

**LAPORAN PROJECT PEMBELAJARAN MESIN DASAR  
SEMESTER GENAP 2024/2025**

IDENTITAS PROYEK	
Judul	Klasifikasi emosi dalam data teks menggunakan metode SVM
Topik	Klasifikasi menggunakan SVM
Identitas Penyusun	1. Shinta Usaila Farachin (23031554160) 2. Ikhrima Atusifah (23031554181) 3. Arina Tri Yuni Wahyuning Tiyas (23031554203)
Kelas	2023 A

## 1. PENDAHULUAN

Pendahuluan penelitian tidak lebih dari 1000 kata yang terdiri dari:

A. Latar belakang dan rumusan permasalahan yang akan diteliti

B. Pendekatan pemecahan masalah

### ○ Latar Belakang (min. 250 kata)

Perkembangan teknologi informasi dan bertambahnya aktivitas manusia di internet telah menyebabkan meningkatnya data teks yang mengandung berbagai ekspresi emosional. Teks-teks ini tidak hanya mencerminkan pendapat, tetapi juga emosi pengguna terhadap suatu permasalahan. Oleh karena itu, kebutuhan akan sistem otomatis yang dapat mengklasifikasikan emosi dari teks menjadi penting, terutama pada bidang bisnis, psikologi, rekomendasi serta monitoring kesejahteraan pengguna internet. Klasifikasi emosi merupakan proses untuk mengelompokkan teks ke berbagai bidang emosi tertentu seperti senang, sedih, marah, takut, dan terkejut. Dalam ranah ini, berbagai metode pembelajaran mesin telah dikembangkan, salah satu algoritma yang sesuai dengan hal ini adalah support factor machine (svm). Svm telah dikenal efektif dalam menangani data yang memiliki dimensi tinggi dan terbukti dapat memberikan solusi yang baik dalam bidang klasifikasi teks, seperti klasifikasi emosi, terutama ketika metode tersebut digabungkan dengan metode seperti TF-IDF dan bag of words [1].

Penggunaan SVM dalam hal klasifikasi emosi terbukti telah menunjukkan performa yang baik, seperti stui alzhbi dkk (2019) pada klasifikasi emosi tweet bahasa arab [2]. Selain itu studi dari Haryadi dan Kusuma (2019) juga menunjukkan bahwa svm masih menjadi salah satu baseline utama dalam klasifikasi bahasa inggris [3]. Sementara rahman dkk (2019) membandingkan berbagai metode klasik dan menemukan bahwa svm memberikan hasil yang baik pada tema berbahasa bengali [4]. Dalam

permasalahan lainnya, model tradisional seperti svm memiliki tempat tersendiri dalam sistem nlp yang efisien dan transparan [5].

Melalui proyek ini, dilakukan eksplorasi terhadap efektivitas metode SVM dalam klasifikasi emosi teks dengan menggabungkan 4 metode ekstraksi fitur yang umum digunakan, yaitu TF-IDF, bow, word2vec dan fasttext, serta mengukur kinerjanya melalui matrix evaluasi seperti akurasi, precision, recall dan f1-score. Proyek ini diharapkan dapat memberikan kontribusi pada pengembangan sistem nlp berbasis pembelajaran mesin untuk bahasa yang terbatas sumber daya anotasinya.

#### ○ **Rumusan Masalah dan Tujuan**

Rumusan masalah

1. Bagaimana cara merepresentasikan data teks ke dalam bentuk fitur numerik yang sesuai agar dapat digunakan oleh algoritma SVM untuk klasifikasi emosi?
2. Seberapa baik algoritma SVM dalam mengklasifikasikan emosi pada data teks berdasarkan matrik evaluasi seperti akurasi, precision, recall, dan f1-score?
3. Bagaimana pengaruh pemilihan metode ekstraksi fitur (seperti TF-IDF dan Bag of Words) terhadap performa model SVM dalam klasifikasi emosi?

