JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 8
No. 3
Desember
2022

Pendekatan *Data Science* untuk Mengukur Empati Masyarakat terhadap Pandemi Menggunakan Analisis Sentimen dan Seleksi Fitur

Fika Hastarita Rachman[#1], Imamah[#2]

[#]*Jurusan Teknik Informatika, Fakultas Teknik, Universitas Trunojoyo Madura*
Bangkalan, Jawa Timur, Indonesia

[1]fika.rachman@trunojoyo.ac.id
[2]i2m@trunojoyo.ac.id

Abstrak— **Empati merupakan kemampuan seseorang untuk turut merasakan penderitaan orang lain. Pandemi covid yang melanda dunia, telah menyisakan banyak kehilangan dan keterpurukan. Penelitian ini bertujuan untuk mengetahui emosi masyarakat terhadap penderitaan sesama menggunakan pendekatan sentimen analisis. Dataset yang digunakan adalah komentar masyarakat di Twitter tentang pandemi Covid dalam rentang waktu November-Desember 2020. Data diambil dengan teknik *crawling* menggunakan library *twint*, didapatkan data sebanyak 2386 komentar, namun komentar yang mengandung empati hanya sebanyak 984 data. Dataset empati kemudian dilabeli oleh tiga orang menggunakan teknik *majority voting*. Hasil pengukuran dataset empati menunjukkan 55,7% komentar masyarakat indonesia mengandung empati positif (berempati), 37,4% empati negatif (tidak berempati), dan 6,9% netral. Untuk membentuk model yang dapat mendeteksi empati secara otomatis, maka digunakan dataset empati sebanyak 400, dengan 200 kelas positif dan 200 kelas negatif, kelas netral tidak digunakan pada penelitian ini karena jumlah data sangat sedikit. Metode *machine learning* yang digunakan untuk membangun model adalah *Support Vector Machine* (SVM) dengan metode ekstraksi fitur reliefF. Berdasarkan penelitian yang dilakukan, akurasi sistem dengan metode SVM tanpa seleksi fitur *ReliefF* adalah 83%. Sedangkan akurasi yang diperoleh sistem dengan seleksi fitur *ReliefF* mencapai 93% dengan penggunaan 85% fitur dari total keseluruhan fitur.**

Kata kunci— **Empati, Klasifikasi, Analisis Sentimen, SVM, ReliefF.**

I. PENDAHULUAN

Data science adalah pendekatan yang banyak digunakan untuk mendapatkan informasi dan pengetahuan dengan memanfaatkan ketersediaan data di internet. Saat ini, *data science* menjadi keilmuan yang populer seiring dengan meluasnya penggunaan internet, kecerdasan buatan, dan media sosial. Media sosial menjadi salah satu sumber data dari penelitian yang berkaitan dengan *data science*. Penelitian sebelumnya telah menggunakan media sosial Twitter untuk mengetahui opini masyarakat tentang tempat wisata di kabupaten Bangkalan[1]. Hasil dari penelitian ini dapat memetakan sentimen positif dan negatif dari wisatawan, sehingga dapat dijadikan rujukan untuk pengembangan tempat wisata di masa depan. Beberapa penelitian juga telah banyak menerapkan data dari sosial media twitter [2][3].

Krisis pandemi Covid-19 telah membuat masyarakat menjadi lebih tertarik untuk menggunakan media sosial. Media sosial banyak digunakan untuk menyalurkan aspirasi, promosi, bahkan kritikan terhadap hal yang tidak disukai. Berdasarkan data yang didapatkan, pengguna media sosial meningkat saat pandemi[4]. Peningkatan penggunaan media sosial ini, kemudian menjadi dasar penelitian untuk mengetahui level empati masyarakat. Menurut Qori'ah dkk tahun 2021 perkembangan teknologi telah membuat perilaku masyarakat berubah, dan menyebabkan mereka kurang mempunyai rasa empati[5].

Penelitian ini menggunakan teknik sentimen analisis yang merupakan metode populer dari *data science* untuk menggali informasi dari komentar publik di twitter dan melihat benarkah perkembangan teknologi telah membentuk mental individualis tanpa empati, atau malah sebaliknya.

Selain popularitas dan penerapan sentimen analisis, motivasi dan kontribusi dari penelitian ini adalah sebagai berikut:

a. Mengekstraksi persepsi publik di Indonesia tentang pandemi Covid-19 dengan mengadopsi teknik kecerdasan komputasional yaitu Sentimen analisis, untuk mengukur empati masyarakat.
b. Membangun model yang dapat mendeteksi otomatis komentar masyarakat apakah memiliki empati positif atau negatif.
c. Mengevaluasi kinerja model menggunakan beberapa metrik performa.

