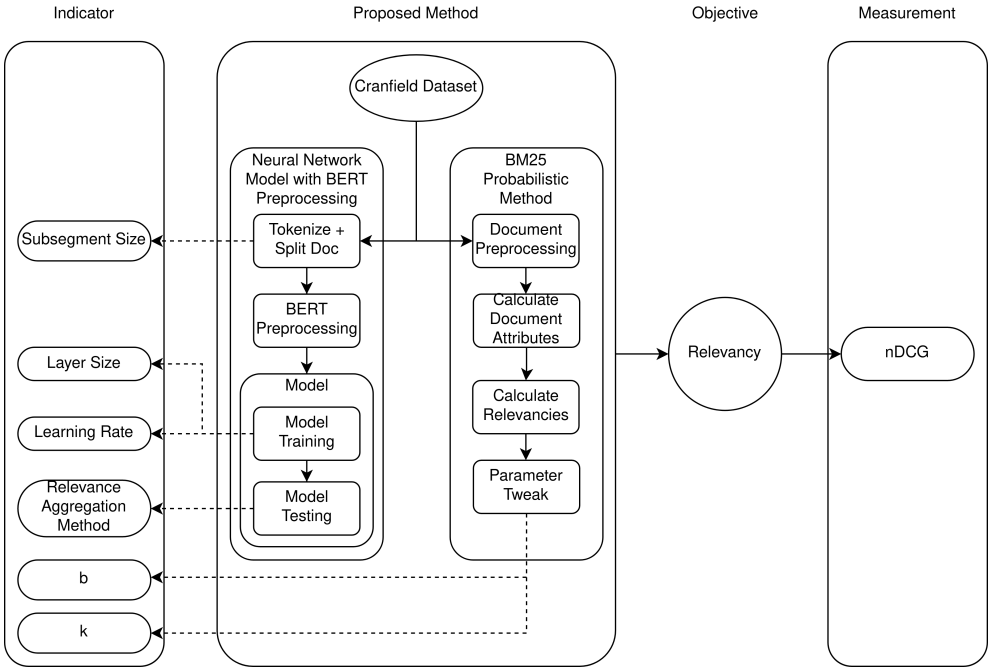
BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini memaparkan analisis masalah yang diatasi beserta pendekatan dan alur kerja dari perangkat lunak yang dikembangkan, mengimplementasikan metode yang digunakan dan hasil yang ditampilkan.

3.1 Analisis Masalah

Penelitian ini bertujuan untuk membangun STBI menggunakan model BERT yang didapat dari tfhub.dev tanpa melakukan *fine-tuning*. Metode yang digunakan berawal dari *document* dan *query preprocessing* menggunakan model BERT *base* dan menggunakan hasilnya sebagai *input* model *neural network*. Hasil penelitian ini merupakan sistem temu balik dokumen yang memakan waktu dan sumber daya yang kecil.

3.2 Kerangka Pemikiran



Gambar 3.1 Kerangka Pemikiran

Gambar 3.1 memperlihatkan metode yang digunakan untuk penelitian yaitu: Metode Probabilistik BM25 dan Metode *Neural Network* dengan BERT *Preprocessing*. Kedua metode melakukan *preprocessing* terhadap data dengan cara yang berbeda. Hasil yang diberikan dari setiap metode adalah nilai relevansi antara dokumen dan *query*.

Indikator merupakan variabel-variabel yang diuji dalam penelitian ini. Variabel-variabel tersebut mempengaruhi hasil akhir dari kedua metode. Karena kedua metode memiliki struktur yang berbeda, maka kedua metode memiliki indikator yang berbeda.

Metode probabilistik BM25 memiliki 2 indikator yang diubah untuk memberikan hasil relevansi terbaik terhadap dataset yang diberikan. Kedua indikator tersebut adalah sebagai berikut:

1. *k* merupakan *hyperparameter* pertama dari metode BM25. Nilai *k* seringkali berada di antara 0 sampai 3. Nilai *k* dapat bernilai lebih dari 3 apabila diperlukan. Nilai *k* digunakan untuk memberikan *weight* terhadap kemunculan kata. Apabila nilai *k* menjadi lebih besar, maka kata-kata yang kemunculannya lebih banyak lebih difokuskan. Apabila *k* bernilai 0, maka jumlah kemunculan sebuah kata tidak dipedulikan. Dengan demikian, apabila *k* bernilai 0, metode hanya melihat apakah sebuah kata ada dalam dokumen atau tidak [3].
2. *b* merupakan *hyperparameter* kedua dari metode BM25. Nilai *b* berada di antara 0 dan 1. Nilai *b* berfokus pada panjang dokumen dibandingkan dengan panjang rata-rata semua dokumen. Apabila nilai *b* bernilai 0 maka panjang dokumen dibandingkan panjang rata-rata semua dokumen tidak diperhitungkan. Sementara nilai 1 artinya panjang dokumen dibandingkan panjang rata-rata semua dokumen diperhitungkan secara keseluruhan [3].

