Emo*Tweet*AI : Sistem Klasifikasi Emosi Dalam Teks Unggahan di Media Sosial Twitter

Theofilus Arifin
Fakultas Teknik Program Studi Teknik Informatika
Universitas Surabaya
Surabaya, Indonesia
theofilusarifin@gmail.com

Jonathan Ryan Darmawan
Fakultas Teknik Program Studi Teknik Informatika
Universitas Surabaya
Surabaya, Indonesia
jonathan.ryan.darmawan@gmail.com

Achmad Nashruddin Riskynanda
Fakultas Teknologi dan Informatika Cerdas Departemen Teknik Informatika
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
achmad.riskynanda01@gmail.com

Abstract*—Dalam era digital yang berkembang pesat, masyarakat semakin aktif berbagi pendapat dan emosi mereka melalui media sosial seperti Twitter. Fenomena ini mencerminkan pergeseran yang signifikan dalam cara kita berkomunikasi dan berinteraksi dalam lingkungan online. Dalam konteks ini, analisis sentimen menjadi hal yang penting untuk dapat memahami dinamika sosial, terutama dalam hal penilaian emosi yang terkandung dalam komunikasi online. Penelitian ini berfokus pada pengembangan model yang dapat mengklasifikasikan dan memahami emosi yang disampaikan melalui tweet seseorang menggunakan arsitektur model SVM, yang secara khusus diadaptasi untuk teks berbahasa Indonesia. Dengan menggunakan model ini, peneliti dapat memperoleh wawasan yang lebih mendalam tentang emosi masyarakat dan bagaimana mereka bereaksi terhadap suatu topik dan peristiwa berdasarkan tweet yang telah mereka tulis pada media sosial twitter.

Keywords—sentimen, analisis, klasifikasi, twitter, emosi

I. PENDAHULUAN

A. Latar Belakang

Seiring berjalannya waktu, berbagai aspek kehidupan manusia terus berkembang. Hal ini dapat meliputi pendidikan, kesehatan, industri, teknologi, dan bidang lainnya. Perkembangan aspek ini tentu saja juga berdampak terhadap tuntutan masyarakat baik kepada pemerintah maupun swasta. Bagi pemerintah, peningkatan kualitas layanan publik sudah diatur pada Undang-Undang Nomor 25 Tahun 2009 tentang pelayanan publik. Terdapat berbagai jenis layanan yang harus dipenuhi oleh pemerintah. Layanan tersebut meliputi Pelayanan barang, jasa, dan administratif. Salah satu indikator keberhasilan pemerintah dalam memberikan layanan adalah tingkat kepuasan masyarakat. Tingkat kepuasan masyarakat dapat diukur salah satunya melalui opini publik terhadap layanan, kebijakan, maupun keputusan yang diambil oleh pemerintah.

Selain pemerintahan, dunia industri juga terus berlomba untuk memberikan yang terbaik bagi masyarakat. Industri akan selalu berusaha untuk meningkatkan baik barang atau layanan yang mereka jual. Salah satu hal yang dapat berperan penting dalam peningkatan kualitas ini adalah *feedback* atau opini masyarakat terhadap barang atau jasa yang disediakan. Dengan adanya opini ini, industri dapat meningkatkan kualitas barang maupun jasa yang mereka jual sesuai dengan keinginan dari masyarakat.

Berdasarkan hal tersebut, opini publik menjadi suatu hal yang sangat penting baik bagi pemerintah maupun industri. Opini masyarakat tersebut dapat ditemukan salah satunya pada media sosial Twitter. Twitter merupakan salah satu situs mikro *blogging* yang cukup populer di Indonesia. Indonesia menempati posisi keenam pengguna Twitter terbanyak di dunia. Pengguna Twitter dapat menyampaikan ekspresi atau opini melalui *tweet*, *Tweet* merupakan tulisan singkat dan sederhana yang dapat menggambarkan pendapat maupun emosi seseorang. Berbagai tulisan seperti komentar, pujian, diskusi, keluhan, argumen, tuntutan, dan berbagai hal lainnya sering kali disampaikan melalui Twitter. Data yang adap pada Twitter sendiri sering kali digunakan dalam proses analisis sentimen [1]. Twitter sendiri menyediakan sebuah API bernama Twitter API yang dapat memberikan berbagai data *tweet* pengguna Twitter. Salah satu contoh opini yang disampaikan melalui Twitter pada 5 September 2023 yang berisi “Kejadian sore ini tgl 5 sept 2023 jam 17.10 di daerah bogangin, kedurus air pdam keruh sekaliii. Mana bisa utk mandi dll, tolong @PDAMSurabaya @e100ss” yang menunjukkan kekecewaan terhadap kualitas layanan dari PDAM Surabaya. Hal ini akan menjadi sulit apabila pemerintah atau perusahaan ingin melihat *trend* secara keseluruhan bagaimana pelayanan mereka.