Tujuan

1. Merepresentasikan data teks ke dalam bentuk fitur numerik yang sesuai menggunakan metode seperti TF-IDF dan Bag of Words (BoW) agar dapat digunakan sebagai input dalam model klasifikasi berbasis SVM.
2. Menerapkan dan mengevaluasi algoritma Support Vector Machine (SVM) dalam mengklasifikasikan emosi pada data teks ke dalam beberapa kelas emosi yang telah ditentukan.
3. Menganalisis dan membandingkan pengaruh pemilihan teknik ekstraksi fitur terhadap performa klasifikasi emosi menggunakan SVM berdasarkan metrik evaluasi seperti akurasi, precision, recall, dan f1-score.

## **2. Metodologi**

Metodologi atau cara untuk mencapai tujuan yang telah ditetapkan ditulis tidak melebihi 1000 kata. Bagian ini berisi metode pre-processing dan/atau metode post processing yang dilengkapi dengan diagram alir penelitian yang menggambarkan apa yang sudah dilaksanakan dan yang akan dikerjakan selama waktu yang diusulkan. Format diagram alir dapat berupa file JPG/PNG. Metode penelitian harus dibuat secara utuh dengan penahapan yang jelas.

#### ○ **Eksplorasi Dataset**

Pemahaman dataset yang dimiliki

Dataset diambil dari kaggle dengan judul datasetnya yaitu klasifikasi emosi Indonesia. Dalam dataset ini berisi kumpulan ulasan produk dari aplikasi tokopedia. Dataset

tersebut berisi ulasan produk dari 29 kategori produk di Tokopedia yang menggunakan bahasa Indonesia. Setiap ulasan produk diberi anotasi dengan satu emosi, yaitu cinta, kebahagiaan, kemarahan, ketakutan, atau kesedihan. Data ini akan digunakan dalam klasifikasi emosi dengan kategori emosi senang, sedih dan lainnya. Pada dataset ini terdiri dari 5400 baris dan 11 kolom yaitu kolom kategori, nama produk, lokasi, harga, peringkat keseluruhan, jumlah terjual, ulasan lengkap, peringkat pelanggan, ulasan pelanggan, sentimen dan emosi. Kolom yang nantinya akan digunakan yaitu kolom ulasan pelanggan serta kolom emosi. Pada kolom emosi terdapat 5 target yaitu happy, sadness, anger, love dan fear. Pada dataset ini masih mengandung tanda baca, emoticon dan kata kata yang tidak baku sehingga perlu diproses lebih lanjut. Berikut link dataset dari mendeley :

<https://data.mendeley.com/datasets/574v66hf2v/1>

#### ○ Langkah Penelitian

Dalam melakukan klasifikasi emosi terhadap data teks seperti ulasan produk diperlukan beberapa tahapan sistematis untuk mendapat hasil yang baik dan efisien. Berdasarkan studi Parvin dkk [1] serta berbagai referensi lain, berikut langkah-langkah yang dilakukan:

##### 1. Pengumpulan data (Data Collection)

Tahap pertama dalam proyek ini yaitu megumpulkan data yang akan dianalisis. Dalam proyek ini data didapat dari kaggle dengan judul klasifikasi emosi Indonesia.

##### 2. Data Preprocessing

Tahapan preprocesing merupakan salah satu tahapan yang penting untuk proyek ini, tahapan ini penting untuk meminimalisir kesalahan atau noise yang dapat mengganggu akurasi model [6]. Selain itu, kualitas preprocessing akan mempengaruhi performa klasifikasi [7]. Pada proyek, proses ini mencakup:

1. Pembersihan missing value dan duplikat : pada proyek ini tidak ada missing value akan tetapi memiliki jumlah duplikat 7, setelah dibersihkan data menjadi 5393 baris.
1. Menghapus karakter khusus, angka, URL dan emoticon yang tidak sesuai
2. Tokenisasi teks menjadi kata-kata individual
3. Mengubah huruf menjadi lowercase agar konsisten
4. Menghapus stopwords
5. Normalisasi teks tau kata

6. Memberi label (kata kerja, kata sifat dll) dan hanya memilih kata yang penting. Tahap ini cukup penting agar model bekerja pada data yang lebih terstruktur dan bersih [8].

