**LAPORAN AKHIR PROYEK**

**Multi-label Emotion Classification using Bi-LSTM with GloVe Word Embedding**



**Disusun oleh:**

|  |  |
| --- | --- |
| **12S18004** | **Rosalia Pane** |
| **12S18008** | **Indah Tri Anastasya Manik** |
| **12S18011** | **Nadya Putri Tambunan** |
| **12S18043** | **Roy Gunawan Napitupulu** |
| **12S18048** | **Rifka Uli Siregar** |

**11S4037 – PEMROSESAN BAHASA ALAMI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

# **DAFTAR ISI**

[**DAFTAR ISI** 2](#_heading=h.gjdgxs)

[**DAFTAR TABEL** 3](#_heading=h.30j0zll)

[**DAFTAR GAMBAR** 4](#_heading=h.1fob9te)

[**BAB I PENDAHULUAN** 5](#_heading=h.3znysh7)

[**1.1.**](#_heading=h.2et92p0) **Latar Belakang** 5

[**1.2.**](#_heading=h.tyjcwt) **Tujuan** 7

[**1.3.**](#_heading=h.3dy6vkm) **Manfaat** 7

[**1.4.**](#_heading=h.1t3h5sf) **Ruang Lingkup** 8

[**BAB II ISI** 8](#_heading=h.2s8eyo1)

[**2.1 Analisis** 8](#_heading=h.17dp8vu)

[**2.1.1 Analisis Data** 8](#_heading=h.3rdcrjn)

[**2.1.2 Analisis Metode** 11](#_heading=h.35nkun2)

[**2.2 Desain** 12](#_heading=h.1ksv4uv)

[**2.2.1 Data Preprocessing** 12](#_heading=h.44sinio)

[**2.2.2 Feature Extraction (TF IDF / GloVe)** 12](#_heading=h.2jxsxqh)

[**2.2.3 Feature Selection** 12](#_heading=h.z337ya)

[**2.2.4 Modelling with Bi-LSTM** 12](#_heading=h.3j2qqm3)

[**2.2.5 Evaluation and Results** 12](#_heading=h.1y810tw)

[**2.3 Implementasi** 12](#_heading=h.4i7ojhp)

[**2.3.1 Data Preprocessing** 12](#_heading=h.2xcytpi)

[**2.3.2 Feature Extraction** 12](#_heading=h.1ci93xb)

[**2.3.3 Feature Selection** 12](#_heading=h.3whwml4)

[**2.3.4 Modeling with Bi-LSTM** 12](#_heading=h.2bn6wsx)

[**2.4 Hasil** 13](#_heading=h.3as4poj)

[**2.4.1 Evaluation BI-LSTM Model –tf idf** 13](#_heading=h.1pxezwc)

[**2.4.2 Evaluation BI-LSTM Model – GloVe** 13](#_heading=h.49x2ik5)

[**2.4.3 Evaluation BI-LSTM Model witk** 13](#_heading=h.2p2csry)

[**2.4.4 BI-LSTM Model with** 13](#_heading=h.147n2zr)

[**2.4.5 Accuracy BI-LSTM Model with** 13](#_heading=h.3o7alnk)

[**BAB 3 PENUTUP** 13](#_heading=h.23ckvvd)

[**3.1 Pembagian Tugas dan Tanggung Jawab** 13](#_heading=h.ihv636)

[**3.2 Kesimpulan** 13](#_heading=h.32hioqz)

[**3.3 Saran** 13](#_heading=h.1hmsyys)

[**DAFTAR PUSTAKA** 1](#_heading=h.41mghml)

# **DAFTAR TABEL**

[Tabel 1. Atribut pada dataset 8](#_heading=h.26in1rg)

# **DAFTAR GAMBAR**

[Gambar 1. Distribusi Label pada Dataset 10](#_heading=h.lnxbz9)

# **BAB I PENDAHULUAN**

Bagian ini menyajikan latar belakang, tujuan, manfaat, dan ruang lingkup pengerjaan proyek.

## **Latar Belakang**

Emosi adalah keadaan pikiran yang berlangsung terus-menerus, yang ditandai dengan gejala mental, fisik dan perilaku. Emosi seseorang dapat diidentifikasi secara langsung melalui ekspresi wajah dan ucapannya. Mendeteksi emosi secara otomatis sangat penting karena dapat diterapkan di berbagai bidang. Misalnya dalam dunia pendidikan, analisis emosi dapat dimanfaatkan untuk lingkungan *e-learning*. Selain itu, dalam bisnis yang digunakan untuk mengidentifikasi keluhan pelanggan [1].