Berdasarkan hal yang sudah dijelaskan sebelumnya, kami mengajukan sebuah sistem sentimen analisis bernama “Emo*Tweet*AI” yang dapat memberikan data berupa emosi seseorang berdasarkan *tweet* yang diunggah. Emosi yang dihasilkan dapat menjadi gambaran bagaimana tanggapan atau opini publik terhadap suatu hal seperti keputusan, kebijakan atau kualitas layanan baik dari pemerintah maupun industri. Sistem ini diharapkan dapat mempermudah pemerintah maupun industri dalam memperoleh opini masyarakat terhadap suatu hal.

B. Tujuan dan Manfaat

Penelitian ini memiliki tujuan antara lain:

1. Merancang sistem analisis sentimen berbasis *Machine Learning* untuk memperoleh emosi seseorang melalui unggahan pada Twitter.
2. Melakukan perbandingan beberapa model *Machine Learning* untuk menentukan model terbaik yang sesuai dengan sistem yang dibuat

Manfaat dari penelitian ini antara lain:

1. Meningkatkan kualitas layanan publik pemerintah kepada masyarakat.
2. Meningkatkan kualitas barang atau jasa dari industri yang dijual kepada masyarakat
3. Meningkatkan kepuasan masyarakat terhadap layanan publik dari pemerintah serta kepuasan pelanggan terhadap barang atau jasa dari industri

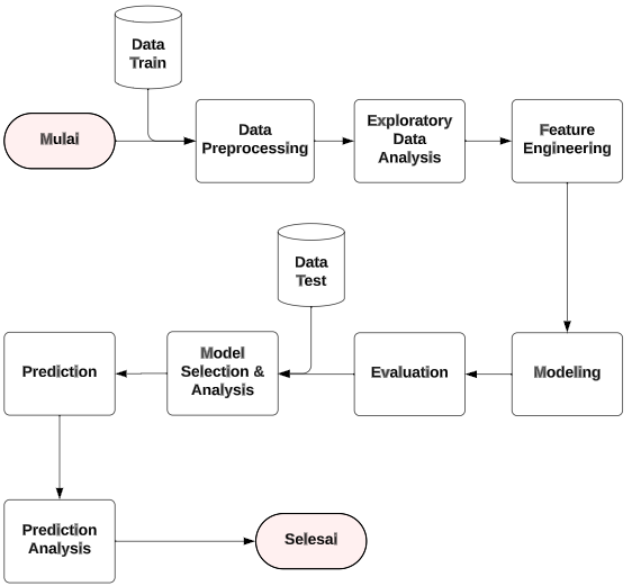
II. STUDI TERKAIT

Analisis sentimen sendiri menjadi tantangan dan kebutuhan bagi banyak pihak. Penelitian [2] melakukan klasifikasi sentimen terhadap layanan publik Pemerintah Kota Makassar yang disampaikan melalui Twitter. Naïve Bayes digunakan sebagai *classifier* untuk menilai apakah tulisan tersebut termasuk ke dalam kategori negatif, netral, atau positif. Sistem pada penelitian tersebut mendapat tingkat akurasi sebesar 91.6%. Studi [3] dilakukan dengan menggunakan Naïve Bayes dan SVM. *Classifier* ini diterapkan pada data opini masyarakat Surabaya di Twitter untuk dikategorikan ke dalam tiga kategori yaitu negatif, positif, dan netral. Akurasi yang didapat oleh Naïve Bayes adalah 78.77% sedangkan SVM adalah 79.81%. Selain itu, studi [4] menggunakan algoritma KNN untuk mengklasifikasi opini pelanggan PLN Jakarta ke tiga kategori yaitu negatif, netral, dan positif. Dari uji yang dilakukan, didapat akurasi klasifikasi sebesar 89.4%.

III. METODOLOGI

A. Usulan Pendekatan

Dalam upaya melakukan klasifikasi emosi dalam teks unggahan yang ada di Twitter, penelitian ini melakukan usulan pendekatan metode yang terdiri dari serangkaian tahap yang dapat dilihat pada Gambar 1. Tahap awal penelitian dimulai dengan proses pembersihan data pada data *train* agar lebih mudah dieksplorasi untuk tahapan berikutnya. Tahapan berikutnya adalah *Exploratory Data Analysis* (EDA) pada data *train*, pada penelitian ini eksplorasi dan identifikasi dilakukan guna mendapatkan *insight* penting dari berbagai label sentimen yang ada. Hal ini bertujuan untuk memahami karakteristik dari tiap label sentimen yang ada. Dengan insight dari data yang sudah bersih, tahap selanjutnya beralih ke *feature engineering*. Pada tahap *feature engineering*, ekstraksi akan dilakukan terhadap fitur-fitur kunci yang menjadi elemen penting dalam memprediksi emosi dalam teks unggahan di platform media sosial Twitter. Setelah proses persiapan data selesai, tahap berikutnya adalah melatih dan membandingkan beberapa model *machine learning* guna mencari model terbaik yang dapat digunakan dalam analisis. Evaluasi dari setiap model yang telah dilatih, memungkinkan kami untuk memilih model yang paling optimal untuk melakukan prediksi label pada data *test* yang belum memiliki label. Pada tahap akhir, analisis terhadap hasil prediksi dilakukan guna mengevaluasi kualitas dan kinerja dari model yang digunakan.