7. Visualisasi kategori emosi

### 3. Exploratory Data Analysis (EDA)

Tahap EDA merupakan tahapan yang penting untuk mengidentifikasi karakteristik linguistik spesifik yang dapat membantu ekstraksi fitur yang efektif [9]. Selain itu, tahapan ini berguna untuk pemodelan unbiased [10]. Pada tahapan ini akan mencakup:

1. Distribusi kelas : menghitung jumlah data untuk tiap kelas emosi pada train, validasi dan test
2. Kata unik per kelas : menghitung jumlah kata unik per kelas
3. Panjang teks : memeriksa maksimum panjang teks dan rata rata per teks per kelas

4. Visualisasi

### 4. Feature Extraction

Pada tahapan ini akan mencakup: bag of words, TF-IDF, word2vec dan fasttext. Selain itu tahap ini juga akan mencakup proses padding sekuens untuk menyamakan panjangnya. Tetapi sebelum melakukan feature extraction akan dilakukan reduksi dimensi menggunakan pca.

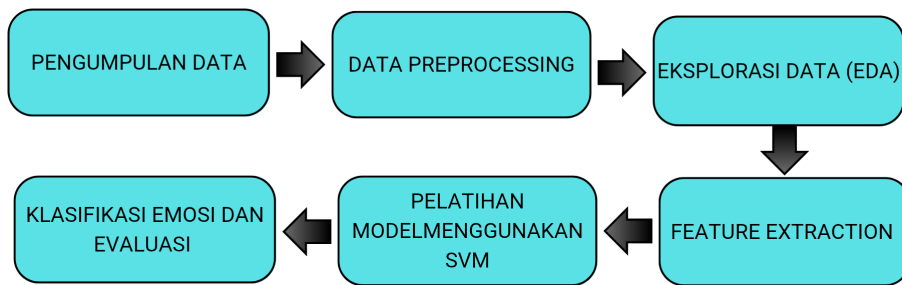
### 5. Pelatihan Model

Data yang sudah diproses kemudian akan dibagi menjadi 80% data latih, 10% data validation dan 10% data uji untuk kemudian dilatih menggunakan algoritma SVM.

### 6. Klasifikasi Emosi dan Evaluasi

Pada tahap ini model kemudian akan digunakan untuk mengklasifikasikan emosi dari data teks. Untuk mengukur performa model digunakan metrik: accuracy (tingkat prediksi yang benar), precision (ketepatan dalam klasifikasi positif), recall (kemampuan dalam menangkap semua data positif) dan F1-Score (keseimbangan antara precision dan recall). Evaluasi ini penting untuk memastikan generalisasi model terhadap data baru.