Dari pengalaman sehari-hari, beberapa emosi tampaknya berbeda dan terjadi secara independen. Emosi yang secara inheren kontradiktif, seperti love dan hate mungkin memerlukan serangkaian kelas yang terpisah untuk mengakomodasikan aspek dari setiap kelas. Disisi lain, emosi yang identik biasanya berada dibawah valensi emosional yang sama dan sering muncul bersamaan dalam situasi tertentu. Oleh karena itu, berbagai emosi ini dapat dikelompokkan bersama. Deteksi emosi, yang berperan sebagai masalah klasifikasi multi-label dapat membantu menjelaskan sifat kompleks dari emosi yang terjadi bersamaan, sehingga memberikan pemahaman tentang karakteristik setiap emosi [2].

Pendeteksian emosi merupakan salah satu masalah yang muncul di bidang *Natural Language Processing* (NLP). NLP digunakan untuk memproses data seperti teks yang terstruktur maupun tidak menjadi pengetahuan bermakna untuk berbagai masalah bisnis. NLP telah banyak digunakan untuk solusi masalah seperti klasifikasi, pemodelan topik, *text generation*, QA *system*, rekomendasi, dan lainnya[3].

Baru-baru ini, masalah klasifikasi *multi-label* telah menarik minat yang cukup besar karena penerapannya ke berbagai *domain*, termasuk klasifikasi teks, klasifikasi adegan dan video, dan bioinformatika [4]. Berbeda dengan masalah klasifikasi tradisional label tunggal (yaitu, multi-kelas atau biner), di mana sebuah *instance* dikaitkan dengan hanya satu label dari satu set label yang terbatas, dalam masalah klasifikasi *multi-label*, sebuah *instance* dikaitkan dengan *subset* dari label [5].

Pendeteksian emosi menggunakan klasifikasi *multi-label* menjadi masalah karena suatu kalimat cenderung melibatkan lebih dari satu kategori emosi. Sehingga, tantangan utama yang muncul adalah bagaimana memodelkan ketergantungan antar label menggunakan pendekatan klasifikasi. Misalnya, emosi dengan label “*angry*” dan “*disgust*” memiliki ketergantungan daripada emosi “*sad*” dan “*joy*” yang saling bertentangan [6]. Analisis emosi melalui sebuah teks tampaknya juga menjadi tantangan karena faktanya bahwa ekspresi tekstual tidak selalu secara langsung melibatkan kata-kata yang berhubungan dengan emosi, tetapi seringkali suatu kalimat perlu dipahami untuk memberikan sebuah makna [7].

Salah satu pendekatan yang dapat digunakan untuk mengatasi masalah klasifikasi *multi-label* adalah metode *Bidirectional Long Short Term Memory* (Bi-LSTM). Metode Bi-LSTM merupakan perkembangan dari model LSTM dengan dua lapisan, dimana lapisan pertama bergerak maju (*forward*) untuk memahami dan memproses dari kata pertama menuju kata terakhir, demikian sebaliknya lapisan atasnya bergerak mundur (*backward*) untuk memahami dan memproses dari kata terakhir menuju kata pertama. Oleh karena itu, Bi-LSTM sangat baik digunakan untuk mengenali pola dalam kalimat, dikarenakan setiap kata dalam kalimat diproses secara sekuensial [8].

Kemudian dalam melakukan klasifikasi, metode Bi-LSTM akan digabungkan dengan salah satu pendekatan pembobotan kata yaitu *pre-trained word embeddings GloVe*. Dimana, *GloVe* merupakan salah satu pendekatan yang memiliki akurasi yang baik untuk memproses pembobotan kata dalam data dibanding model *word embeddings* lain seperti CBOW dan *skip-grams*. Secara keseluruhan, *GloVe* mengungguli model lain dalam hal analogi kata, kemiripan kata dan tugas *named entity recognition* [9].