Gambar 1. Diagram alir usulan pendekatan untuk klasifikasi emosi dalam teks unggahan di media sosial Twitter

B. Dataset

Dataset yang digunakan dalam penelitian ini adalah data teks yang merupakan hasil unggahan pada media sosial Twitter. *Dataset* terdiri dari 5153 yang dibagi menjadi 4153 data *train* dan 1000 data *test* dengan dua kolom yaitu kolom label dan *tweet*. Kolom label merupakan kategori emosi dari teks hasil unggahan sedangkan kolom *tweet* merupakan kalimat teks hasil unggahan. Kolom label terdiri dari 5 jenis kategori emosi yaitu *anger*, *fear*, *joy*, *love*, dan *sadness*. Sampel pada masing-masing kategori label dapat dilihat pada tabel 1.

Tabel 1 Sampel data untuk masing-masing label pada dataset

Label	*Tweet*
Fear	Lebih menyeramkan kalo punya grup WA keluarga yg isinya sharing2 hadist & ayat Al-Qur'an DAN grup WA lingkungan paroki gereja.... Serem kalo salah reply
Joy	*Tweet*: Hepibesdey canteeeekk [USERNAME] panjang umur, jadi pribadi yang jauh lebih baik, semoga apa yang di cita-citakan terwujud dan langgeng sama [USERNAME]
Love	*Tweet*: Happy annyversarry yg pertama kalinya .. Trsrah org mau blng apa, aku gk pdli.. Ini yg aku rasain slma 1blan.. Aku cukup bhagia sm km smuanya baik" aja, jgn ada mslh yy sayang, cemburu krna org lain wajarkan.. Dan smg kita juga makin mngrti satu sama lain.
Sadness	*Tweet*: beli kaos banyak dengan ukuran yg udah downgrade, kucariin kok ngga ada ternyata masuk ke lemari Bapak, mbak ART ku kayaknya ga notice aku udah kurusan.
Anger	*Tweet*: [USERNAME] tolong aplikasinya diberesin. Saya tadi pesan, katanya server error. Eh muncul 4 taksi. Khan kasian drivernya, saya musti cancel.

C. Data Preprocessing

Data *Preprocessing* merupakan kumpulan langkah-langkah dengan tujuan membuat data lebih dimengerti oleh mesin agar mengurangi ambiguitas pada langkah *feature extraction*. Data *Preprocessing* merupakan langkah yang sangat penting dan akan memengaruhi efisiensi langkah-langkah selanjutnya [5]. Pada data *Preprocessing* kami melakukan pemrosesan terhadap data *train* dan data *test*

secara bersamaan guna memastikan tiap proses dilakukan untuk data *train* dan data *test*. Pemisahan data *train* dan *test* akan dilakukan setelah data preprocesing selesai. Metode *Preprocessing* data yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Penghapusan data kosong.
   Penghapusan dilakukan data yang memiliki teks kosong pada kolom *tweet*. Hal ini dilakukan karena data kosong tidak memberikan kontribusi apa pun dalam analisis dan dapat mengganggu kinerja algoritma pemodelan.
2. Penghapusan data duplikat.
   Pengapusan data duplikat dilakukan hanya pada data *train*. Hal ini dilakukan untuk menghilangkan data yang identik dalam *dataset* agar tidak terjadi *overfitting* saat pelatihan model.
3. Penghapusan mention dan link.
   Pada *dataset* terdapat key*word* [USERNAME] dan juga [URL] yang merupakan hasil dari crawling data pada media sosial Twitter. Key*word* ini perlu dihapus karena tidak memiliki makna pada analisis sentimen yang akan dilakukan. Contoh hasil proses penghapusan mention dan link dapat dilihat pada tabel 2.

Tabel 2. Contoh hasil proses penghapusan mention dan link.

Sebelum	Sesudah
Hepibesdey canteeeekk [USERNAME] panjang umur	Hepibesdey canteeeekk panjang umur
klo milih pemimpin bukan petahana nanti jadinya gub tidak tahu & males kerja spt ini [URL]	klo milih pemimpin bukan petahana nanti jadinya gub tidak tahu & males kerja spt ini

4. Penghapusan *hashtag*.
   Proses ini menghapus simbol "#" pada *dataset* dan memisahkan kata pada *hashtag*. Contoh hasil proses penghapusan *hashtag* dapat dilihat pada tabel 3.

Tabel 3. Contoh hasil proses penghapusan hashtag.

