**Klasifikasi Emosi Multi Label Menggunakan Pendekatan Algoritma Support Vector Machine**

Proposal Proyek Akhir PBA

Disampaikan Sebagai Bagian Dari Persyaratan Kelulusan Matakuliah

Pemrosesan Bahasa Alami

Oleh :

|  |  |
| --- | --- |
| 11S18003 | Febby Irene Siringoringo |
| 11S18011  11S18027  11S18032  11S18057  11S18059 | Awit Hutabalian  Romauli Siagian  Rido Sitanggang  Marlyse Elysabeth Sitorus  Kristiane Purba |
|  |  |



Institut Teknologi Del

2021/2022

# 

# DAFTAR ISI

[**DAFTAR ISI i**](#_gjdgxs)

[**DAFTAR GAMBAR ii**](#_30j0zll)

[**DAFTAR TABEL iii**](#_1fob9te)

[**BAB 1 PENDAHULUAN 1**](#_3znysh7)

[1.1](#_2et92p0) Latar Belakang 1

[1.2](#_tyjcwt) Perumusan Masalah 2

[1.3](#_3dy6vkm) Tujuan Penelitian 2

[1.4](#_1t3h5sf) Ruang Lingkup Penelitian 2

[1.5](#_4d34og8) Metode Penelitian 3

[1.6](#_2s8eyo1) Sistematika Penyajian 3

[**BAB 2 TINJAUAN PUSTAKA 4**](#_17dp8vu)

[2.1](#_3rdcrjn) Sub Bab 4

[2.2](#_26in1rg) Sub Bab 4

[2.3](#_lnxbz9) Sub Bab 4

[**Lampiran Rangkuman Paper 5**](#_1ksv4uv)

[**DAFTAR REFERENSI 6**](#_35nkun2)

# DAFTAR GAMBAR

[Gambar 1.1 Hubungan Era Kelima Komputer dan Era Kelima Manajemen 2](#_44sinio)

[Gambar 2.2 *Value Chain* Nike 4](#_2jxsxqh)

**DAFTAR TABEL**

[Tabel 2.1 Karakteristik A Berdasarkan B 4](#_z337ya)

# PENDAHULUAN

Pada bab ini akan menjelaskan latar belakang penelitian, rumusan permasalahan penelitian, tujuan penelitian, ruang lingkup penelitian, hasil yang diharapkan dalam penelitian dan tahapan penelitian dalam proyek mata kuliah Pemrosesan Bahasa Alami (PBA)

## Latar Belakang

Natural Language Processing (NLP) merupakan salah satu bidang ilmu dalam kategori keilmuan Artificial Intelligence (AI). NLP fokus pada pengolahan bahasa alami yang diterapkan pada piranti komputer, sehingga komputer dapat memahami apa yang diinginkan oleh pengguna. Bahasa alami tidak memiliki compiler atau interpreter untuk dapat dipahami sebagai halnya bahasa pemrograman, akan tetapi ketika bahasa alami ini diterapkan dalam NLP maka dapat menggunakan bahasa pemrograman sebagai instruksi dari manusia kepada komputer. NLP digunakan dalam aplikasi-aplikasi yang membutuhkan pengolahan bahasa. Model aplikasinya diantaranya berupa aplikasi yang dapat digunakan untuk meringkas dokumen, penerjemahan bahasa asing, hingga aplikasi mengenai ucapan pengguna. Pada penelitian kali ini, pengaplikasian NLP yang akan dibahas adalah *emotion recognition*.

Agar konten yang menggambarkan *emotion* dapat dipahami dan dicari, terdapat suatu metode yang cukup efektif untuk menangani masalah ini yaitu dengan cara pelabelan. Pelabelan dilakukan dengan memberikan tanda atau label terhadap suatu teks. Pelabelan dipilih karena label merupakan bagian dari struktur hirarkis dari teks, sehingga dapat mewakili konten dari suatu teks. Jika data yang akan dianalisis banyak, maka klasifikasi multi-label dapat dijadikan solusi untuk melabelkan konten ke dalam beberapa kelas sekaligus. Setelah konten terlabeli, untuk mengetahui suatu konten masuk ke dalam label mana, perlu dilakukan proses klasifikasi.

Emosi seseorang dapat dilihat dari segi ucapan, ekspresi wajah, nada suara dan juga teks yang ditulis. Pengekspresian emosi dapat diketahui dalam sebuah kalimat. Emosi dalam kalimat dapat diidentifikasi melalui kata sifat. Klasifikasi emosi merupakan tugas penting dalam Natural Language Processing (NLP). Secara otomatis menyimpulkan emosi merupakan langkah awal untuk aplikasi seperti chatbots emosional (Zhou et al, 2018). Pada pengklasifikasian emosi memilih lebih dari 1 emosi dalam teks. Pengklasifikasian emosi sangat diperlukan untuk mengetahui suasana hati atau perasaan orang tersebut. Pengekspresian emosi dapat kita lihat dari postingan , chat ataupun komentar di media sosial.