Oleh karena itu, berdasarkan uraian dari permasalahan sebelumnya, penulis berfokus pada pengklasifikasian emosi *multi-label*, yang bertujuan untuk mengembangkan sistem otomatis untuk mengkategorikan kalimat ke dalam netral dan 27 emosi seperti *admiration*, *amusement*, *anger* dan emosi lainnya. Penggunaan pendekatan Bi-LSTM dan *word embedding GloVe* akan membantu dalam membentuk matriks *embedding* pada masalah klasifikasi *multi-label* dengan menggunakan *dataset GoEmotions* yang diperoleh dari *Hugging Face.*

## **Tujuan**

Tujuan dari proyek *multi-label emotion classification* ini, antara lain:

1. Menerapkan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dengan *GloVe* sebagai *word embedding* dalam menganalisis teks apakah teks tersebut termasuk ke dalam teks dengan beberapa label (*multi-label*).
2. Untuk mengetahui bagaimana tingkat akurasi menggunakan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dengan *GloVe* sebagai *word embedding* dalam melakukan pengklasifkasian *multi-label*.

## **Manfaat**

Berikut adalah manfaat dari pembuatan *multi-label emotion classification* menggunakan metode Bi-LSTM dan *GloVe* antara lain:

1. Mengetahui cara dan proses klasifikasi *multi-label emotion* menggunakan metode Bi-LSTM dan *GloVe* sebagai *word embedding*.
2. Mengetahui tingkat akurasi menggunakan metode Bi-LSTM dan *GloVe* sebagai *word embedding* dalam melakukan klasifikasi *multi-label emotion*.

## **Ruang Lingkup**

Ruang lingkup dalam pengerjaan proyek ini yaitu menggunakan metode Bi-LSTM dan *word embedding* *GloVe* dengan menggunakan *dataset GoEmotions* yang diperoleh dari *Hugging Face* [10]*.*

# **BAB II ISI**

Pada bab ini mencakup analisis yaitu analisis terhadap data dan analisis terhadap metode.

## **2.1 Analisis**

Pada *subbab* ini dijelaskan analisis yang dilakukan terhadap data dan metode yang digunakan dalam pengimplementasian multi-label klasifikasi emosi.

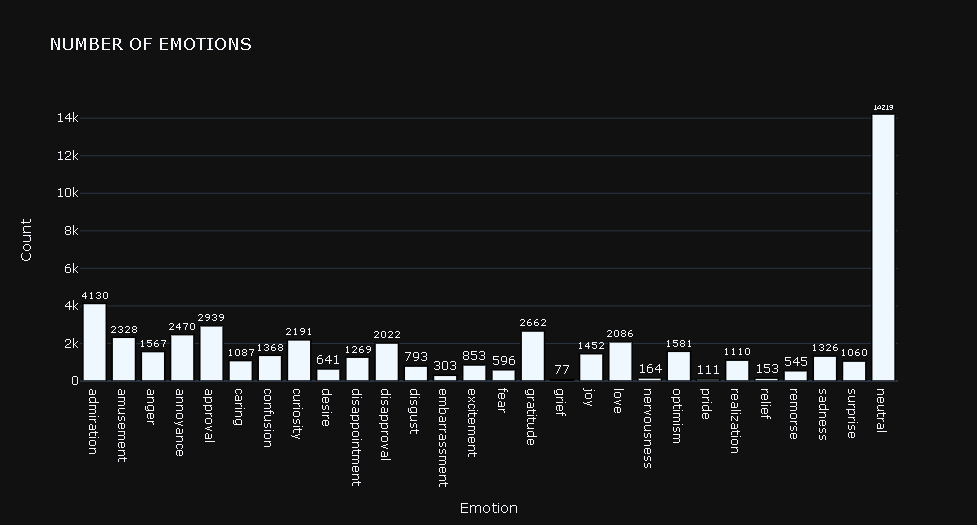
### **2.1.1 Analisis Data**

*Dataset* yang digunakan dalam proyek ini menggunakan *dataset GoEmotions* yang diperoleh dari *Hugging Face* [10]. *Dataset GoEmotions* terdiri dari 43410 baris dan 37 kolom. *Dataset* tersebut telah diberikan beberapa kelas/label yaitu pada setiap teks dalam data tersebut. Pada Tabel 1 berikut menampilkan gambaran dari dataset yang digunakan yang terdiri dari nama atribut, tipe atribut dan keterangan.