Sebelum	Sesudah
#BulanBungKarno #JuniBulanBungKarno	bulan bung karno juni bulan bung karno

5. *Lowercasing*.
   Proses ini mengubah semua huruf dalam teks menjadi huruf kecil untuk menghindari perbedaan kasus yang tidak relevan. Contoh hasil proses *lowercasing* dapat dilihat pada tabel 4.

Tabel 4. Contoh hasil proses lowercasing.

Sebelum	Sesudah
Waktu kalian jd oposisi Era Pak Harto, Pak SBY	waktu kalian jd oposisi era pak harto, pak sby

6. Penghapusan string emoticon.
   Proses ini menghapus emotikon atau ekspresi wajah dari teks. Contoh hasil proses penghapusan string emoticon dapat dilihat pada tabel 5.

Tabel 5 Contoh hasil proses penghapusan string emoticon.

Sebelum	Sesudah
bentar2 marahan begitu deh :(	bentar2 marahan begitu deh

7. Penghapusan emoji
   Proses ini melakukan penghapusan terhadap karakter emoji atau gambar simbol dari *dataset* yang ada. Emoji bukan merupakan teks sehingga penghapusan emoji membantu menjaga fokus pada teks yang sebenarnya dan mencegah gangguan dari karakter-karakter visual yang tidak relevan. Contoh hasil proses penghapusan emoji dapat dilihat pada tabel 6.

Tabel 6. Contoh hasil proses penghapusan emoji.

Sebelum	Sesudah
momen di mana kamu merasa seperti amarahmu meresap ke dalam dirimu seperti racun. 💢😤	momen di mana kamu merasa seperti amarahmu meresap ke dalam dirimu seperti racun.

8. Penghapusan tanda baca
   Proses ini menghapus karakter-karakter yang tidak dibutuhkan dalam proses klasifikasi teks. Karakter tanda baca yang dihapus dan digantikan oleh spasi contohnya seperti koma, titik, tanda tanya, tanda seru, dan karakter lain. Khusus untuk karakter petik satu penghapusan tanda baca tidak diganti dengan spasi. Contoh hasil proses penghapusan tanda baca dapat dilihat pada tabel 7.

Tabel 7. Contoh hasil proses penghapusan tanda baca.

Sebelum	Sesudah
nyante dulu.. sruput kopi biar tambah bijak...	nyante dulu sruput kopi biar tambah bijak
ayat al-qur'an dan grup wa lingkungan	ayat al quran dan grup wa lingkungan

9. Pengubahan Bahasa Inggris ke Bahasa Indonesia
   Proses ini mengubah kata dalam Bahasa Inggris menjadi Bahasa Indonesia menggunakan *library google translate*. Perubahan ini dilakukan karena sentimen analisis yang dikerjakan dalam Bahasa Indonesia namun beberapa orang Indonesia kerap mencampur Bahasa Inggris dalam teks unggahan pada media sosial Twitter. Dalam rangka memastikan tidak ada kata berbeda yang memiliki makna yang sama, proses pengubahan dari Bahasa Inggris ke Bahasa Indonesia dilakukan. Contoh hasil proses pengubahan Bahasa Inggris ke Bahasa Indonesia dapat dilihat pada tabel 8.

Tabel 8. Contoh hasil proses pengubahan dari bahasa Inggris ke bahasa Indonesia.

Sebelum	Sesudah
first impression kami terhadap ka abel	kesan pertama kami terhadap ka abel
selesai 100 persen bye bye bye harapan indah	selesai 100 persen selamat tinggal selamat tinggal selamat tinggal harapan indah

10. Penghapusan karakter berulang.
    Kehadiran karakter berulang dalam teks data dapat menghambat efisiensi proses ekstraksi fitur dalam pemodelan data. Sebagai contoh dalam *feature extraction*, ketika melakukan *token*isasi kata "cintaaa" potensi perbedaan nilai mungkin muncul dibandingkan dengan kata "cinta" apabila karakter berulangnya tidak dihapus. Selain itu, keberadaan karakter berulang yang tidak dihapus dapat menghasilkan sejumlah besar fitur yang

pada dasarnya identik, namun diperlakukan sebagai fitur yang berbeda. Hal ini mengakibatkan keberadaan data yang redundan dan tidak memberikan informasi yang relevan. Contoh hasil proses penghapusan karakter berulang dapat dilihat pada tabel 9.

Tabel 9. Contoh hasil proses penghapusan karakter berulang.

Sebelum	Sesudah
hepibesdey canteeeekk panjang umur	hepibesdey cantek panjang umur

11. Penghapusan data numerik.
Penghapusan data numerik dalam analisis sentimen perlu dilakukan karena data numerik seperti angka atau kode sering kali tidak memiliki makna dalam konteks analisis sentimen dan dapat mengganggu proses pemodelan teks. Hal ini membantu meningkatkan akurasi analisis sentimen dengan fokus pada informasi tekstual yang relevan, menghindari gangguan dari data numerik yang tidak memberikan wawasan terkait sentimen yang diungkapkan dalam teks. Contoh hasil proses penghapusan data numerik dapat dilihat pada tabel 10.