Metode *neural network* dengan BERT *preprocessing* memiliki 4 indikator. Indikator yang digunakan adalah sebagai berikut:

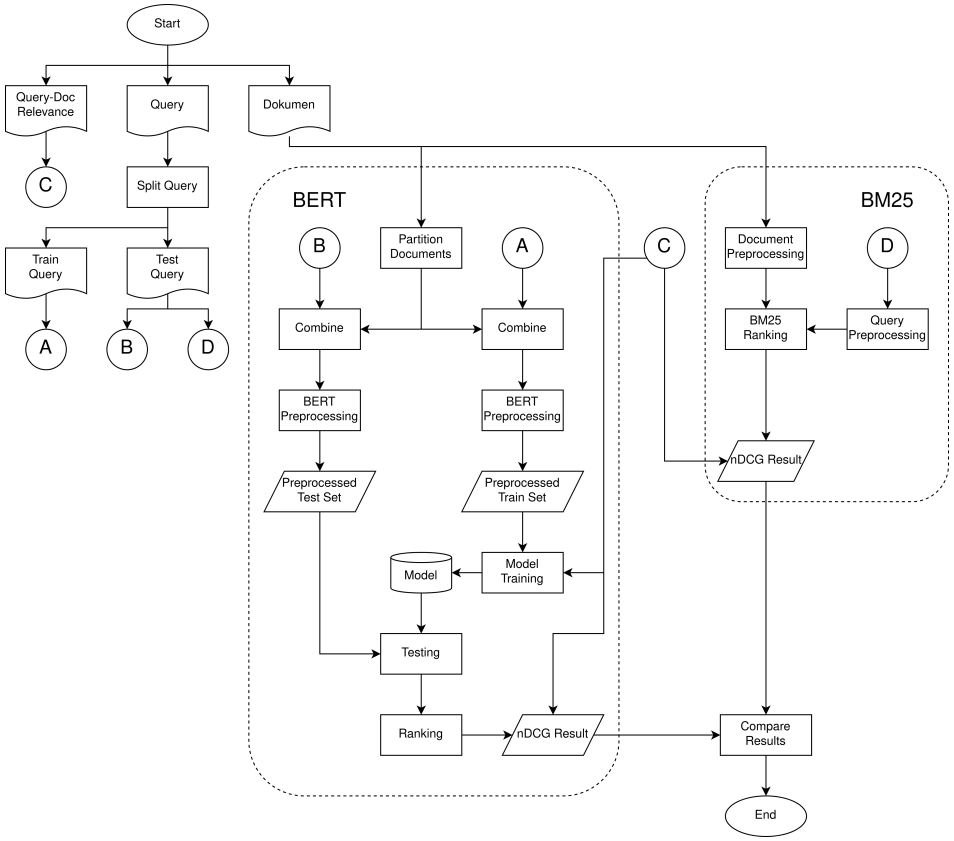
1. Ukuran *subsegment* merupakan indikator yang mempengaruhi jumlah *subsegment* yang dihasilkan pada saat *preprocessing* data. Indikator tersebut diperlukan karena BERT kesulitan mengolah data dengan jumlah *token* yang lebih besar dari 512. Ukuran *subsegment* yang kecil memberi dampak buruk terhadap kemampuan BERT untuk mempelajari konteks. Selain itu, ukuran *subsegment* kecil membuat data latih semakin banyak. Jumlah data yang banyak membantu proses *training* sehingga model *neural network* menjadi lebih akurat. Karena ukuran *subsegment* yang besar dan kecil memiliki kelebihan masing-masing, maka ukuran *subsegment* digunakan sebagai indikator dalam penelitian.
2. Ukuran *layer* merupakan indikator jumlah *neuron* yang digunakan pada setiap *layer* dalam arsitektur *neural network* yang digunakan. Ukuran optimal sebuah layer tidaklah sama untuk setiap *neural network* serta berada di antara ukuran

input dan ukuran *output*. Jumlah *neuron* yang banyak meningkatkan hasil dari model. Kekurangannya, jumlah *neuron* yang besar berpotensi untuk membuat model menjadi *overfit* terhadap data dan proses pelatihan membutuhkan *resource* yang lebih besar. Maka dalam penelitian, jumlah neuron dalam setiap *layer* menjadi indikator [15].

3. *Learning rate* merupakan nilai yang digunakan dalam fungsi *optimizer* untuk menentukan seberapa cepat model *neural network* belajar. Nilai *learning rate* yang besar mempercepat model untuk mendekati hasil yang optimal. Kekurangannya, nilai yang besar berpengaruh agar nilai optimal dilewatkan oleh model. Nilai optimal yang lebih baik ditemukan dengan menggunakan *learning rate* yang lebih kecil. Kekurangannya, proses tersebut memakan waktu yang lebih lama dalam tahap *training*. Karena *learning rate* yang besar dan kecil mempengaruhi model, maka *learning rate* menjadi indikator dalam penelitian.
4. Metode penggabungan nilai relevansi merupakan metode yang digunakan untuk menghitung nilai relevansi dokumen. Dalam tahap *preprocessing*, dokumen yang besar dipecah menjadi beberapa *subsegment*. Model *neural network* yang dilatih menghasilkan nilai relevansi antara *subsegment* dan *query*. Metode digunakan untuk menggabungkan nilai-nilai *subegment* dari dokumen yang sama menjadi satu. Dalam penelitian sebelumnya [1], terdapat beberapa metode yang digunakan. Penelitian ini menambah beberapa metode lain melihat metode yang paling optimal dalam menggabungkan nilai-nilai relevansi dari beberapa *subsegment* sebuah dokumen menjadi satu.

