Multi-Label Emotion Classification using Logistic Regression Method with Dataset GoEmotions

# Proposal Tugas Pemrosesan Bahasa Alami Oleh :

|  |  |
| --- | --- |
| 11S18045 | Gorga Siagian |
| 11S18056 | Sri Intan Sinaga |



11S4037 - Pemrosesan Bahasa Alami Fakultas Informatika dan Teknik Elektro Institut Teknologi Del

2021/2022

## DAFTAR ISI

[DAFTAR ISI 2](#_Toc89274628)

[1.3. Tujuan 4](#_Toc89274629)

[1.4. Manfaat Penelitian 4](#_Toc89274630)

[1.5. Ruang Lingkup Penelitian 4](#_Toc89274631)

[2.1 Preprocessing Data 5](#_Toc89274632)

[2.1.1. *Text Cleaning* 5](#_Toc89274633)

[2.1.2*. Case Folding* 5](#_Toc89274634)

[*2.1.3. Tokenization* 5](#_Toc89274635)

[2.1.4. Stopwords Removal 6](#_Toc89274636)

[2.1.5. Stemming 6](#_Toc89274637)

[2.2 TF-IDF 6](#_Toc89274638)

[2.3 *Logistic Regression* 6](#_Toc89274639)

[2.4 Klasifikasi 7](#_Toc89274640)

[2.5 Sosial Media 7](#_Toc89274641)

**BAB 1 PENDAHULUAN**

* 1. **Latar Belakang Masalah**

Emosi dapat diidentifikasi dari berbagai sumber seperti teks, ekspresi wajah, gambar, pidato, lukisan, lagu, dan sebagainya. Dengan semakin populernya media sosial online, orang-orang suka mengekspresikan emosi mereka atau berbagi acara yang bermakna dengan orang lain di platform jejaring sosial seperti *twitte*r, *facebook*, catatan pribadi, blog, novel, email, pesan obrolan, dan berita utama .Emosi adalah perasaan kuat yang berasal dari suasana hati atau interaksi seseorang satu sama lain.

Deteksi emosi di jejaring sosial online menguntungkan banyak aplikasi seperti layanan iklan yang dipersonalisasi, sistem saran, dll. Emosi dapat diidentifikasi dari berbagai sumber seperti teks, ekspresi wajah, gambar, pidato, lukisan, lagu, dll. Deteksi emosi dapat dilakukan dengan berbagai teknik dalam pembelajaran mesin. Teknik deteksi emosi berfokus pada klasifikasi multi-kelas sambil mengabaikan konsistensi beberapa label emosi dalam satu contoh pada komputer melalui pembelajaran mesin(*machine learning*).Komputer dapat memantau emosi pengguna untuk menyarankan musik atau film yang sesuai dalam interaksi komputer manusia . Selain itu, output dari sistem penambangan emosi dapat berfungsi sebagai input ke sistem lain

Deteksi emosi adalah masalah klasifikasi multi-label yang membutuhkan prediksi beberapa emosi skor dari urutan data yang diberikan. Setiap data urutan yang diberikan dapat memiliki lebih dari satu emosi, sehingga masalahnya dapat diajukan sebagai masalah klasifikasi multi-label daripada sebagai masalah klasifikasi multi-kelas. Pembelajaran mesin dan pembelajaran mendalam digunakan dalam hal ini penelitian untuk memecahkan masalah tersebut.Pada tugas proyek pemrosesan bahasa alami kali ini,kami melakukan implementasi untuk deteksi emosi pada dataset GoEmotion yang disediakan oleh Hugging Face.

Pada dataset tersebut diambil dari komentar reddit, dimana implementasi yang kami lakukan deteksi emosi dari dataset tersebut dalam bentuk emosi positif, negatif, dan netral. Untuk model yang digunakan untuk melakukan klasifikasi deteksi emosi tersebut menggunakan *Logistic Regression* dimana model klasifikasi *Logistic Regression* tersebut dapat memodelkan probabilitas untuk masalah klasifikasi dengan dua kemungkinan hasil. *Logistic Regression* merupakan perpanjangan dari model linear regression untuk masalah klasifikasi.