Penelitian sentimen analisis sebelumnya [6] mencoba untuk menganalisa sentimen dengan hanya menggunakan ekstraksi fitur TF-IDF tanpa proses seleksi fitur. Proses

seleksi fitur dapat meningkatkan akurasi sistem, karena fitur yang digunakan merupakan fitur terpilih yang penting [7]. Sehingga dalam penelitian ini digunakan proses seleksi fitur dengan menggunakan metode ReliefF.

Artikel ini disusun dalam lima bagian. Bagian pertama berisi latar belakang, tujuan dan kontribusi dari penelitian. Bagian 2 menyajikan penelitian terkait yang dilakukan peneliti sebelumnya. Metode yang digunakan dijelaskan pada bagian 3. Skenario dan hasil uji coba dijelaskan pada bagian 4. Selanjutnya, bagian 5 merangkum hasil penelitian yang dilakukan.

II. KAJIAN PUSTAKA

Pada bagian ini dijelaskan mengenai konsep dasar analisis sentimen, justifikasi penggunaan SVM dan ekstraksi fitur reliefF. Analisis sentimen merupakan teknik untuk menganalisa opini, sikap, perasaan, atau penilaian terhadap suatu peristiwa atau topik yang diekspresikan secara tekstual[3]. Prinsip dasar analisis sentimen adalah mengelompokkan teks yang terdapat di dalam suatu kalimat, kemudian membuat kesimpulan apakah pendapat yang dikemukakan dalam kalimat tersebut bersifat positif, negatif atau netral. Sentimen analisis dapat menyatakan suatu emosional mengandung makna sedih, gembira, atau marah[8].

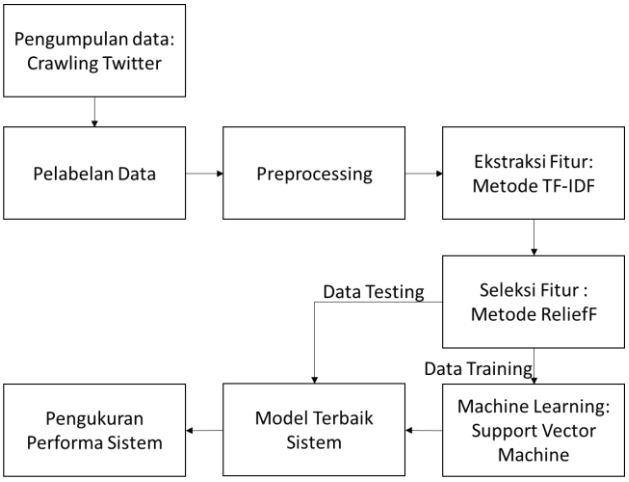
Analisis sentimen membutuhkan *machine learning* untuk mengembangkan model yang dapat melakukan deteksi komentar secara otomatis. Salah satu metode yang banyak digunakan adalah Support Vector Machine (SVM). Dimas dkk menggunakan SVM dan Query Expansion untuk menganalisa komentar masyarakat tentang pembelian produk. Dataset dibagi menjadi dua, kelas positif dan negatif. Akurasi yang dihasilkan sebesar 96,25%[9]. Rachmad dkk juga menggunakan SVM untuk menganalisa sentimen dari para pengguna Gopay, dan dihasilkan akurasi sebesar 89,17%[10]. SVM banyak digunakan dalam penelitian text mining karena mampu menghasilkan model klasifikasi yang baik meskipun dilatih dengan himpunan data yang relatif sedikit[11].

Seleksi fitur diperlukan untuk mengurangi kompleksitas atribut yang kurang relevan dengan tujuan meningkatkan hasil akurasi[12]. Penelitian Yujia dkk menemukan bahwa metode seleksi fitur dengan menggunakan *chi-square* mampu meningkatkan akurasi klasifikasi teks[13]. Oman dkk juga telah membuktikan bahwa seleksi fitur dengan algoritma genetika mampu meningkatkan akurasi pada klasifikasi cerita pendek[14]. Pada penelitian ini, metode seleksi fitur yang digunakan adalah ReliefF. Metode ini dipilih dengan pertimbangan bahwa algoritma ReliefF berdasarkan penelitian yang dilakukan oleh Elsayyad mampu menghasilkan fitur yang lebih spesifik. Hal ini karena ReliefF adalah algoritma seleksi fitur yang sederhana, efektif dan mampu untuk mengidentifikasi fitur relevan berdasarkan interaksi antar fitur[15].

Berdasarkan kajian pustaka pada penelitian sebelumnya, sehingga digunakan SVM dan ekstraksi fitur ReliefF untuk membuat model *machine learning* yang dapat memprediksi jenis empati ke dalam kelas positif (berempati) dan negatif (tidak berempati).