Objektif dari penelitian ini adalah sebuah STBI yang memberikan nilai relevansi yang akurat antara dokumen dan query yang diberikan. Nilai relevansi yang digunakan dalam dataset adalah nilai angka yang berada di antara 0 sampai 4. Karena nilai relevansi yang ada dalam dataset bukan boolean, maka nilai yang digunakan untuk mengukur hasil sistem adalah *normalized discounted cumulative gain* (nDCG). Secara spesifik, nilai nDCG yang diambil hanya melihat 20 dokumen pertama dalam hasil pemberian peringkat (nDCG@20).

3.3 Urutan Proses *Global*



Gambar 3.2 Urutan Proses Global

Diagram pada Gambar 3.2 memperlihatkan urutan proses global dari penelitian. *Dataset Cranfield Experiment* berisi seluruh dokumen yang digunakan, *query* yang digunakan, dan nilai relevansi antara dokumen dan *query* yang memiliki jangkauan nilai dari 0-4. Nilai relevansi tersebut digunakan sebagai *golden answer* dari *dataset*. Dalam penelitian ini, *query* yang digunakan dipecah menjadi 2 *subset* yaitu *subset training* dan *subset testing*.

3.3.1 Proses *Global* Metode BM25

Dataset yang didapat diolah agar siap digunakan. Untuk metode BM25, dokumen dan *query* dibersihkan, simbol dan *stopwords* dibuang, kata-katanya diubah ke bentuk dasar (*lemmatization*), dan kata-kata tersebut disederhanakan dengan metode *stemming*.

Setelah dokumen dan *query* dibersihkan, nilai relevansi antar dokumen dan *query* dikalkulasikan dengan fungsi BM25. Fungsi tersebut memiliki dua parameter bebas *b* dan *k* yang diuji nilai optimalnya. Dokumen-dokumen diberikan peringkat berdasarkan nilai relevansi tersebut. Dari urutan dokumen tersebut dikalkulasi nilai nDCG. Nilai tersebut dibandingkan dengan nilai nDCG

dari hasil metode *neural network* dengan BERT *preprocessing*.

3.3.2 Proses *Global* Metode *Neural Network* dengan BERT *Preprocessing*

Untuk metode *neural network* dengan BERT *preprocessing*, *dataset* yang didapat diolah agar siap digunakan sebagai data latih model. Pasangan *query* dan dokumen diubah menjadi *label* pada proses *training*. Setelah itu, dokumen-dokumen yang digunakan dipecah menjadi beberapa bagian. Hal ini dilakukan karena BERT dibatasi untuk mengolah 512 *token* sebelum *resource* yang dibutuhkan menjadi sangat berat. Penelitian menguji panjang *subsegment* yang tepat untuk dokumen agar menghasilkan hasil yang terbaik. Dokumen-dokumen yang telah dipecah digabungkan dengan judul dokumen apabila dokumen tersebut memiliki judul. Kemudian, dokumen-dokumen yang telah dipartisi digabung dengan *query* dan diubah menjadi *token* yang dimasukkan ke dalam model BERT untuk membuat vektor representasi dari pasangan dokumen dan *query*.

Model yang dilatih menerima sebuah masukan yaitu representasi vektor dari *subsegment* sebuah dokumen dan *query*. Masukkan tersebut kemudian digabungkan ke dalam *deep neural network* yang terdiri dari beberapa *layer*. *Layer* terakhir dari arsitektur berupa *layer sigmoid* yang memberikan nilai antara 0 dan 1 sebagai nilai relevansi antara *subsegment* dokumen dan *query*. Nilai-nilai dari berbagai *subsegment* yang merupakan bagian dari satu dokumen digabungkan agar didapat nilai relevansi yang merepresentasikan seluruh dokumen. Penelitian menguji metode penggabungan relevansi dari *subsegment* yang memberikan nilai terbaik.

Evaluasi model dilakukan dengan *query validation* dan *testing*. Proses evaluasi satu *query* dilaksanakan dengan memasukkan nilai *embedding* pasangan *subsegment* dokumen dan *query* yang telah dipreproses oleh BERT ke dalam model. model kemudian mengeluarkan nilai yang merepresentasikan relevansi antara *subsegment* dan *query*. Nilai-nilai relevansi dari semua *subsegment* digabungkan berdasarkan dokumen asalnya. Kemudian, dokumen-dokumen tersebut diurutkan berdasarkan nilai relevansi. Setelah hasil urutan dokumen tersebut sudah didapat, nilai evaluasi nDCG dikalkulasi untuk melihat performa model terhadap satu *query*. Hasil nDCG dari setiap *query* dari *test set* dirata-ratakan untuk mengkalkulasikan nilai nDCG dari keseluruhan *test set*. Nilai nDCG yang didapat dibandingkan dengan nilai nDCG dari metode BM25.