* 1. **Rumusan Masalah**

Berdasarkan Latar Belakang masalah yang telah disampaikan maka rumusan masalah yang diperoleh yaitu :

* + 1. Bagaimana pengklasifikasian emosi berdasarkan dataset GoEmotion menggunakan algoritma *Logistic Regression*?

### Tujuan

Dari Rumusan masalah yang diperoleh ,terdapat tujuan dalam penelitian yaitu:

* + 1. Memperoleh implementasi algoritma *Logistic Regression* dalam klasifikasi emosi berdasarkan pada dataset Go Emotions.

### Manfaat Penelitian

Manfaat penelitian untuk Tugas Pemrosesan Bahasa Alami yaitu :

* + 1. Dapat melakukan implementasi dalam deteksi emosi dalam dataset GoEmotion dengan mengklasifikasikan nya dalam bentuk emosi positif, negatif, dan netral.
    2. Dapat memahami pemikiran secara logis terkait menganalisis masalah dengan pengolahan bahasa alami.

### Ruang Lingkup Penelitian

Ruang lingkup penelitian Tugas Pemrosesan Bahasa Alami yaitu :

* + 1. Data yang diolah merupakan data GoEmotions yang telah disediakan oleh Hugging Face.
    2. Implementasi ini menerima inputan dalam bentuk kalimat Bahasa inggris dan melakukan normalisasi dalam bentuk huruf.
    3. Pemrograman menggunakan bahasa pemrograman *Python*.
    4. Fitur ekstraksi yang digunakan adalah TF-IDF*.*

**BAB 2 PENDAHULUAN**

Bab 2 ini membahas mengenai terkait tinjauan pustaka dan langkah-langkah yang dilakukan pada penelitian ini.

## 2.1 Preprocessing Data

Pada tahap ini data disiapkan menjadi data yang siap untuk dianalisis. Data yang diperoleh ditampilkan dalam keadaaan apa adanya, sehingga mungkin berupa data yang tidak diinginkan karena merupakan data yang tidak relevan. Data yang tidak relevan tersebut terkadang merupakan data yang sulit ditangani. Oleh karena itu, penghapusan bagian data yang tidak relevan yang sangat disarankan dilakukan dan merupakan tahap yang harus dilewati, sehingga efisiensi sistem dapat terwujud. Berikut struktur yang terdapat pada data tweet yaitu User ID (disimbolkan dengan @), URL, teks, tanggal dan waktu, lokasi, file multimedia (gambar, video dll), emoticon, hashtag (disimbolkan dengan #). Masing-masing bagian struktur tersebut memiliki signifikansi sendiri selama analisis sentimen tetapi ada beberapa data yang tidak memiliki efek signifikan sehingga disarankan untuk menghilangkan data ini .

Nama pengguna dalam *Reddit* , ini untuk memberitahu siapa yang menulis komentar tersebut. Pada proses deteksi teks tidak ada pengaruh yang signifikan dari nama pengguna sehingga dengan menerapkan filter, nama pengguna dikeluarkan dari training data dan testing data. Ada batasan karakter dalam komentar *Reddit* sehingga pengguna menyertakan beberapa tautan URL untuk menjelaskannya dengan lebih baik. URL ini (umumnya dimulai dengan http://) tidak diperlukan selama training data dan testing data karena tidak mengandung informasi berguna di dalamnya yang dapat digunakan selama analisis sentimen.

Tahapan untuk melakukan *preprocessing data* terdapat 5 yaitu *text cleaning,case folding,tokenization,stopword removal,*dan *stemming*.

### 2.1.1. *Text Cleaning*

*Text Cleaning* adalah tahapan karakter dan tanda baca yang tidak perlu dihilangkan pada teks,dimana berfungsi mengurangi *noise* pada dataset.Contoh karakter dihilangkan seperti URL,tag

,tanda baca,seperti titik(.),koma(,),dan tanda baca lainnya pada text.