Berikut diagram alur tahapan klasifikasi emosi pada data teks menggunakan svm



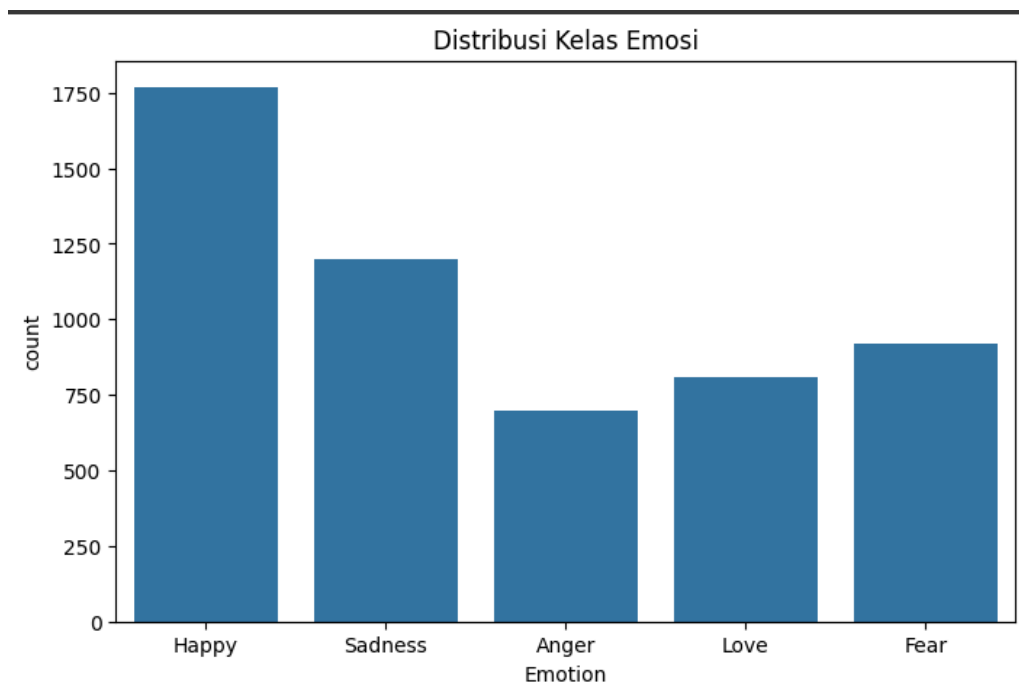
### 3. Hasil dan Analisis

#### 3.1 Exploratory Data Analysis

Berikut adalah hasil EDA proyek ini:

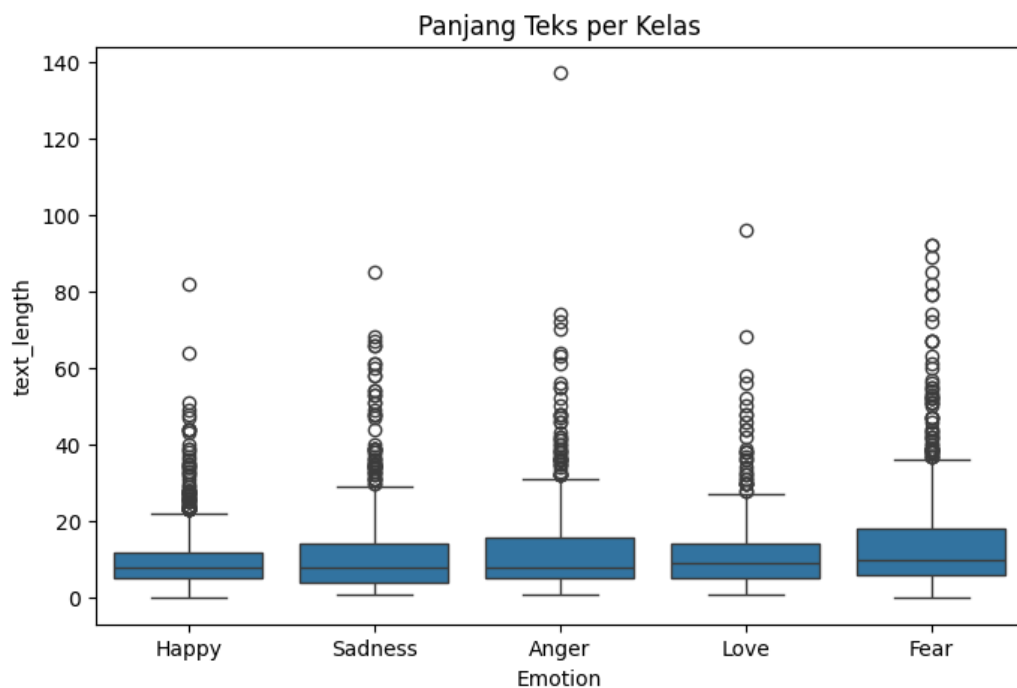
##### 1. Distribusi kelas

Distribusi pada proyek ini menunjukkan bahwa kelas happy memiliki data sebanyak sekitar 1767 kata, diikuti sadness, fear, love dan anger yang paling sedikit dengan jumlah kurang dari 750. Karena data ini tidak seimbang maka dilakukan penyeimbangan data. Setelah dilakukan keseimbangan data, semua data mempunyai jumlah yang sama yaitu 1413.



##### 2. Panjang teks

Pada proyek ini terlihat bahwa median panjang teks relatif mirip di antara semua kelas, yaitu antara 10 - 15 panjang satuan. Akan tetapi, ada cukup banyak outlier di setiap kelas, terutama kelas anger, yang memiliki teks dengan panjang hingga hampir 140. Hal ini, menunjukkan adanya variasi besar dalam panjang teks. Oleh karena itu, perlu dilakukan normalisasi atau padding. Setelah dilakukan padding panjang teks menjadi panjang maksimal 137.