Tabel 1. Atribut pada *dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| No | Nama Atribut | Tipe Atribut | Keterangan |
| 1 | *text* | Kategorikal | Komentar yang berasal dari Reddit |
| 2 | *id* | Kategorikal | Berisi id text |
| 3 | *author* | Kategorikal | Nama pengguna penulis komentar di Reddit |
| 4 | *subreddit* | Kategorikal | Subreddit tempat komentar berasal |
| 5 | *link\_id* | Kategorikal | ID tautan komentar |
| 6 | *parent\_id* | Kategorikal | ID *parent* dari komentar |
| 7 | *created\_utc* | Numerik | *Timestamp* komentar |
| 8 | *rater\_id* | Numerik | ID unik dari annotator |
| 9 | *example\_very\_unclear* | Kategorikal | Apakah annotator menandai label yang tidak jelas atau sulit untuk diberi label (dalam hal ini mereka tidak memilih label emosi) |
| 10 | *admiration* | Kategorikal | Berisi pengelompokan *text* antara *admiration* (1) dan *non-admiration* (0) |
| 11 | *amusement* | Kategorikal | Berisi pengelompokan *text* antara *amusement*(1) dan *non-amusement* (0) |
| 12 | *anger* | Kategorikal | Berisi pengelompokan *text* antara *anger*(1) dan *non-anger* (0) |
| 13 | *annoyance* | Kategorikal | Berisi pengelompokan *text* antara *annoyance*(1) dan *non-annoyance* (0) |
| 14 | *approval* | Kategorikal | Berisi pengelompokan *text* antara *approval* (1) dan *non-approval* (0) |
| 15 | *caring* | Kategorikal | Berisi pengelompokan *text* antara *caring* (1) dan *non-caring* (0) |
| 16 | *confusion* | Kategorikal | Berisi pengelompokan *text* antara *confusion*(1) dan *non-confusion* (0) |
| 17 | *curiosity* | Kategorikal | Berisi pengelompokan *text* antara *curiosity* (1) dan *non-curiosity* (0) |
| 18 | *desire* | Kategorikal | Berisi pengelompokan *text* antara *desire* (1) dan *non-desire* (0) |
| 19 | *disappointment* | Kategorikal | Berisi pengelompokan *text* antara *disappointment* (1) dan *non-disappointment* (0) |
| 20 | *disapproval* | Kategorikal | Berisi pengelompokan *text* antara *disapproval* (1) dan *non-disapproval* (0) |
| 21 | *disgust* | Kategorikal | Berisi pengelompokan *text* antara *disgust* (1) dan *non-disgust* (0) |
| 22 | *embarrassment* | Kategorikal | Berisi pengelompokan *text* antara *embarrassment* (1) dan *non-embarrassment* (0) |
| 23 | *excitement* | Kategorikal | Berisi pengelompokan *text* antara *excitement* (1) dan *non-excitement* (0) |
| 24 | *fear* | Kategorikal | Berisi pengelompokan *text* antara *fear* (1) dan *non-fear* (0) |
| 25 | *gratitude* | Kategorikal | Berisi pengelompokan *text* antara *gratitude* (1) dan *non-gratitude* (0) |
| 26 | *grief* | Kategorikal | Berisi pengelompokan *text* antara *grief* (1) dan *non-grief* (0) |
| 27 | *joy* | Kategorikal | Berisi pengelompokan *text* antara *joy* (1) dan *non-joy* (0) |
| 28 | *love* | Kategorikal | Berisi pengelompokan *text* antara *love* (1) dan *non-love* (0) |
| 29 | *nervousness* | Kategorikal | Berisi pengelompokan *text* antara *nervousness* (1) dan *non-nervousness* (0) |
| 30 | *optimism* | Kategorikal | Berisi pengelompokan *text* antara *optimism* (1) dan *non-optimism* (0) |
| 31 | *pride* | Kategorikal | Berisi pengelompokan *text* antara *pride* (1) dan *non-pride* (0) |
| 32 | *realization* | Kategorikal | Berisi pengelompokan *text* antara *realization* (1) dan *non-realization* (0) |
| 33 | *relief* | Kategorikal | Berisi pengelompokan *text* antara *relief* (1) dan *non-relief* (0) |
| 34 | *remorse* | Kategorikal | Berisi pengelompokan *text* antara *remorse* (1) dan *non-remorse* (0) |
| 35 | *sadness* | Kategorikal | Berisi pengelompokan *text* antara *sadness* (1) dan *non-sadness* (0) |
| 36 | *surprise* | Kategorikal | Berisi pengelompokan *text* antara *surprise* (1) dan *non-surprise* (0) |
| 37 | *neutral* | Kategorikal | Berisi pengelompokan *text* antara *neutral* (1) dan *non-neutral* (0) |

Berikut adalah distribusi label pada *dataset* ditunjukkan pada Gambar 1.



