Multi-Label Emotion Classification using Naive Bayes 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](#_TOC_250000)

[BAB 1](#_bookmark0) [3](#_bookmark0)

[PENDAHULUAN](#_bookmark1) [3](#_bookmark1)

|  |  |  |
| --- | --- | --- |
| **1.1.** | [**Latar Belakang Masalah**](#_bookmark1) | [3](#_bookmark1) |
| **1.2.** | [**Rumusan Masalah**](#_bookmark2) | [3](#_bookmark2) |
| **1.3.** | [**Tujuan**](#_bookmark3) | [4](#_bookmark3) |
| **1.4.** | [**Manfaat Penelitian**](#_bookmark4) | [4](#_bookmark4) |
| **1.5.** | [**Ruang Lingkup Penelitian**](#_bookmark5) | [4](#_bookmark5) |
| **BAB 2** |  | 5 |

**TINJAUAN PUSTAKA** 5

* 1. ***Preprocessing Data*** 5
  2. [**Word Embedding**](#_bookmark6)[6](#_bookmark6)
  3. [**Naive Bayes**](#_bookmark7)[6](#_bookmark7)
  4. **Klasifikasi 7**
  5. **Sosial Media 7**

[**PEMBAGIAN TUGAS**](#_bookmark8)[8](#_bookmark8)

**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 koeksistensi 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 Naive Bayes dimana model klasifikasi Naive Bayes tersebut dapat menghitung probabilitas posterior suatu kelas berdasarkan distribusi kata dalam dokumen. Sebelum melakukan klasifikasi untuk deteksi emosi pada teks,terlebih dahulu melakukan tahapan text preprocessing, text classification, dan fitur ekstraksi kata dalam teks.

* 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 Naive Bayes?

### Tujuan

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

* + 1. Memperoleh implementasi algoritma Naive Bayes 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 *word embedding.*

**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 Word Embedding

Word embedding merupakan kumpulan nama dari pemodelan bahasa dan teknik ekstraksi fitur pada natural language processing (NLP) dimana setiap kata atau phrasa dari suatu kosakata akan dipetakan menjadi vektor yang berupa bilangan real. Word embedding kerap digunakan dalam neural networks, reduksi dimensi pada matriks kemunculan kata, model probabilistik, dan lain lain. Metode word embedding ini juga digunakan sebagai input untuk meningkatkan performa pada pengolahan bahasa manusia seperti parsing sintaktik dan analisa sentimen. (Mikolov, Sutskever, et al. 2013) mengusulkan dua model log-linear untukmenghitung word embeddings dari suatu dataset secara efisien, yaitu bag-of-words dan skip gram. Continous bag-of words (CBOW) model memprediksi kata saat ini berdasarkan konteks kata. Sedangkan skip-gram memprediksi kata-kata yang berada disekitar kata yang diberikan sesuai dengan kedekatan antara masing-masing vektor kata

## 2.3 Naive Bayes

Naive Bayes adalah pengklasifikasi paling sederhana dan paling umum digunakan. Model klasifikasi Naive Bayes menghitung probabilitas posterior suatu kelas berdasarkan distribusi kata dalam dokumen. Hal itu bergantung pada representasi dokumen yang sangat sederhana sebagai Bag of words. Model ini bekerja dengan menggunakan perhitungan probabilitas dengan cara membandingkan nilai probabilitas suatu sampel berada di suatu kelas dengan nilai probabilitas di kelas lainnya dengan menggunakan konsep dasar teorema Bayes. Metode ini dapat digunakan untuk *text classification, spam filtering*, dan *hybrid recommender system.* Metode ini dapat digunakan untuk *text classification, spam filtering*, dan *hybrid recommender system* [3].Persamaan (2.1) menunjukkan bahwateorema Bayes menggunakan frekuensi untuk mengestimasi probabilitas prior dan posterior.

|  |  |
| --- | --- |
| 𝑝𝑟𝑖𝑜𝑟 × 𝑙𝑖𝑘𝑒𝑙𝑖ℎ𝑜𝑜𝑑  𝑝𝑜𝑠𝑡𝑒𝑟𝑖𝑜𝑟 =  𝑒𝑣𝑖𝑑𝑒𝑛𝑐𝑒  Keterangan:  *prior*: nilai probabilitas sebelum melakukan eksperimen  *posterior*: probabilitas setelah adanya percobaan | (2.1) |

##### *likelihood*: nilai yang menyatakan seberapa sering senuah nilai diberikan bahwa x telah terobservasi

*evidence*: banyaknya sampel

##### Teorema Bayes menunjukkan seberapa sering A terjadi jika diberikan kejadian B,ketika diketahui seberapa sering B terjadi jika diberikan kejadian A, dan seberapa besar kemungkinan A dan B berada pada mereka sendiri [2]. Persamaan (2.2) menunjukkan formula yang digunakan dalam teorema Bayes.

|  |  |
| --- | --- |
| 𝑃(𝐵|𝐴)𝑃(𝐴)  𝑃(𝐴|𝐵) =  𝑃(𝐵)  Keterangan:  P (A | B) : probabilitas A jika diberikan kejadian B P (B | A) :probabilitas B jika diberikan kejadian A P (A) :probabilitas A  P (B) :probabilitas B. | (2.2) |

Metode klasifikasi dilakukan untuk mengambil input *x* dan sekumpulan output tertentu dari kelas *Y=y1,y2,..,ym* dan mengembalikan sebuah kelas yang diprediksi *y ∈ Y*. Dalam klasifikasi teks, *C* merupakan kelas, *y* merupakan *output variable*, *d* merupakan dokumen, dan *x* merupakan *input variable.*

## 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 |  |  |  |  |  |  |  |