3.4 Analisis Manual

Pada bagian ini dilakukan analisis tahapan proses dengan melakukan perhitungan manual.

3.4.1 Proses Ranking dengan Metode BM25

Gambar 3.3 memperlihatkan contoh sebuah dokumen dari *dataset*.

Title : piston theory - a new aerodynamic tool for the aeroelastician .
Body : piston theory - a new aerodynamic tool for the aeroelastician . representative applications are described which illustrate the extent to which simplifications in the solutions of high-speed unsteady aeroelastic problems can be achieved through the use of certain aerodynamic techniques known collectively as /piston theory ./ based on a physical model originally proposed by hayes and lighthill, piston theory for airfoils and finite wings has been systematically developed by landahl, utilizing expansions in powers of the thickness ratio and the inverse of the flight mach number m . when contributions of orders and are negligible, the theory predicts a point-function relationship between the local pressure on the surface of a wing and the normal component of fluid velocity produced by the wing's motion . the computation of generalized forces in aeroelastic equations, such as the flutter determinant, is then always reduced to elementary integrations of the assumed modes of motion . essentially closed-form solutions are given for the bending- torsion and control-surface flutter properties of typical section airfoils at high mach numbers . these agree well with results of more exact theories wherever comparisons can be fairly made . moreover, they demonstrate the increasingly important influence of thickness and profile shape as m grows larger, a discovery that would be almost impossible using other available aerodynamic tools . the complexity of more practical flutter analyses-e.g., ... thermoelastic interaction problems .

Gambar 3.3 Contoh Dokumen

Dokumen digabungkan dan dibersihkan isinya agar *token* yang unik semakin sedikit dan *token* yang tidak memberikan informasi tambahan dibuang. Contoh hasil diperlihatkan pada Gambar 3.4.

'piston', 'theori', 'new', 'aerodynam', 'tool', 'aeroelastician', 'piston', 'theori', 'new', 'aerodynam', 'tool', 'aeroelastician', 'repres', 'applic', 'describ', 'illustr', 'extent', 'simplif', 'solut', 'high', 'speed', 'unsteadi', 'aeroelast', 'problem', 'achiev', 'use', 'certain', 'aerodynam', 'techniqu', 'known', 'collect', 'piston', 'theori', 'base', 'physic', 'model', 'origin', 'propos', 'hay', 'lighthil', 'piston', 'theori', 'airfoil', 'finit', 'wing', 'systemat', 'develop', 'landahl', 'util', 'expans', 'power', 'thick', 'ratio', 'invers', 'flight', 'mach', 'number', 'contribut', 'order', 'neglig', 'theori', 'predict', 'point', 'function', 'relationship', 'local', 'pressur', 'surfac', 'wing', 'normal', 'compon', 'fluid', 'veloc', 'produc', 'wing', 'motion', 'comput', 'gener', 'forc', 'aeroelast', 'equat', 'flutter', 'determin', 'alway', 'reduc', 'elementari', 'integr', 'assum', 'mode', 'motion', 'essenti', 'close', 'form', 'solut', 'given', 'bend', 'torsion', 'control', 'surfac', 'flutter', 'properti', 'typic', 'section', 'airfoil', 'high', 'mach', 'number', 'agre', 'well', 'result', 'exact', 'theori', 'wherev', 'comparison', 'fairli', 'made', 'moreov', 'demonstr', 'increasingli', 'import', 'influenc', 'thick', 'profil', 'shape', 'grow', 'larger', 'discoveri', 'would', 'almost', 'imposs', 'use', 'avail', 'aerodynam', 'tool', 'complex', 'practic', 'flutter', 'analysi', 'e', 'g', ... 'thermoelast', 'interact', 'problem'

Gambar 3.4 Contoh dokumen yang telah dibersihkan

Setelah semua dokumen selesai melewati tahap *preprocessing*, informasi-informasi yang ada mengenai dokumen-dokumen tersebut diambil. Gambar 3.5 memperlihatkan contoh informasi seluruh dokumen yang diperlukan untuk analisis manual.

Document Length: 203 Average Document Length: 102.84	
b: 0.5 k: 0.5	Token: progress Frequency: 0 IDF: 4.38
Token: made Frequency: 2 IDF: 1.38	Token: research Frequency: 1 IDF: 3.17
Token: unsteadi Frequency: 2 IDF: 3.40	Token: aerodynam Frequency: 6 IDF: 2.05

Gambar 3.5 Informasi dari dokumen

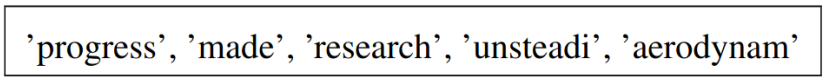
Setelah semua informasi mengenai dokumen didapat, proses pemberian peringkat dimulai. Proses dimulai dengan memberikan sebuah *query* sebagai *input*. Contoh *query* diperlihatkan pada Gambar 3.6.

what progress has been made in research on unsteady aerodynamics .