### 3. Kata unik per kelas

Hasil EDA dari jumlah kata unik yaitu diketahui bahwa kelas sadness memiliki jumlah kata unik terbanyak (3095), diikuti oleh fear (3450), sementara love memiliki jumlah kata unik paling sedikit (2332). Hal ini menunjukkan bahwa ekspresi kesedihan dan ketakutan cenderung lebih bervariasi atau kompleks dibanding dengan ekspresi cinta. Sebaliknya emosi cinta digambarkan mempunyai kata kata yang konsisten atau terbatas.

```
Happy: 3095 kata unik
Sadness: 3528 kata unik
Anger: 2617 kata unik
Love: 2332 kata unik
Fear: 3450 kata unik
```

## 3.2 Hasil Eksperimen

### Hasil evaluasi cross validation

Hasil dari cross-validation memperlihatkan nilai akurasi rata-rata sebesar 0.5965 dengan standar deviasi  $\pm 0.0093$ , yang menunjukkan kinerja model yang cukup konsisten di setiap fold. Angka ini menunjukkan bahwa model dapat secara konsisten mengidentifikasi pola emosi dalam teks. Dalam laporan klasifikasi, tampak bahwa model menunjukkan kinerja paling baik dalam mendeteksi emosi Bahagia, dengan nilai f1-score 0.77, precision 0.71, dan recall 0.83. Ini menunjukkan bahwa model sangat efektif dalam mengenali dan mengklasifikasikan teks yang memiliki muatan emosi positif.

Sementara itu, untuk emosi Marah dan Takut, model mencatat precision 0.45 dan 0.36, dengan recall masing-masing 0.49 dan 0.46, serta nilai f1-score sebesar 0.45 dan 0.40. Ini menunjukkan bahwa meskipun berada dalam kelas dengan jumlah data yang lebih sedikit, model tetap dapat mengenali dan mengklasifikasikan teks yang mengandung kemarahan dan ketakutan dengan baik. Emosi Cinta dan Kesedihan juga memperoleh perhatian yang seimbang dengan nilai f1-score masing-masing 0.61 dan 0.56.

Secara keseluruhan, tingkat akurasi klasifikasi tercatat 0.60, dengan nilai rata-rata makro (macro average) untuk precision, recall, dan f1-score masing-masing 0.56, serta nilai berbobot (weighted average) masing-masing 0.60. Nilai-nilai ini menunjukkan bahwa model berhasil membagi perhatian secara cukup seimbang di semua label emosi, dan tidak terlalu mendominasi satu kelas tertentu. Ini menunjukkan keberhasilan dalam merancang pipeline klasifikasi teks yang dapat menangkap nuansa emosi secara komprehensif, baik dari emosi positif maupun negatif.

```
=== Evaluasi Cross-Validation untuk Gabungan (TF-IDF + BoW + Word2Vec + FastText) (5-Fold StratifiedKFold) ===
Akurasi Rata-rata Gabungan (TF-IDF + BoW + Word2Vec + FastText): 0.5965 (+/- 0.0093)

Classification Report (Rata-rata hasil CV):
      precision    recall  f1-score   support

   Anger         0.41      0.49      0.45         699
    Fear         0.41      0.39      0.40         920
   Happy         0.75      0.78      0.77        1770
    Love         0.63      0.58      0.61         809
   Sadness       0.60      0.56      0.58        1202

 accuracy         0.60         0.60         0.60        5400
  macro avg       0.56         0.56         0.56        5400
 weighted avg     0.60         0.60         0.60        5400
```

## Hasil evaluasi SVM

Hasil penilaian menunjukkan bahwa model berhasil memperoleh akurasi sebesar 62,41%. Ini menunjukkan kinerja klasifikasi yang memuaskan dalam mengenali berbagai macam emosi pada data uji. Dalam laporan klasifikasi, emosi Happy menunjukkan performa tertinggi dengan f1-score 0.79, precision 0.76, dan recall 0.83, yang mengindikasikan bahwa model sangat sensitif terhadap teks yang mencerminkan kebahagiaan. Akurasi ini krusial dalam aplikasi nyata seperti analisis sentimen atau pemantauan psikologis yang berbasis teks.