Deteksi emosi membantu menjelaskan sifat kompleks dari emosi yang muncul secara bersamaan , sehingga memberikan pemahaman tentang karakteristik setiap emosi. Analisis emosi memiliki beberapa manfaat dalam berbagai aplikasi dalam perdagangan, kesehatan masyarakat, kesejahteraan sosial dan lain-lain. Analisis ini menunjukkan sikap terhadap suatu target atau topik. Seringkali ketika kita membaca teks atau email, seringkali kita tidak mengetahui emosi yang terkandung pada teks tersebut sehingga perlu dilakukan klasifikasi emosi dengan menggunakan multi label seperti emosi kegembiraan, kemarahan, kecemasan , kekecewaan, ketidaksetujuan dan lain-lain. Sehingga perlu dilakukan penelitian untuk melakukan klasifikasi emosi dengan multi label dalam sebuah kalimat.

Support Vector Machine (SVM) merupakan metode yang banyak digunakan pada proses pengklasifikasian data. Support Vector Machine (SVM) memiliki keunggulan dalam mengolah data dengan dimensi tinggi tanpa mengurangi performanya. Metode SVM digunakan karena metode ini sangat cepat dan efektif pada klasifikasi data teks (Feldman & Sanger, 2007).

## Rumusan Masalah

Rumusan masalah dari proyek ini yaitu sebagai berikut:

1. Bagaimana penerapan algoritma Support Vector Machine dalam pengklasifikasian emosi *multi-label* berdasarkan skor prediksi emosi yang dihasilkan terhadap data *training* dan *testing*?
2. Bagaimana hasil dari pengklasifikasian emosi *multi-lable* dengan menggunakan algoritma SVM?

## Tujuan Penelitian

Tujuan penelitian dari proyek ini adalah sebagai berikut :

1. Memperoleh model klasifikasi emosi dengan multi label menggunakan algoritma SVM.
2. Memperoleh hasil klasifikasi emosi multi label dengan meninjau dengan confuse matrix.

## Ruang Lingkup

Ruang lingkup dari pengerjaan proyek ini adalah sebagai berikut:

1. Penelitian ini menggunakan pendekatan algoritma Support Vector Machine (SVM)
2. Dataset yang digunakan di proyek ini berasal dari <https://huggingface.co/datasets/go_emotions>
3. Dataset sudah diberi label terlebih dahulu pada setiap kelas emosi dataset yang akan digunakan.

## Hasil yang diharapkan

Hasil yang diharapkan dari pengerjaan proyek ini adalah

1. Menerapkan *Support Vector Machine* dalam melakukan klasifikasi multi label untuk emosi berbasis teks.
2. Penerapan model pada sebuah *Command Line based simulator*.

## Tahapan Penelitian

Rencana kegiatan yang akan dilakukan selama pengerjaan proyek ini adalah sebagai berikut:

1. Studi Literatur

Pada tahap ini mengumpulkan informasi yang terkait dengan topik dan metode yang peneliti gunakan dalam penelitian. Informasi berupa paper, jurnal, buku, informasi dari internet, dan lain sebagainya yang dapat menjadi referensi untuk peneliti dalam melakukan penelitian. Informasi dapat diperoleh dari perpustakaan dan internet dengan sumber informasi yang terpercaya.

1. Pengumpulan Data

Dalam melakukan penelitian diperlukan data yang digunakan berasal dari <https://huggingface.co/datasets/go_emotions>

1. Analisis dan Perancangan Sistem

Pada tahap ini menganalisis masalah dan tujuan yang akan dibahas, metode yang digunakan menurut informasi yang diperoleh dan menyesuaikan dengan data yang dikumpulkan. Kemudian membangun perancangan (implementasi) terhadap penggunaan data dengan algoritma.

1. Implementasi Sistem

Setelah dilakukan analisis data dan perancangan, maka tahap selanjutnya adalah implementasi untuk algoritma Support Vector Machine (SVM)

1. Penulisan Laporan

Tahap terakhir dalam penelitian ini adalah laporan akhir yang berisi kesimpulan dari penelitian yang telah dilakukan.

# Tinjauan Pustaka

Pada bab ini akan dijelaskan tentang tinjauan pustaka dan landasan teori yang ada terkait topik penelitian.