Gambar 3.6 Contoh *Query*

Query tersebut diolah dengan cara yang sama dengan dokumen menjadi

seperti Gambar 3.7.



Gambar 3.7 *Query* yang telah dibersihkan

Setelah *query* diolah, nilai relevansi sebuah dokumen dikalkulasi. Persamaan BM25 dari Persamaan 2.3 digunakan untuk menghitung nilai relevansi dokumen berdasarkan *query*. Hasil yang didapat adalah 0.0766. Dokumen-dokumen tersebut kemudian diurutkan berdasarkan nilai relevansi tersebut. Tabel 3.1 memperlihatkan contoh lima dokumen teratas berdasarkan nilai relevansi yang didapat.

Tabel 3.1 Lima Dokumen dengan Nilai Relevansi Tertinggi dari Hasil Metode BM25

Document ID	Relevance Score
892	0.57084
899	0.29026
137	0.27565
1109	0.22942
586	0.19855

Dokumen-dokumen yang telah diberi diurutkan dibandingkan dengan nilai relevansi dari *dataset* yang digunakan sebagai *golden answer* untuk mendapatkan nilai performa model. Nilai yang dikalkulasi merupakan nilai *normalized Discounted Cumulative Gain*. Tabel 3.2 memperlihatkan contoh nilai relevansi dari dokumen, Tabel 3.3 memperlihatkan nilai DCG, iDCG, dan nilai nDCG dari 5 dokumen teratas (nDCG@5).

Tabel 3.2 Label Relevansi Dokumen

Dokumen	Nilai Relevansi
199	4
593	3
594	4
892	1
899	3

Tabel 3.2 Label Relevansi Dokumen (lanjutan)

Dokumen	**Nilai Relevansi**
902	1
903	3
1109	2
1289	2

Tabel 3.3 Contoh Perhitungan nDCG

i	**Relevansi (DCG)** Urutan : 892, 899, 137, 1109, 586	**Relevansi (iDCG)** Urutan : 199, 594, 593, 899, 903	**DCG :** $rel_i/log2(i+1)$	**iDCG :** $rel_i/log2(i+1)$
1	1	4	1	4
2	3	4	1.89	2.52
3	0	3	0	1.5
4	2	3	0.86	1.29
5	0	3	0	1.16
	Total		**3.75**	**10.47**
	nDCG @ 5		**0.35**	

Untuk menemukan nilai nDCG dari seluruh *query*, nilai nDCG untuk masing-masing *query* dirata-rata.

3.4.2 *Preprocess* Dokumen dan *Query* dengan BERT

Gambar 3.8 memperlihatkan contoh sebuah dokumen dari *dataset*.

Title : piston theory - a new aerodynamic tool for the aeroelastician .
Body : piston theory - a new aerodynamic tool for the aeroelastician . representative applications are described which illustrate the extent to which simplifications in the solutions of high-speed unsteady aeroelastic problems can be achieved through the use of certain aerodynamic techniques known collectively as /piston theory ./ based on a physical model originally proposed by hayes and lighthill, piston theory for airfoils and finite wings has been systematically developed by landahl, utilizing expansions in powers of the thickness ratio and the inverse of the flight mach number m . when contributions of orders and are negligible, the theory predicts a point-function relationship between the local pressure on the surface of a wing and the normal component of fluid velocity produced by the wing's motion . the computation of generalized forces in aeroelastic equations, such as the flutter determinant, is then always reduced to elementary integrations of the assumed modes of motion . essentially closed-form solutions are given for the bending- torsion and control-surface flutter properties of typical section airfoils at high mach numbers . these agree well with results of more exact theories wherever comparisons can be fairly made . moreover, they demonstrate the increasingly important influence of thickness and profile shape as m grows larger, a discovery that would be almost impossible using other available aerodynamic tools . the complexity of more practical flutter analyses-e.g., ... thermoelastic interaction problems .

Gambar 3.8 Contoh Dokumen

Langkah pertama dalam *preprocessing* dokumen adalah memecah *body* menjadi beberapa *subsegment* yang ukurannya lebih kecil. Untuk itu, seluruh teks dipisah berdasarkan *whitespace* (spasi dan *newline*) seperti pada Gambar 3.9.