Gambar 1. Distribusi Label pada *Dataset*

Setiap label dari dataset *GoEmotions*, selanjutnya akan dilakukan pengklasifikasian dengan di mapping ke dalam 7 label. Berikut daftar label yang digunakan untuk klasifikasi.

* *anger: anger, annoyance, disapproval*
* *disgust: disgust*
* *fear: fear, nervousness*
* *joy: joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring*
* *sadness: sadness, disappointment, embarrassment, grief, remorse*
* *surprise: surprise, realization, confusion, curiosity*
* *neutral: neutral*

### **2.1.2 Analisis Metode**

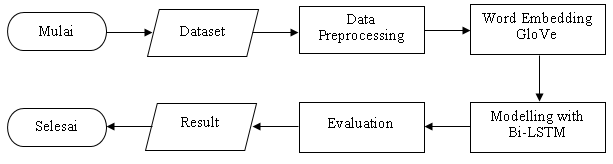
Pada klasifikasi kelas (seperti *binary* dan *multi-class*), *output* yang diterima akan tunggal dari beberapa opsi yang diberikan. *Binary class* akan memberikan kemungkinan kelas "N" sebanyak 2 (N = 2) sementara *multi-class* akan memberi kemungkinan sebanyak N > 2. Berbeda dengan *Multi-label text classification*, jenis pengklasifikasian ini memprediksi beberapa kemungkinan label yang akan dilibatkan dalam teks tertentu. Artinya, *output* yang dihasilkan akan lebih dari satu.

*Bidirectional Long Short-Term Memory* (Bi-LSTM) adalah perkembangan dari model LSTM dimana terdapat dua lapisan yang prosesnya saling berkebalikan arah, model ini sangat baik untuk mengenali pola dalam kalimat karena setiap kata dalam kalimat diproses secara sekuensial. Dengan adanya lapisan dua arah yang saling berlawanan ini maka model dapat memahami dan mengambil perspektif dari kata terdahulu dan kata terdepan, sehingga proses pembelajaran akan semakin dalam yang berdampak pada model akan lebih memahami konteks pada klasifikasi emosi tersebut. *GloVe* merupakan salah satu pendekatan yang memiliki akurasi yang baik untuk memproses pembobotan kata dalam data dibanding model *word embeddings* lain seperti CBOW dan *skip-grams*. Sehingga dalam melakukan klasifikasi, metode Bi-LSTM akan digabungkan dengan salah satu pendekatan pembobotan kata yaitu *pre-trained word embeddings GloVe*.

Analisis *multi-label emotion classification* pada komentar Reddit yang terdapat dalam *dataset GoEmotions* diklasifikasikan dengan metode Bi-LSTM yang akan digabungkan dengan salah satu pendekatan pembobotan kata yaitu *pre-trained word embeddings GloVe*.

## **2.2 Desain**

Pada subbab ini dijelaskan desain pemrosesan bahasa alami yaitu yang ditampilkan dalam bentuk *flowchart* atau diagram alir seperti ditunjukkan pada Gambar



### **2.2.1 *Data Preprocessing***

*Preprocessing* adalah proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur. Tahap preprocessing mengubah data tekstual menjadi data yang siap dijadikan model text mining [15]. Lirik dari dataset yang sudah dibuat dilakukan preprocessing untuk memperbaiki struktur dan menghindari data yang tidak sempurna. Ada beberapa tahapan yang biasa dilakukan pada tahap ini, yaitu *tokenization*, *stop-word removal*, *lowercase conversion*, dan *lemmatization* [16].

#### **2.2.1.1 *Data Cleaning***

Data yang diperoleh dari dataset memiliki beberapa *noise* yang perlu dibersihkan, misalnya string yang kosong (*incomplete data/missing value*). Data cleaning dilakukan ketika data yang diperoleh tidak lengkap (*missing value*), terdapat *error* (*noisy data*) dan juga tidak konsisten. Data cleaning perlu dilakukan karena ketiga masalah di atas dapat mengakibatkan hasil prediksi dalam klasifikasi menjadi tidak akurat. Untuk mengatasi *noisy* data dilakukan beberapa cara yakni *binning*, *regression*, *clustering* dan *semi supervised method* [11].