Tabel 10. Contoh hasil proses penghapusan data numerik.

Sebelum	Sesudah
yang penting tidak ngejer2 cowok orang	yang penting tidak ngejer cowok orang

12. Penghapusan kata yang terdiri dari satu huruf
Penghapusan kata yang terdiri dari satu huruf perlu dilakukan karena kata tersebut merupakan *noise* dan tidak memberikan kontribusi signifikan terhadap pemahaman sentimen dalam teks. Dengan menghapus kata yang terdiri dari satu huruf saja, analisis sentimen dapat dilakukan dengan fokus pada kata-kata yang lebih informatif. Contoh hasil proses penghapusan kata yang terdiri dari satu huruf dapat dilihat pada tabel 11.

Tabel 11. Contoh hasil proses penghapusan kata yang terdiri dari satu huruf data numerik.

Sebelum	Sesudah
persibku menang lagi o o o o o	persibku menang lagi

13. Penghapusan kata *slang*
Proses ini mengubah kata yang tidak baku menjadi baku. Hal ini membuat kata-kata pada *dataset* ter-standarisasi dengan baik dan mengacu pada kata yang baku guna memungkinkan analisis sentimen yang lebih akurat. Pada penelitian ini terdapat 3931 kata *slang* dan kata baku seharusnya yang sudah dikumpulkan dari berbagai sumber maupun dibuat secara manual dari hasil pengamatan terhadap *dataset*. Contoh hasil proses penghapusan kata *slang* dapat dilihat pada tabel 12.

Tabel 12. Contoh hasil proses penghapusan kata slang.

Sebelum	Sesudah
gw suka banget nih filmnya	saya suka banget ini filmnya

14. Stemming
Stemming adalah proses yang dilakukan dengan menghapus semua imbuhan (*affixes*) yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*), dan kombinasi dari awalan dan akhiran (*confixes*) pada setiap kata sehingga tersisa hanya kata dasarnya [6]. Stemming dapat mengurangi jumlah fitur serta meningkatkan performa model klasifikasi dengan menjadikan beberapa fitur dengan bentuk yang berbeda menjadi hanya satu fitur [7]. Tabel 13 menunjukkan contoh hasil proses stemming.

Tabel 13 Contoh hasil proses stemming.

Sebelum	Sesudah
kakak mengucapkan akan bermain bola sore ini	kakak ucap akan main bola sore ini

15. Penghapusan *stopword*
Proses ini adalah tahap untuk menghapus *stopword* yang merupakan kumpulan kata atau fitur yang sering ditemukan atau berulang namun memiliki pengaruh yang sangat kecil hingga tidak sama sekali dalam proses klasifikasi sehingga perlu untuk dihapus. Jenis kata yang banyak dikategorikan ke dalam *stopword* adalah kata hubung (konjungsi) dan kata ganti [7]. hasil proses penghapusan *stopword* dapat dilihat pada tabel 14.

Tabel 14. Contoh hasil proses penghapusan stopword

Sebelum	Sesudah
saya suka kucing dan anjing	saya suka kucing anjing

D. Feature Extraction

1. *Tokenizing*
Tokenizing merupakan proses memecah teks menjadi bagian-bagian yang terpisah yang disebut *token*. Proses *tokenizing* dilakukan dengan mendeteksi karakter yang menjadi pembatas antar kata seperti spasi untuk dijadikan penanda saat memisahkan teks menjadi kata-kata penyusunnya. Tujuan dari proses ini adalah menyederhanakan teks menjadi *input* yang ringkas untuk proses klasifikasi.
2. *Sequencing*
Sebelum melakukan proses klasifikasi dokumen teks harus terlebih dahulu diubah ke bentuk vektor. Pengubahan ke vektor ini dibutuhkan agar mesin dapat mengidentifikasi kombinasi kata-kata yang ada. Proses mengubah teks menjadi *sequence* ini dilakukan dalam beberapa langkah yaitu membuat kata-kata kamus, mengkonversi kata menjadi angka, dan *padding*. Langkah pertama yang dilakukan adalah membuat kamus kata. Setiap kata unik yang ada di kalimat akan direpresentasikan menjadi sebuah angka. Setelah kamus kata telah terbentuk, langkah selanjutnya adalah mengkonversi kata menjadi angka. Proses ini dilakukan dengan mengkonversi setiap kata yang ada menjadi angka representatifnya pada kamus kata. Langkah selanjutnya adalah melakukan *padding*. *Padding* adalah proses menambahkan angka 0 pada awal atau akhir dari setiap *sequence* sehingga semua *sequence* memiliki panjang yang sama.