['piston', 'theory', '-', 'a', 'new', 'aerodynamic', 'tool', 'for', 'the', 'aeroelastician', '.', 'representative', 'applications', 'are', 'described', 'which', 'illustrate', 'the', 'extent', 'to', 'which', 'simplifications', 'in', 'the', 'solutions', 'of', 'high-speed', 'unsteady', 'aeroelastic', 'problems', 'can', 'be', 'achieved', 'through', 'the', 'use', 'of', 'certain', 'aerodynamic', 'techniques', 'known', 'collectively', 'as', '/piston', 'theory', './', 'based', 'on', 'a', 'physical', 'model', 'originally', 'proposed', 'by', 'hayes', 'and', 'lighthill,', 'piston', 'theory', 'for', 'airfoils', 'and', 'finite', 'wings', 'has', 'been', 'systematically', 'developed', 'by', 'landahl,', 'utilizing', 'expansions', 'in', 'powers', 'of', 'the', 'thickness', 'ratio', 'and', 'the', 'inverse', 'of', 'the', 'flight', 'mach', 'number', 'm', '.', 'when', 'contributions', 'of', 'orders', 'and', 'are', 'negligible,', 'the', 'theory', 'predicts', 'a', 'point-function', 'relationship', 'between', 'the', 'local', 'pressure', 'on', 'the', 'surface', 'of', 'a', 'wing', 'and', 'the', 'normal', 'component', 'of', 'fluid', 'velocity', 'produced', 'by', 'the', "wing's", 'motion', '.', 'the', 'computation', 'of', 'generalized', 'forces', 'in', 'aeroelastic', 'equations,', 'such', 'as', 'the', 'flutter', 'determinant,', 'is', 'then', 'always', 'reduced', 'to', 'elementary', 'integrations', 'of', 'the', 'assumed', 'modes', 'of', 'motion', '.', 'essentially', 'closed-form', 'solutions', 'are', 'given', 'for', 'the', 'bending-', 'torsion', 'and', 'control-surface', 'flutter', 'properties', 'of', 'typical', 'section', 'airfoils', 'at', 'high', 'mach', 'numbers', '.', 'these', 'agree', 'well', 'with', 'results', 'of', 'more', 'exact', 'theories', 'wherever', 'comparisons', 'can', 'be', 'fairly', 'made', '.', 'moreover,', 'they', 'demonstrate', 'the', 'increasingly', 'important', 'influence', 'of', 'thickness', 'and', 'profile', 'shape', 'as', 'm', 'grows', 'larger,', 'a', 'discovery', 'that', 'would', 'be', 'almost', 'impossible', 'using', 'other', 'available', 'aerodynamic', 'tools', '.', 'the', 'complexity', 'of', 'more', 'practical', 'flutter', 'analyses-e.g.,', ... , 'thermoelastic', 'interaction', 'problems', '.']

Gambar 3.9 *tokenized Document*

Kemudian, *subsegment* dari dokumen diperoleh dengan menggabungkan setiap 150 *token* yang ada dan bergeser sejauh 75 *token*. Apabila setelah bergeser ke akhir dokumen dan ukurannya tidak tepat 150 *token*, maka *subsegment* terakhir terdiri dari 150 *token* terakhir. Kumpulan *token* tersebut digabungkan dengan *title* dan menjadi satu *subsegment* dari dokumen. Gambar 3.10 memperlihatkan 2 *subsegment* pertama dari dokumen tersebut. Teks tebal merupakan judul dari dokumen dan teks miring berwarna biru merupakan teks yang *overlap* dari dua *subsegment*.

piston theory - a new aerodynamic tool for the aeroelastician . piston theory - a new aerodynamic tool for the aeroelastician . representative applications are described which illustrate the extent to which simplifications in the solutions of high-speed unsteady aeroelastic problems can be achieved through the use of certain aerodynamic techniques known collectively as /piston theory ./ based on a physical model originally proposed by hayes and lighthill, piston theory for airfoils and finite wings has been systematically developed by landahl, utilizing expansions in powers of *the thickness ratio and the inverse of the flight mach number m . when contributions of orders and are negligible, the theory predicts a point-function relationship between the local pressure on the surface of a wing and the normal component of fluid velocity produced by the wing's motion . the computation of generalized forces in aeroelastic equations, such as the flutter determinant, is then always reduced to elementary integrations of the assumed modes of motion*	**piston theory - a new aerodynamic tool for the aeroelastician .** *the thickness ratio and the inverse of the flight mach number m . when contributions of orders and are negligible, the theory predicts a point-function relationship between the local pressure on the surface of a wing and the normal component of fluid velocity produced by the wing's motion . the computation of generalized forces in aeroelastic equations, such as the flutter determinant, is then always reduced to elementary integrations of the assumed modes of motion .* essentially closed-form solutions are given for the bending- torsion and control-surface flutter properties of typical section airfoils at high mach numbers . these agree well with results of more exact theories wherever comparisons can be fairly made . moreover, they demonstrate the increasingly important influence of thickness and profile shape as m grows larger, a discovery that would be almost impossible using other available aerodynamic tools . the complexity of more practical flutter analyses-e.g.