### 2.1.2*. Case Folding*

Tahap *Case Folding* yaitu proses menyamakan huruf teks ke dalam huruf kecil,berarti jika terdapat huruf besar di dalam teks maka diubah ke dalam bentuk huruf kecil .

### *2.1.3. Tokenization*

Proses *Tokenization* adalah memecah kalimat menjadi bagian atau kata-kata.Hasil kata dari proses disebut token.proses *tokenization* yang dilakukan seperti menghilangkan tanda baca yang tidak diperlukan.Terdapat beberapa model *tokenization* yang digunakan yaitu *unigram,bigram,trigram,*dan *ngram*.

### 2.1.4. Stopwords Removal

Merupakan proses menghilangkan kata-kata yang muncul dalam jumlah banyak tetapi dianggap tidak memiliki arti.Alasan menghapus *stopword* yang berhubungan dengan *text mining* yaitu penggunaannya yang terlalu umum,sehingga pengguna bisa fokus pada kata lain yang jauh lebih penting.

### 2.1.5. Stemming

*Stemming* adalah proses mengubah kata menjadi kata dasar menurut kaidah bahasa indonesia yang benar.

## 2.2 TF-IDF

Term Frequency — Inverse Document Frequency atau TF — IDF adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Secara sederhana, metode TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen. Pada Term Frequency (TF), terdapat beberapa jenis formula yang dapat digunakan :

1. TF biner (binary TF), hanya memperhatikan apakah suatu kata atau term ada atau tidak dalam dokumen, jika ada diberi nilai satu (1), jika tidak diberi nilai nol (0).
2. TF murni (raw TF), nilai TF diberikan berdasarkan jumlah kemunculan suatu term di dokumen. Contohnya, jika muncul lima (5) kali maka kata tersebut akan bernilai lima (5).
3. TF normalisasi, menggunakan perbandingan antara frekuensi sebuah term dengan nilai maksimum dari keseluruhan atau kumpulan frekuensi term yang ada pada suatu dokumen.
4. TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit term dalam query, namun mempunyai frekuensi yang tinggi.

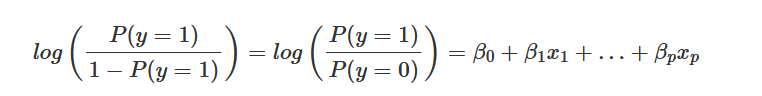
Lalu, yang berikutnya atau yang kedua adalah **IDF (Inverse Document Frequency)**. Metode IDF merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan.Berbeda dengan TF yang semakin sering frekuensi kata muncul maka nilai semakin besar, dalam IDF, **semakin sedikit** frekuensi kata muncul dalam dokumen, maka **makin besar** nilainya.

## 2.3 *Logistic Regression*

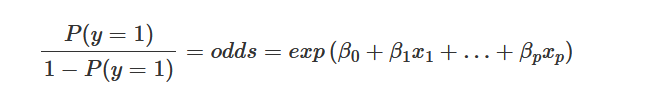
Logistic regression adalah bentuk khusus regresi yang diformulasikan untuk melakukan klasifikasi data ke dalam dua group (prediksi group) dan menjelaskan variabel dependen biner (kategorikal/non-metric) Logistic regression cocok digunakan bila kita ingin memprediksi keanggotan variabel independen (prediktor) dalam dua grup saja, misal grup orang yang menderita diabetes atau tidak menderita diabetes. Bentuk umum logistic regression:

𝑌 =𝑋1 +𝑋2 +𝑋3+...+𝑋𝑛

Dimana Y = biner non-metric dan 𝑋1 +𝑋2 +𝑋3+...+𝑋𝑛 = non-metric atau metric. Logistic regression tidak memerlukan asumsi normalitas dari variabel dependen. Ketika variabel dependen berbentuk kategorikal biner maka distribusinya adalah binomial. Interpretasi bobot dalam logistic regression berbeda dari interpretasi bobot dalam linear regression, karena hasil dalam logistic regression adalah probabilitas antara 0 dan 1. Bobot tidak lagi mempengaruhi probabilitas secara linier. Jumlah tertimbang ditransformasikan oleh fungsi logistik menjadi probabilitas. Oleh karena itu sangat perlu merumuskan kembali persamaan untuk interpretasi sehingga hanya suku linier yang berada di ruas kanan rumus.