E. Exploratory Data Analysis

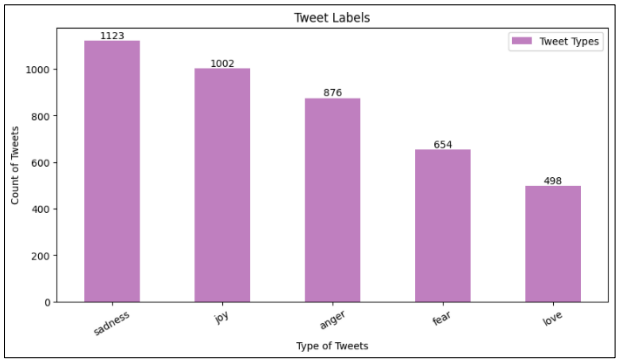
1. Hasil data *Preprocessing*
Berdasarkan tabel 15 dapat dilihat karakteristik dari *tweet* yang ada di Indonesia. Karakteristik yang paling banyak

terlihat pada data tersebut adalah penggunaan *slang word* sebanyak 4278 data. Penggunaan Bahasa Inggris juga banyak terlihat. Sebanyak 4242 data menggunakan Bahasa Inggris. Penggunaan emoticon maupun emoji tidak terlalu banyak digunakan pada data yang ada, hanya terdapat 390 emoji dan 190 emoticon yang ada pada data di dataset yang digunakan.

Tabel 15. Banyak data yang berubah untuk tiap jenis data Preprocessing

Jenis Data *Preprocessing*	Banyak data yang berubah
Penghapusan data kosong pada *train* saja	2
Penghapusan data duplikat pada *train* saja	120
Penghapusan mention dan link	2176
Penghapusan *hashtag*	898
Lowercasing	3773
String emoticon	190
Penghapusan emoji	390
Penghapusan tanda baca	3963
Pengubahan Bahasa Inggris ke Bahasa Indonesia	4242
Penghapusan karakter berulang	3587
Penghapusan data numerik	2175
Penghapusan kata yang terdiri dari satu huruf	1491
Penghapusan kata *slang*	4278
Stemming	5030
Penghapusan *Stopword*	5137

2. Persebaran Data
   Selanjutnya perhitungan banyak data pada masing-masing label dilakukan untuk pengamatan lebih lanjut tentang persebaran data. Banyak data untuk masing-masing kategori dapat dilihat pada Gambar 2.



Gambar 2. Visualisasi bar chart banyak data tweet untuk masing-masing label pada dataset

Dapat dilihat bahwa label *sadness* memiliki data paling banyak yaitu 1123 data sedangkan label *love* memiliki data paling sedikit yaitu 498 data.

3. *Word Statistics*
   Word statistics adalah pendekatan analisis data eksploratif terhadap setiap kata dalam setiap kategori label yang ada. Dalam konteks penelitian ini, kita menghitung frekuensi dan persentase kemunculan kata-kata pada label tertentu. Penelitian ini membagi data sesuai dengan kategori label yang ada kemudian menghitung 7 kata teratas yang sering muncul. Dengan memanfaatkan persentase kemunculan kata pada label tertentu, kita dapat mengamati kemungkinan bahwa kalimat yang mengandung kata tersebut cenderung diklasifikasikan ke dalam suatu label tertentu.

- *Sadness*
  Kata yang paling sering muncul pada label *sadness* adalah "lihat", "lemah", "hati", "sedih", "sakit", "hilang", dan "salah". Variabilitas yang signifikan dalam frekuensi munculnya kata-kata utama ini, serta tidak adanya satu kata yang mendominasi sepenuhnya, mengindikasikan bahwa *tweet* yang diunggah dapat diklasifikasikan sebagai *sadness* berdasarkan beberapa kata kunci tertentu.

Tabel 16. Word Statistics pada label sadness.

Kata	Frekuensi Pada Label	Persentase Keseluruhan Pada Label
lihat	104	9.74%
teman	101	9.46%
hati	98	9.18%
sedih	97	9.08%
sakit	84	7.87%
hilang	82	7.68%
salah	77	7.21%

- *Joy*
  Kata yang paling sering muncul pada label *joy* adalah "kasih", "terima", "selamat", "teman", "moga", "nya", dan "alhamdulilah". Persentase kata yang satu dengan yang lain tidak terpaut jauh sehingga tidak ada satu kata mutlak yang mendominasi pada label *joy*.

Tabel 17. Word Statistics pada label joy.

Kata	Frekuensi Pada Label	Persentase Keseluruhan Pada Label
kasih	124	12.46%
terima	123	12.36%
selamat	96	9.65%
teman	96	9.65%
moga	94	9.45%
nya	75	7.54%
alhamdulilah	71	7.14%

- *Anger*
  Kata yang paling sering muncul pada label *anger* adalah "marah", "pakai", "salah", "nya", "indonesia", "bilang", dan "suka". Persentase kata yang satu dengan yang lain tidak terpaut jauh sehingga tidak ada satu kata mutlak yang mendominasi pada label *Anger*.

Tabel 18. Word Statistics pada label anger.