## 

## Multi-label Text Classification

*Multi-label text classification* merupakan bagian dari *supervised classification* dimana masing-masing data dapat dikaitkan lebih dari satu label. Klasifikasi *multi-label* adalah tugas mengkategorikan teks ke dalam satu atau lebih kategori. Kinerja pada klasifikasi *multi-label* terbatas ketika ada data kecil berlabel dan mengarah pada kesulitan menangkap hubungan semantik. Dibutuhkan teknik klasifikasi teks multilabel yang dapat mengelompokkan label dari teks.

Pelabelan data teks cukup memakan waktu namun penting untuk klasifikasi teks otomatis. Khususnya, membuat beberapa label secara manual untuk dokumen mungkin tidak praktis ketika jumlah data sangat besar, sehingga perlu untuk melatih pengklasifikasian teks multi label. Untuk meminimalkan *human-labeling effort*, ada beberapa pendekatan pembelajaran aktif multi-label yang dapat mengurangi data berlabel yang diperlukan tanpa mengorbankan akurasi klasifikasi. Pendekatan dalam memperhitungkan informasi multi-label, dan memilih data tidak berlabel yang dapat menghasilkan penguranagan terbesar dari kehilangan model yang diharapkan, ini dapat dioptimalkan dengan *Support Vector Machines* (SVM).

* 1. **Tahapan Preprocessing**

Tahap preprocessing merupakan tahapan awal dalam mempersiapkan data mentah sebelum dilakukan proses lain. Tujuan tahap preprocessing adalah untuk kemudahan dan penyeragaman pembacaan. Tahap preprocessing dilakukan dengan cara mengeleminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Tahap preprocessing terdiri dari beberapa tahap diantaranya:

### 2.2.1 Text Cleaning

Tahap ini bertujuan untuk menghilangkan kata-kata dan karakter-karakter yang tidak bermakna. Penghilangan data ini bertujuan mengurangi noise data dan missing value. Contohnya tanda baca, symbol dan karakter-karakter yang tidak dikenali. Contoh atribut yang tidak dikenali yaitu link yang diawali dengan atribut “http”, atribut @ dan symbol-simbol seperti ~!#$%^&\*()\_+:{}[]|. Ketika atribut-atribut tersebut dihilangkan dari data maka tidak berpengaruh terhadap hasil klasifikasi lalu digantikan dengan menggunakan karakter spasi.

|  |  |
| --- | --- |
| **Input** | **Output Process** |
| Ustadz Abdul Somad Dihadiahi Mobil Alphard oleh Anak-Anaknya | Ustadz Abdul Somad Dihadiahi Mobil Alphard oleh Anak Anaknya |
| Hina Panglima TNI dan Jokowi di Medsos, Dokter di Sumbar Ditangkap | Hina Panglima TNI dan Jokowi di Medsos Dokter di Sumbar Ditangkap |

### 2.2.2 Case Folding

Setelah tahapan *text cleaning* selesai, maka dilakukan tahapan *case folding*. Tahapan *case folding* merupakan tahapan yang dilakukan untuk menghilangkan huruf besar atau mengubah atau mengkonversikan text menjadi bentuk standar, dimana hanya menerima huruf kecil atau *lowercase* saja (a-z) dan karakter-karakter kecuali alfabet akan dihilangkan dan dianggap sebagai *delimiter*. *Delimiter* merupakan urutan satu atau lebih karakter yang digunakan untuk menentukan batas pemisah. Tujuan dilakukannya tahap *case folding* yaitu untuk menghindari redundansi kata, karena jika data ditulis dengan huruf awal kapital dan huruf non kapital tidak ada terdeteksi bahwa datanya memiliki arti yang berbeda. Berikut contoh dari *case folding*:

|  |  |
| --- | --- |
| **Input** | **Output Process** |
| Ustadz Abdul Somad Dihadiahi Mobil Alphard oleh Anak-Anaknya | ustadz abdul somad dihadiahi mobil alphard oleh anak anaknya |
| Hina Panglima TNI dan Jokowi di Medsos, Dokter di Sumbar Ditangkap | hina panglima tni dan jokowi di medsos dokter di sumbar ditangkap |

**2.2.3 Tokenisasi**

Tokenisasi adalah sebuah proses yang membagi teks, dimana teks dapat berupa kalimat, paragraf atau dokumen kedalam beberapa bagian (token). Contohnya pada kalimat “Aku suka makan sate kacang.” menjadi ‘aku’, ‘suka ‘, ‘makan’, ‘sate’, ‘kacang’

|  |  |
| --- | --- |
| **Input** | **Output Process** |
| Ustadz Abdul Somad Dihadiahi Mobil Alphard oleh Anak-Anaknya | ‘ustadz’, ‘abdul’ ,’somad’,’ dihadiahi’,’ mobil’,’ alphard’,’ oleh ‘,’anak’,’ anaknya’ |
| Hina Panglima TNI dan Jokowi di Medsos, Dokter di Sumbar Ditangkap | ‘hina’,’ panglima’,’ tni’,’ dan ‘,’jokowi’,’di’,’ medsos’,’ dokter’,’ di’,’ sumbar ‘,’ditangkap’ |