Rumus diatas disebut dengan istilah dalam fungsi log() “odds” (probabilitas kejadian dibagi dengan probabilitas tidak ada kejadian) dan dibungkus dalam logaritma yang disebut log odds. Rumus ini menunjukkan bahwa regression models logistik merupakan model linier untuk log odds. Rumus tersebut juga dapat mengetahui bagaimana prediksi berubah ketika salah satu fitur xj diubah 1 unit. Untuk melakukan ini, pertama-tama kita dapat menerapkan fungsi exp() ke kedua sisi persamaan:

Logistic regression tidak memerlukan asumsi normalitas dari variabel independen. Selain itu variabel independen dapat bertipe metric maupun non-metric.Tidak seperti *discriminant analysis*, yang bisa digunakan juga untuk prediksi variabel dependen biner, *logistic regression* tidak memerlukan pengecekan keseimbangan matriks variance-covariance di antara 2 grup →bahkan ketika terpenuhi, *logistic regression* lebih disukai.

## 2.4 Klasifikasi

Klasifikasi adalah suatu kegiatan mengelompokkan. Di mana klasifikasi sangat dibutuhkan dalam perpustakaan, karena klasifikasi bertujuan untuk mengelompokkan satu koleksi yang sejenis, yang pengelompokannya berdasarkan judul, pengarang, dan lain sebagainya.

## 2.5 Sosial Media

Media sosial adalah fasilitas jaringan sosial online yang menghubungkan profil pengguna dengan orang lain atau kelompok lainnya sebagai media komunikasi. Media sosial ada dalam berbagai macam bentuk, diantaranya termasuk sosial network, forum internet, weblogs, social blogs, micro blogging, wikis, podcasts, gambar, video, rating, dan bookmark sosial. Salah satu forum sosial adalah Reddit. Sebagai media jejaring sosial, dalam Reddit pengguna bisa menemukan berita terbaru, topik-topik yang sedang tren, dan meme-meme terbaik di internet. Pengguna juga dapat mencari dan menemukan berbagai komunitas dengan topik beraneka ragam. Aktivitas yang dapat dilakukan seperti mengunggah cerita, gambar, dan tautan menarik untuk bahan diskusi dengan mudah.

**PEMBAGIAN TUGAS**

**Table 1 Deskripsi Tugas**

|  |  |  |
| --- | --- | --- |
| **No** | **Pekerjaan** | **Keterangan** |
| 1 | Studi Literatur | Melakukan studi literatur terkait penelitian yang berasal dari  jurnal dan *paper*. |
| 2 | Pengumpulan Data | Melakukan pengumpulan data berupa kalimat-kalimat berbahasa Indonesia dengan teknik *scraping* dari beberapa buku, jurnal dan  *website*. |
| 3 | Analisis Pengolahan Data | Melakukan analisis mengenai pengolahan data yang berupa  *preprocessing* yang dilakukan pada data |
| 4 | Analisis Algoritma | Melakukan analisis mengenai algoritma yang digunakan untuk  melakukan normalisasi. |
| 5 | Implementasi | Melakukan implementasi. |
| 6 | Pembahasan, Kesimpulan dan Saran | Membuat penjelasan dari pengujian model yang telah dilakukan serta membuat kesimpulan akhir dari tugas akhir dan saran untuk penelitian selanjutnya. |

**Tabel 2 Pembagian Tugas**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Nama** | **Pek erja**  **an** | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **1** | Gorga Siagian |  |  |  |  |  |  |  |
| **2** | Sri Intan Sinaga |  |  |  |  |  |  |  |