1109/ICACSIS.2017.8355039

Chollet, F. (2017). *Deep Learning with Python* (1st ed.). Manning Publications Co.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Ghosh, S., Roy, S., & Bandyopadhyay, S. K. (2012). A tutorial review on Text Mining Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(4), 223–233. www.ijarcce.com

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *10*, 1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Habibi, R., Setyohadi, D. B., & Wati, E. (2016). Analisis Sentimen Pada Twitter Mahasiswa Menggunakan Metode Backpropagation. *Jurnal Informatika*, *12*(1), 103–109. https://doi.org/10.21460/inf.2016.121.462

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).

Kohavi, R. (2001). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. *14*.

Kramer, A. D. I. (2010). An unobtrusive behavioral model of “gross national happiness.” *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 287–290. https://doi.org/10.1145/1753326.1753369

Lin, Y., Wang, X., & Zhou, A. (2016). Opinion spam detection. *Opinion Analysis for Online Reviews*, *May*, 79–94. https://doi.org/10.1142/9789813100459\_0007

Liu, I., & Sari, Y. A. (2019). Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(5), 4914–4922.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Studies in Computational Intelligence*, *881 SCI*, 928–940. https://doi.org/10.1007/978-3-030-36687-2\_77

Osinga, D. (2018). Deep Learning Cookbook. In *Book* (Issue June).

Putri, C. A. (2020). Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *6*(2), 181–193. https://doi.org/10.35957/jatisi.v6i2.206

Rahayu Ponilan, I., Herdiani, A., & Selviandro, N. (2016). *Pengukuran Happiness Index Masyarakat Kota Bandung pada Media Sosial Twitter Menggunakan Pendekatan Ontologi Top-Down Hierarchy*. *September*, 17–22. https://doi.org/10.21108/indosc.2016.113

Sidiq, H. K., Kusumo, D. S., & Sardi, I. L. (2019). Mendeteksi Cyberhate pada Twitter Menggunakan Text Classification dan Crowdsourced Labeling. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(4), 315. https://doi.org/10.22146/jnteti.v8i4.530

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017*-*Decem*(Nips), 5999–6009.

Zulfa, I., & Winarko, E. (2017). Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(2), 187. https://doi.org/10.22146/ijccs.24716

# **LAMPIRAN**

**Lampiran 1. *Problem Tree* Penelitian**