III. METODOLOGI

Bagian ini akan menjelaskan metodologi yang digunakan untuk dapat mencapai kontribusi yang disebutkan pada bagian pertama. Langkah-langkah penelitian digambarkan pada gambar 1.



Gambar 1. Diagram alur penelitian

Data masukan sistem adalah data hasil crawling dari Twitter yang berupa Tweet, kemudian diproses dalam beberapa tahapan dari tahap Pelabelan Data sampai dengan Tahap Pengukuran Performa Sistem.

A. Pengumpulan dan Pelabelan Data

TABEL I
TEKNIK PELABELAN DATA MANUAL

No	Komentar	P1	P2	P3	Label
1.	Sedih bgt dapet kabar om adi meninggal dunia stlh 5 hari dirawat grgr covid. Istrinya meninggal setahun yg lalu. Skrg 3 anaknya yatim piatu. Ketiganya belum nikah. Yg paling kecil masih SMA. Ngga bisa bayangin rasanya jadi mereka. Gabisa nganter ayah ke peristirahatan trkhir	+	+	-	+
2.	alhamdulillah covid bawa berkah	-	-	-	-

Dataset yang digunakan adalah komentar masyarakat di twitter tentang pandemi Covid dalam rentang waktu November-Desember 2020. Kata kunci yang digunakan adalah "covid" dengan filter data hanya komentar berbahasa indonesia. Data diambil dengan teknik crawling menggunakan library *twint*, didapatkan data sebanyak 2386 komentar, namun komentar yang mengandung empati hanya sebanyak 984 data. Data terkumpul setelah proses crawling, tetapi ini masih berupa komentar tanpa label.

Suatu dataset harus memiliki label sebagai *groundtruth* sehingga butuh proses pelabelan data.

Proses pelabelan data harus dengan teliti karena akan sangat berpengaruh terhadap hasil klasifikasi. Dari 3 label yang diberikan oleh 3 orang pada masing-masing *tweet*, selanjutnya akan dilakukan *majority voting*, dimana label yang paling banyak akan dianggap label yang sebenarnya. Contoh proses *majority voting* dapat dilihat pada tabel 1. P1, P2 dan P3 adalah simbol dari orang yang pertama, orang kedua dan orang ketiga.

Pada tabel 1 diberikan contoh dua komentar, pada komentar pertama, P1 (pakar 1) menyatakan bahwa kalimat tersebut mengadung empati positif (berempati), P2 juga menyatakan hal yang sama. Namun P3 memberikan pendapat yang berbeda, dalam proses majority voting, maka diambil label terbanyak untuk dijadikan sebagai label data. Pada komentar kedua, P1, P2 dan P3 sepakat menyatakan bahwa komentar kedua tidak mengandung empati, menganggap bahwa covid membawa berkah hanya pada sebagian orang, sedangkan banyak dari masyarakat yang kehilangan keluarga, pekerjaan karena Covid. Jika merasakan bahagia di saat banyak yang sedih, seharusnya bentuk empati adalah dengan menyembunyikan dan tidak mengekspos kebahagiaan. Berdasarkan penelitian ini, dapat disimpulkan bahwa proses pelabelan dataset empati tidak dapat dilabeli dengan *lexicon* karena kata positif seperti "berkah", dalam konteks kalimat kedua menyebabkan kalimat berlabel negatif. Sebaliknya kata "sedih" pada komentar pertama bukanlah negatif, namun menunjukkan empati.

B. Preproses Data (Data Preprocessing)

Sebelum melakukan proses analisa, dokumen *tweet* harus melalui tahap *preprocessing* terlebih dahulu untuk memperoleh format data yang sesuai, karena format data akan sangat mempengaruhi keoptimalan hasil analisis[16]. Tahap *preprocessing* dokumen *tweet* terdiri dari 5 tahap, yaitu *cleaning, case folding, tokenizing, stopwords removing,* dan *stemming*.

Tahap *cleaning* bertujuan untuk mengurangi *noise* pada data dengan cara nama akun, angka, RT, *hashtag*, duplikat, emotikon, tanda baca, dan juga *hyperlink.* Tahap kedua pada *preprocessing* adalah tahap *case folding*. *Case folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar, biasanya dalam formatihuruf kecil atau *lowercase.* Hanya huruf 'a' isampai dengan 'z' yang diterima. Tahap *tokenizing* adalah tahap pemisahan kalimat menjadi per kata. Setiap kata nantinya akan dihimpun menjadi susunan *array.* Tahap selanjutnya yang dilakukan adalah penghapusan *stopword (stopwords removal)* atau kata bersifat umum dan sering muncul namun tidak memiliki pengaruh terhadap sentimen. Tahap selanjutnya adalah proses stemming yang merupakan proses merubah kata menjadi kata dasar dengan menghapus imbuhan berupa awalan, sisipan, maupun akhiran. Library yang digunakan pada tahap stemming pada penelitian ini adalah Sastrawi yang me. menggunakan algoritma Nazief dan Andriani dalam proses stemmingnya.