Untuk emosi Kemarahan, model mencatat precision sebesar 0.48, recall 0.45, dan f1-score 0.46, yang menunjukkan bahwa model dapat mengidentifikasi ekspresi kemarahan dengan tingkat presisi yang cukup baik. Hal yang sama terjadi pada emosi Fear, dengan nilai f1-score 0.45 dan presisi 0.40. Emosi Cinta dan Kesedihan juga menunjukkan hasil yang cukup baik, masing-masing dengan f1-score 0.64 dan 0.56. Ini menunjukkan bahwa model mampu menggeneralisasi kinerjanya ke berbagai ekspresi emosional tanpa terlalu condong ke satu kelas.

Rata-rata makro untuk precision, recall, dan f1-score adalah 0.58, sementara nilai berbobot untuk masing-masing metrik itu mencapai 0.62. Nilai-nilai ini menunjukkan bahwa model beroperasi secara proporsional terhadap seluruh kategori, bukan hanya menitikberatkan pada kelas mayoritas. Dengan demikian, penilaian ini menunjukkan bahwa pendekatan pelatihan model telah memberikan kinerja yang seimbang dan konsisten dalam konteks klasifikasi emosi yang rumit dan bernuansa.

=== Evaluasi untuk Gabungan (TF-IDF + BoW + Word2Vec + FastText) ===				
Akurasi Gabungan (TF-IDF + BoW + Word2Vec + FastText): 0.6241				
	precision	recall	f1-score	support
Anger	0.48	0.43	0.46	83
Fear	0.44	0.45	0.45	75
Happy	0.76	0.83	0.79	181
Love	0.62	0.61	0.62	72
Sadness	0.61	0.57	0.59	129
accuracy			0.62	540
macro avg	0.58	0.58	0.58	540
weighted avg	0.62	0.62	0.62	540

### 3. Daftar Pustaka

Sitasi disusun dan ditulis berdasarkan sistem nomor sesuai dengan urutan pengutipan, mengikuti format APA. Hanya pustaka yang disitasi pada usulan penelitian yang dicantumkan dalam Daftar Pustaka. Pustaka yang disitasi maksimal 8 tahun terakhir sebanyak minimal 10 pustaka.

[1]. Parvin, T., Sharif, O., & Hoque, M. M. (2022). Multi-class textual emotion categorization using ensemble of convolutional and recurrent neural network. *SN Computer Science*, 3(1), 62.



- [2]. Alzu'bi, S., Badarneh, O., Hawashin, B., Al-Ayyoub, M., Alhindawi, N., & Jararweh, Y. (2019, October). Multi-label emotion classification for arabic tweets. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 499-504). IEEE..
- [3]. Haryadi, D., & Kusuma, G. P. (2019). Emotion detection in text using nested long short-term memory. *International Journal of Advanced Computer Science and Applications*, 10(6).
- [4]. Rahman, M. A., & Seddiqui, M. H. (2019). Comparison of classical machine learning approaches on bangla textual emotion analysis. arXiv preprint arXiv:1907.07826.
- [5]. Sun, B., & Chen, H. (2021). A survey of k nearest neighbor algorithms for solving the class imbalanced problem. *Wireless Communications and Mobile Computing*, 2021(1), 5520990.
- [5]. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- [6]. Alswaidan, N., & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8), 2937-2987.
- [7]. Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: a review. arXiv preprint arXiv:1806.00674.
- [8]. Distant, D., Faralli, S., Rittinghaus, S., Rosso, P., & Samsami, N. (2022). Domainsenticnet: An ontology and a methodology enabling domain-aware sentic computing. *Cognitive computation*, 14(1), 62-77.
- [9]. Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., & Saha, T. (2020, May). Multi-domain tweet corpora for sentiment analysis: resource creation and evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 5046-5054).
- [10]. Samy, A. E., El-Beltagy, S. R., & Hassanien, E. (2018). A context integrated model for multi-label emotion detection. *Procedia computer science*, 142, 61-71.