##### **2.2.1.1.1 Clean Text**

Pada tahap *clean text* dilakukan beberapa proses *data cleaning*, seperti:

* *Clean emoji*

Tahapan ini merupakan tahapan *preprocessing* untuk menghapus emoji karena tidak dapat dianalisis. Dilakukan dengan mengkodekan *string* menggunakan pengkodean ASCII yang selanjutnya akan didekodekan untuk menghapus emoji.

* *Make text lowercase*

*Lower Casing* atau *case folding* atau adalah salah satu tahap dari *preprocessing* untuk *text mining*, dimana semua huruf diubah menjadi huruf kecil untuk mencegah sensitivitas huruf besar-kecil. Dengan cara ini, kita dapat meningkatkan kinerja *classifier* tanpa mempertimbangkan ketidakkonsistenan teks.

* *Remove text in square brackets*

Karakter [ dan ] merupakan karakter khusus dalam regex. Karakter tersebut digunakan untuk menampilkan daftar karakter yang cocok. Proses ini akan menghapus teks dalam kurung {}, kotak [], dan/atau bulat (), serta kurung itu sendiri.

* *Remove links*

Penggunaan *remove links* akan membantu dalam menghapus *url* atau *link* yang terdapat dalam teks. Munculnya *remove links* membuat data tidak efektif dan tidak memiliki arti.

* *Remove punctuation*

Penggunaan *punctuation* (tanda baca) seperti spasi, tanda konvensional atau tipografi tertentu biasanya membantu pembaca untuk memahami teks tertulis. Tetapi dalam pemrosesan data, tanda baca tersebut perlu dihapus untuk menghilangkan bagian data yang tidak membantu, atau *noise*.

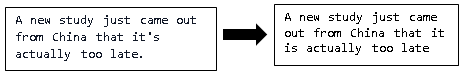
* *Remove words containing numbers*

Selama proses pembersihan data biasanya diperlukan penghapusan angka dari data di *Natural Language Processing*. Misalkan data memiliki string abcd1234efg567, dan dilakukan penghapusan digit/angka dari string untuk mendapatkan string seperti abcdefg.



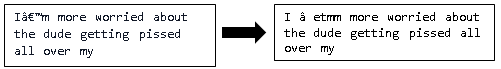
##### **2.2.1.1.2 Clean Contractions**

Kata-kata yang ditulis dengan apostrof (‘) disebut sebagai *contractions*. Tujuannya adalah untuk membakukan teks akar lebih masuk akal. Misalnya: *don’t* menjadi *do not, can’t* menjadi *cannot*.



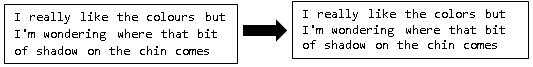
##### **2.2.1.1.3 Clean Special Characters**

Proses ini bertujuan untuk menghapus spesial karakter yang terdiri dari *non alphanumeric* karakter seperti ! ++ << [ % , <=<= ] & — <><> | ‘ . = ~ ( / == ~= ) /! >>// >>= ! { ? ` \* } @ : ; ^ |= &= += -= = /= \*=

****

##### **2.2.1.1.4 Correct Spelling**

Kesalahan ejaan sering terjadi dan telah banyak fitur perangkat lunak yang menyediakan perbaikan dari kesalahan tersebut. Python menawarkan banyak modul yang bertujuan untuk membuat penulisan pemeriksa ejaan sederhana menjadi mudah.

****

##### **2.2.1.1.5 Remove Space**

Terkadang sebuah data sering memiliki karakter spasi di depan, di akhir, atau beberapa karakter spasi yang disematkan dimana karakter ini terkadang bisa menyebabkan hasil yang tidak diharapkan saat mengurutkan, memfilter, atau mencari data. Keberadaan spasi berlebih juga bisa menjadi kendala dalam pengolahan data. Maka pada tahap ini tanda spasi yang berlebih akan dihapus untuk membenahi teks-teks tersebut supaya terlihat lebih rapi dan lebih konsisten yang akhirnya akan mempermudah dalam pengolahan data lebih lanjut.