Gambar 3.10 *Subsegments*

Setelah dokumen dipecah menjadi beberapa *subsegment*, setiap *subsegment* diubah menjadi *token* sebelum masuk ke dalam model BERT. Gambar 3.11 memperlihatkan contoh *subsegment* pertama yang telah diubah menjadi *token*. *token* tersebut disimpan dalam *RaggedTensor* 3 dimensi. Dimensi pertama merepresentasikan setiap dokumen, dimensi kedua merepresentasikan setiap *token*, dimensi ketiga merepresentasikan setiap komponen dari *token*.

```
tf.RaggedTensor [[[16733], [3399], [1011], [1037], [2047], ... , [8290],
[3471], [1012]]]
```

Gambar 3.11 *tokenized Subsegment*

Dimensi ketiga diperlukan karena terdapat kata-kata yang merupakan

perubahan suatu bentuk dari kata dasar. Apabila kata-kata tersebut pernah ditemui model BERT pada proses pelatihan, maka kata tersebut diberikan *token* yang unik. Sebagai contoh, pada Tabel 3.4, kata *Apple* dan *Apples* masing-masing pernah ditemui oleh model BERT saat training. Sehingga masing-masing kata memiliki *token* yang unik. Apabila kata belum pernah ditemui sebelumnya, maka kata tersebut diberikan beberapa *token* yang merepresentasikan satu kata tersebut. Sebagai contoh, Tabel 3.4 memperlihatkan bahwa kata *milestones* terdiri dari dua *token* dimana *token* pertama merupakan *token* yang merepresentasikan *milestone* dan *token* kedua merepresentasikan sifat jamak (berjumlah lebih dari satu) dari kata.

Tabel 3.4 Contoh Kata-kata dan *token* yang Didapat

Kata	***tokens***
Apple	6207
Apples	18108
Milestone	19199
Milestones	19199, 2015

Setelah *subsegment* yang ada telah diubah menjadi *token*, *query* yang digunakan diubah menjadi *token* dengan langkah yang sama. Karena ukuran *query* lebih kecil daripada dokumen, maka *query* tidak dipecah menjadi *subsegment*. Contohnya diperlihatkan pada Tabel 3.5.

Tabel 3.5 *Query tokenization*

Query	what progress has been made in research on unsteady aerodynamics .
tokenize Query	tf.RaggedTensor [[[2054], [5082], [2038], [2042], [2081], [1999], [2470], [2006], [4895, 25647, 2100], [28033, 2015], [1012]]

Setelah *query* dan dokumen diubah menjadi *token* dalam *RaggedTensor*, keduanya digabungkan menjadi *token* dalam satu dimensi agar diolah oleh model. Contohnya diperlihatkan pada Gambar 3.12. *token* pertama 101 merupakan *token* yang merepresentasikan *start sequence*. *token* biru merepresentasikan *query*. *token* 102 pertama merupakan *separator* untuk memisahkan *query* dan *subsegment* dokumen. *token* merah merupakan *token* yang merepresentasikan *subsegment*

dokumen. Terakhir, *token* 102 muncul sekali lagi untuk merepresentasikan *end sequence*. Nilai 0 yang ada di belakang adalah *padding* agar *query* dan *subsegment* dokumen memiliki ukuran yang selalu sama untuk dimasukkan ke dalam model.

[**101**, *101, 2054, 5082, 2038, 2042, 2081, 1999, 2470, 2006, 4895, 25647, 2100, 28033, 2015, 1012* **102**, *16733, 3399, 1011, ... , 8290, 3471, 1012* **102**, 0, 0, 0 ... , 0 ,0 ,0]

Gambar 3.12 *tokenized Query Subsegment Pair*

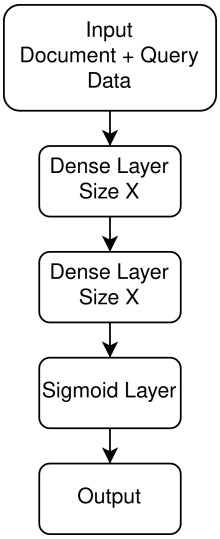
Token yang telah disiapkan dimasukkan ke dalam model BERT untuk mendapatkan nilai yang merepresentasikan *subsegment* dokumen dan *query* yang dimasukkan. Model menghasilkan 768 nilai yang digunakan sebagai data *X* untuk model. Jumlah nilai yang dihasilkan adalah 768 karena model BERT yang digunakan, BERT *Base*, memiliki 768 *neuron* pada setiap *layer*. Hasilnya menjadi seperti Gambar 3.13.

[-0.9942417 , -0.92802274, -0.99996585, ... , -0.99919695, -0.96231484, 0.9887495]

Gambar 3.13 *Neural Network Input*

BERT memberikan dua hasil untuk digunakan. Hasil pertama merupakan nilai *embedding* yang merepresentasikan setiap *token* dalam dokumen. Nilai kedua merupakan hasil *pool* dari hasil pertama. Agar model yang digunakan menjadi lebih ringan, hasil kedua yang dipakai. Hasil tersebut merupakan vektor yang merepresentasikan keseluruhan isi dari *query* dan *subsegment* dokumen. Vektor tersebut dimasukkan ke dalam model *neural network* untuk *training* agar model memberikan nilai relevansi antara *subsegment* dan *query*.