**2.2.4 Pembobotan TF IDF**

Tahap selanjutnya yaitu menghitung bobot dari setiap data dengan menggunakan penghitungan bobot TF-IDF. Tahap TF-IDF adalah sebuah metode pembobotan *term* yang sudah berkali-kali digunakan sebagai pembandingpada metode pembobotan yang lain. Tujuan utama dilakukannya pembobotan TF-IDF yaitu untuk meningkatkan hasil klasifikasi algoritma SVM pada penelitian ini.

Tahapan pada TF-IDF ini adalah mencari jumlah kata yang diketahui (tf) saat sudah dikalikan dengan banyaknya data/dokumen di mana kata tersebut muncul (idf). TF-IDF pada SVM digunakan agar data bisa dianalisa.

Untuk menghitung *Term Frequency* kita menggunakan persamaan berikut

Tf(t,d)

Keterangan :

tf = banyak kata t muncul dalam satu dokumen

Untuk mencari fr(x,t) kita memperoleh dari persamaandibawah ini

r(x,t)

Untuk Menghitung IDF maka digunakan persamaan berikut

idf =

Keterangan:

i = jumlah variabel

j = jumlah data

N = jumlah dokumen

df ij = banyaknya dokumen j yang memuat kata i

Untuk menghitung TF IDF menggunakan persamaan dibawah ini

Keterangan:

Wij = besar bobot dari kata i pada dokumen ke-j

tf ij = banyaknya kata i pada dokumen j

df ij = banyaknya dokumen j yang memuat kata i

* 1. **Algoritma Support Vector Machine (SVM)**

*Support Vector Machine* (SVM) merupakan algoritma klasifikasi yang bekerja dengan cara mencari *hyperplane* yaitu klasifikasi dengan cara membagi data menjadi dua kelas menggunakan garis vektor. SVM bertujuan untuk mendapatkan hipotesis berupa bidang pemisah yang memiliki generalisasi yang baik serta meminimalisir rata-rata *error* (*empirical risk)*. Algoritma ini bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan untuk menemukan data yang menjadi dua kelas menggunakan garis vektor atau disebut *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. SVM dapat mengatasi masalah klasifikasi dan regresi dengan *linier* maupun *non-linear.*

Kelebihan pada algoritma *Support Vector Machine* (SVM) adalah sebagai berikut:

1. Generalisasi didefinisikan sebagai kemampuan suatu method dalam melakukan pengklasifikasian suatu pola dimana tidak termasuk didalamnya daya yang dipakai dalam fase *learning.*
2. *Curse of Dimensionality* didefinisikan sebagai sebuah permasalahan yang dihadapi suatu metode dalam mengestimasikan parameter.
3. Feasibility SVM dapat diimplementasikan dengan mudah karena proses penentuan vektor dapat dirumuskan dalam QP Problem. QP problem dapat diselesaikan dengan metode sekuensial pada SVM.

Kekurangan pada algoritma *Support Vector Machine* (SVM) adalah sebagai berikut:

1. Untuk permasalahan dalam skala besar, yaitu jumlah sampel yang diolah, SVM sulit dipakai. Ketika data dilatih yang digunakan sangat besar, akan mempengaruhi waktu pelatihan dan ukuran memori yang dibutuhkan.
2. Menggunakan fungsi pemisah yang memisahkan data ke dalam dua kelas, jika kelas yang ingin dipisahkan lebih dari dua maka diperlukan modifikasi.

**2.3.1 Structural Risk Minimization (SRM)**

Structural Risk Minimization (SRM) bertujuan untuk mengklasifikasikan data yang tidak terdapat dalam data *training* dengan benar dan menjamin batas atas dari generalisasi pada data pengujian dengan cara mengontrol fleksibilitas dari hipotesis hasil *learning*.

* 1. **Corpus**

Corpus merupakan kumpulan teks yang disimpan dalam bentuk digital atau elektronik dengan tujuan penggunaannya sebagai sumber data dalam penelitian linguistik. Dalam pembuatan sampel bahasa corpus, ada beberapa hal yang perlu diperhatikan yaitu orientasi bahasa, kriteria pemilihan sampel dan sifat sampel.

## 

# DAFTAR REFERENSI

Mutisari,K., Adiwijaya, Said Al Faraby.2020.Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines, *IEEE International.*

Gunawan, K.I., & Santoso, J. (2021). Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia. *Journal of Information System, Graphics, Hospitality and Technology, 3*(01), 29-38, http://doi.org/10.37823/insight.v3i01.126