Kata	Frekuensi Pada Label	Persentase Keseluruhan Pada Label
marah	101	12.07%
pakai	71	8.48%
salah	66	7.89%
nya	54	6.45%
indonesia	52	6.21%
bilang	51	6.09%
suka	50	5.97%

- *Fear*
  Kata yang paling sering muncul pada kategori label *fear* adalah "takut", "seram", "teman", "lihat", "pikir",

"rumah", dam "nya". Kata "takut" memiliki persentase yang paling tinggi dan hampir selalu ada pada unggahan *tweet* dengan label *fear*. Persentase kata "takut" juga terpaut jauh dibandingkan kata-kata lainnya. Hal ini mengindikasikan bahwa *tweet* dengan kata "takut" kemungkinan besar akan diklasifikasikan sebagai label *fear*.

Tabel 19. Word Statistics pada label fear.

Kata	Frekuensi Pada Label	Persentase Keseluruhan Pada Label
takut	487	75.62%
seram	70	10.87%
teman	68	10.56%
lihat	67	10.40%
pikir	62	9.63%
rumah	51	7.92%
nya	42	6.52%

- *Love*

Kata yang paling sering muncul pada label *love* adalah "cinta", "sayang", "orang", "kasih", "hati", "jatuh", dan "ku". Kata "cinta" dan "sayang" merupakan kata yang memiliki persentase tertinggi dan hampir selalu muncul pada unggahan *tweet* dengan label *love*. Persentase kedua data tersebut juga terpaut jauh dibandingkan kata-kata lainnya. Hal ini mengindikasikan bahwa *tweet* dengan kata "cinta" dan "sayang" kemungkinan besar akan diklasifikasikan sebagai label *love*.

Tabel 20. Word Statistics pada label love.

Kata	Frekuensi Pada Label	Persentase Keseluruhan Pada Label
cinta	402	82.72%
sayang	268	55.14%
orang	73	15.02%
kasih	67	13.79%
hati	59	12.14%
jatuh	57	11.73%
ku	49	10.08%

F. Modeling

Pada penelitian ini digunakan 7 model untuk melakukan klasifikasi emosi terhadap data yang ada. Model yang digunakan adalah Naïve Bayes, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), LightGBM, dan Long Short Term Memory (LSTM). Naive Bayes dapat melakukan klasifikasi secara sederhana dan efektif serta dapat digunakan pada berbagai jenis klasifikasi. Random Forest sering kali digunakan pada data yang memiliki fitur kategorikal dengan *missing value* serta dapat memberikan informasi mengenai fitur mana yang paling berpengaruh terhadap prediksi yang dilakukan [8]. SVM dapat digunakan untuk klasifikasi pada data yang memiliki dimensi yang besar dengan efektif [9]. XGBoost sering kali digunakan pada data dengan skala yang besar. Model ini dapat menghasilkan prediksi yang lebih umum sehingga meminimalkan terjadinya *overfitting* [10]. CatBoost merupakan model yang dapat memprediksi data kategorikal dengan keberagaman yang besar serta ukuran yang besar [11]. LightGBM merupakan sebuah model yang dapat digunakan pada klasifikasi dan regresi dengan hasil prediksi yang presisi, stabil, dan efisien [12]. LSTM merupakan salah satu model yang paling baik dan dinamis dalam melakukan klasifikasi pada berbagai data [13].

G. Uji Coba dan Evaluasi

Uji coba dilakukan untuk memastikan bahwa model yang digunakan dalam sistem klasifikasi emosi dalam teks unggahan di media sosial Twitter yang telah ditentukan dengan komputasi yang optimal. Evaluasi ini dilakukan dengan menggunakan data yang telah diberikan anotasi label untuk dibandingkan dengan hasil deteksi model juga membandingkan performa pemrosesan oleh model. Metrik evaluasi yang digunakan adalah akurasi. Akurasi adalah pengukuran proporsi dari prediksi yang benar (*true*) dibandingkan dengan total prediksi yang dilakukan [14]. Akurasi dapat dihitung menggunakan persamaan 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive*, dan FN adalah *False Negative*.

IV. HASIL DAN PEMBAHASAN

Berdasarkan uji coba dari setiap model yang telah dilatih, didapatkan performa dari model seperti yang tertera pada tabel 21. Performa akurasi yang tinggi menunjukkan bagaimana ketepatan model dalam mengklasifikasikan suatu kalimat ke dalam label kelas yang diharapkan. Pada tabel juga disajikan lama waktu pelatihan pada tiap model.

Tabel 21. Perbandingan akurasi dan waktu pelatihan tiap model.

Model	Akurasi	Waktu Pelatihan (s)
Naïve Bayes	58,31%	0.0063
Random Forest	64,68%	4.06
Support Vector Machine	65,42%	1.71
LightGBM	33,99%	0.80
Extreme Gradient Boosting	31,10%	2.62
Categorical Boosting	33,08%	61.17
LSTM	59.88%	357.23

Waktu pelatihan model adalah faktor penting dalam mengevaluasi efisiensi model. Dalam hal ini, Naive Bayes memiliki waktu pelatihan yang sangat singkat, sekitar 0.0063 detik. Ini adalah waktu yang sangat cepat dibandingkan dengan model lainnya. Di sisi lain, CatBoost dan LSTM memiliki waktu pelatihan yang sangat tinggi, masing-masing sekitar 61.17 detik dan 357.23 detik (sekitar 5 menit). Waktu pelatihan yang lama seperti ini bisa menjadi masalah terutama dalam pengembangan model secara praktis.