3.4.3 Arsitektur Model



Gambar 3.14 Arsitektur Model

Model yang digunakan untuk STBI memiliki arsitektur seperti Gambar 3.14. Nilai representasi dokumen dan *query* yang didapat dari BERT digunakan sebagai *input* dari model. *Input* dimasukkan ke dalam 2 *layer dense* dengan fungsi aktivasi *relu* dan kemudian masuk ke dalam *sigmoid layer* untuk memprediksi nilai relevansi antara *subsegment* dan *query*. *Output* yang didapat berupa nilai dari 0 hingga 1. Nilai 1 artinya *subsegment* dan *query* 100 % relevan sementara 0 artinya *subsegment* dokumen tidak relevan sama sekali.

3.4.4 Evaluasi Model

Evaluasi model yang telah dilatih dilakukan dengan beberapa langkah. *Metric* evaluasi yang diteliti adalah nilai nDCG dari *test set*. *Test set* yang digunakan adalah nilai representasi pasangan dokumen dan *query* yang digunakan pada *test set*. Nilai representasi tersebut diperoleh dari model BERT sebagai tahap *preprocessing*. Penelitian menguji metode yang tepat untuk mengambil nilai *subsegment* yang digunakan. Untuk penelitian ini, nilai yang digunakan adalah nilai *subsegment* pertama, nilai maksimum dari seluruh *subsegment* sebuah dokumen, dan nilai rata-rata dari seluruh *subsegment* setiap dokumen. Tabel 3.6 memperlihatkan contoh hasil relevansi dan metode penggabungan yang digunakan.

Tabel 3.6 Contoh Nilai Relevansi Dokumen

Subsegment ID	**Relevance Score**	**First Score**	**Max Score**	**Mean Score**	**Doc ID**
753a	0.1565	0.1565	0.1581	0.1573	753
753b	0.1581				
892a	0.1611	0.1611	0.1611	0.1611	892
899a	0.1613	0.1613	0.1613	0.1560	899
899b	0.1507				
902a	0.1616	0.1616	0.1616	0.1375	902
902b	0.1008				
902c	0.1502				
1289a	0.1034	0.1034	0.1617	0.1420	1289
1289b	0.1611				
1289c	0.1617				

Nilai-nilai relevansi yang didapat digunakan untuk memberikan peringkat terhadap dokumen-dokumen. Dokumen dengan nilai relevansi yang lebih besar ditempatkan pada peringkat yang lebih tinggi. Tabel 3.7 memperlihatkan contoh urutan dokumen berdasarkan metode penggabungan yang berbeda.

Tabel 3.7 Proses Mengurutkan Dokumen

Document ID	**First Score**	**Max Score**	**Mean Score**
Document 753	0.1565	0.1581	0.1573
Document 892	0.1611	0.1611	0.1611
Document 899	0.1613	0.1613	0.1560
Document 902	0.1616	0.1616	0.1375
Document 1289	0.1034	0.1617	0.1420
Document Order	Document 902 Document 899 Document 892 Document 753 Document 1289	Document 1289 Document 902 Document 899 Document 892 Document 753	Document 892 Document 753 Document 899 Document 1289 Document 902

Dokumen-dokumen yang telah diberi peringkat dibandingkan dengan nilai relevansi dari *dataset* yang digunakan sebagai *golden answer* untuk mendapatkan nilai performa model. Nilai yang dikalkulasi merupakan nilai *normalized Discounted Cumulative Gain*. Sebagai contoh, Tabel 3.8 memperlihatkan nilai relevansi dari dokumen, Tabel 3.9 memperlihatkan nilai DCG (hasil dari metode penggabungan *subsegment max*), iDCG, dan nilai nDCG dari 5 dokumen teratas (nDCG@5).

Tabel 3.8 Label Relevansi Dokumen

Dokumen	**Nilai Relevansi**
199	4
593	3
594	4
892	1
899	3
902	1
903	3
1109	2
1289	2

Tabel 3.9 Contoh perhitungan nDCG

i	**Relevansi (DCG)** Urutan : 1289, 902, 899, 892, 753	**Relevansi (iDCG)** Urutan : 199, 594, 593, 899, 903	**DCG :** $rel_i/log2(i+1)$	**iDCG :** $rel_i/log2(i+1)$
1	2	4	2	4
2	1	4	0.63	2.52
3	3	3	1.5	1.5
4	1	3	0.43	1.29
5	0	3	0	1.16
	Total		**4.65**	**10.47**
	nDCG @ 5		**0.43**	

Untuk menemukan nilai nDCG dari seluruh *query*, nilai nDCG untuk masing-masing *query* dirata-rata.

Tabel 3.3 dan Tabel 3.9 memperlihatkan bahwa hasil nDCG@5 dari metode *neural network* dengan BERT *preprocessing* memberikan hasil 8% lebih baik *query* dari Tabel 3.5.