Dalam hal akurasi, SVM adalah model terbaik dengan akurasi 65.42%, diikuti oleh Random Forest dengan akurasi sekitar 64.68%. LSTM juga memiliki akurasi yang baik yaitu 59.88%. Sementara itu, Naive Bayes memiliki akurasi yang dengan nilai sekitar 58.31%. Model XGBoost, CatBoost, dan LightGBM memiliki akurasi yang lebih rendah, masing-masing sekitar 31.10%, 33.08%, dan 33.99%.

Pada penelitian ini SVM dan Random Forest merupakan model dengan tingkat akurasi yang tertinggi dibandingkan model lainnya. Namun model SVM merupakan pilihan model yang terbaik karena selain memiliki akurasi yang

paling tinggi, waktu pelatihan yang dibutuhkan lebih cepat dibandingkan model Random Forest.

V. KESIMPULAN DAN SARAN

Berdasarkan pengujian yang telah dilakukan, dapat disimpulkan bahwa SVM adalah model yang paling baik dalam hal waktu pelatihan dan memberikan akurasi yang cukup tinggi. Melalui pengembangan model klasifikasi emosi menggunakan SVM yang telah diadaptasi untuk teks berbahasa Indonesia, penelitian ini memberikan solusi yang potensial untuk memahami dinamika sosial dalam berbagai konteks, termasuk pelayanan publik dan industri. Hasil *Exploratory Data Analysis* (EDA) juga memberikan wawasan yang berharga tentang kata-kata kunci yang sering muncul dalam setiap kategori label, yang dapat membantu *stakeholder* dalam merancang strategi komunikasi dan layanan yang lebih responsif terhadap sentimen masyarakat. Namun, studi ini juga memiliki beberapa kendala, seperti ketidakseimbangan jumlah data pada setiap label emosi yang dapat memengaruhi kinerja model.

Saran untuk penelitian mendatang adalah untuk mengatasi ketidakseimbangan data dengan teknik *oversampling* atau *undersampling* untuk meningkatkan akurasi model. Selain itu, memperbaiki kemampuan dalam mengatasi kata-kata slang dan bahasa campuran dalam teks dapat menjadi fokus penting untuk peningkatan keakuratan analisis sentimen di media sosial berbahasa Indonesia. Selanjutnya, studi ini dapat diperluas dengan menggunakan data eksternal pada platform media sosial lainnya, sehingga hasilnya dapat lebih relevan dan berlaku secara lebih luas.

DAFTAR PUSATAKA

[1] A. Mittal and S. Patidar, "Sentiment Analysis on Twitter Data: A Survey," in *Proceedings of the 7th International Conference on Computer and Communications Management*, 2019, pp. 91–95, doi: 10.1145/3348445.3348466.

[2] R. Rosdiana, T. Eddy, S. Zawiyah, and N. Y. U. Muhammad, "Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar," *Proceeding SNTEI*, pp. 87–93, 2019.

[3] N. Y. A. Faradhillah, "Analisis Sentimen Terhadap Kinerja Pelayanan Publik Di Kota Surabaya Berdasarkan Klasifikasi Komentar Di Media Sosial Dengan Menggunakan Algoritma Naïve Bayes." Institut Teknologi Sepuluh Nopember, 2016.

[4] M. S. Alrajak, I. Ernawati, and I. Nurlaili, "Analisis Sentimen Terhadap Pelayanan PT. PLN Di Jakarta Pada Twitter Dengan Algoritma K-Nearest Neighbor (K-NN)," in *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, 2020, vol. 1, no. 2, pp. 110–122.

[5] K. Balasaravanan and M. Prakash, "Detection of dengue disease using artificial neural network based classification techniquetion," *Int. J. Eng. Technol.*, vol. 7, no. 1, pp. 13–15, 2018, doi: 10.14419/ijet.v7i1.3.8978.

[6] M. S. H. Simarangkir, "Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia," *J. Inkofar*, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.

[7] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.

[8] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–14, 2018, doi: 10.1186/s12859-018-2264-5.

[9] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 803–855, 2019.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[11] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. big data*, vol. 7, no. 1, pp. 1–45, 2020.

[12] J. Yan *et al.*, "LightGBM: accelerated genomically designed crop breeding through ensemble learning," *Genome Biol.*, vol. 22, pp. 1–24, 2021.

[13] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks," *arXiv Prepr. arXiv1909.09586*, 2019.

[14] H. Dalianis, "Evaluation Metrics and Evaluation," *Clin. Text Min.*, no. 1967, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6.