C. Ekstraksi Fitur Menggunakan TF-IDF

Ekstraksi fitur merupakan tahapan untuk merubah *term* atau kata menjadi data numerik[17]. Pada penelitian ini, metode ekstraksi fitur yang dipakai TF-IDF *(Term Frequency and Inverse Document Frequency*). TF-IDF bekerja dengan cara memberi bobot atau nilai untuk setiap *term* (t) pada dokumen. Variabel *term frequency* (tf) menyimpan nilai penghitungan jumlah *term* (t) pada satu dokumen (d), sedangkan dokumen frekuensi (df) adalah variabel untuk menyimpan hasil perhitungan dari jumlah dokumen dimana *term* (t) muncul. Variabel *inverse document frequency* (idf) berperan dalam pengurangan bobot *term* jika terlalu sering muncul dalam semua dokumen. Jika suatu *term* terlalu sering muncul pada sebagian besar dokumen maka akan mengganggu pencarian terhadap *term* yang unik.

Proses pembobotan *term* dengan TF-IDF yaitu:

1. Mencari nilai *weighting term frequency* menggunakan rumus:

$$W_{tf\ t,d} = \begin{cases} 1 \\ 0, \end{cases} + log_{10} tf_{t,d}, if tf_{t,d} > 0 \quad (1)$$

2. Mencari nilai *document frequency*.
3. Mencari nilai *inverse document frequency* menggunakan rumus:

$$idf_t = log_{10}\left(\frac{N}{df_t}\right) \quad (2)$$

4. Hitung TF-IDF menggunakan persamaan:

$$W_{t,d} = W_{tf,td} \times idf_t \quad (3)$$

Dimana variabel W_{tftd} merupakan bobot frekuensi, $tf_{t,d}$ adalah frekuensi dari term, idf_t adalah nilai *inverse document frequency*, N merupakan jumlah dokumen, df_t adalah jumlah dokumen yang memuat term(t), dan $W_{t,d}$ merupakan nilai bobot TF-IDF.

D. Algoritma ReliefF

Seleksi fitur adalah proses mengidentifikasi fitur yang memiliki pengaruh yang besar dengan mengesampingkan fitur yang tidak relevan. Dengan semakin sedikitnya data, memungkinkan algoritma pada tahap selanjutnya beroperasi lebih cepat dan lebih akurat[15].

Algoritma ReliefF merupakan salah satu metode seleksi fitur yang banyak digunakan[15]. Sama halnya dengan metode seleksi fitur yang lain, metode ini juga digunakan untuk menemukan fitur-fitur penting pada keseluruhan fitur dataset. Algoritma ReliefF menggunakan teknik pembobotan untuk menghitung signifikansi fitur, dan akan dipilih fitur yang memiliki nilai diatas ambang batas (treshold) [16]. Algoritma ReliefF memanfaatkan konsep *closest neighbors* untuk mengukur bobot fitur berdasarkan interaksi antar fitur[18].

Untuk mendapatkan bobot dari setiap fitur, perlu ditentukan terlebih dahulu satu fitur untuk dijadikan nilai tengah (R), fitur ini dipilih secara acak[19]. Setelah nilai R sudah ditentukan, *ReliefF* akan mengindentifikasi tetangga

terdekat dari R yang disebut *near hit* dan *near miss. Near hit (H_j)* adalah tetangga terdekat yang berada dalam satu kelas dengan atribut R, sedangkan *near miss (M_j)* adalah tetangga terdekat yang berbeda kelas[20]. Perhitungan jarak antar atribut dihitung menggunakan *Manhattan Distance*[21].

$$r = \sum_{i=1}^{N} |pi - qi| \quad (4)$$

Dimana,
r = *manhattan distance*
pi = koordinat titik pertama pada dimensi i
qi = koordinat kedua pada dimensi i
N = jumlah dimensi

Setelah *near hit* dan *near miss* diketahui, maka bobot fitur dapat dihitung. Bobot fitur yang akan dihasilkan adalah antara -1 dan 1. Fitur yang bernilai positif adalah fitur yang relevan. *Output* pada proses ini adalah urutan ranking fitur berdasarkan perhitungan bobotnya[22]. Rumus untuk menghitung bobot fitur adalah sebagai berikut.

$$W = \frac{P(beda\ nilai\ fitur \mid contoh\ kelas\ yang\ berbeda)}{P(beda\ nilai\ fitur \mid contoh\ kelas\ yang\ sama)} \quad (5)$$

E. Support Vector Machine (SVM)

SVM merupakan metode *machine learning* yang diperkenalkan oleh Vladimir dkk pada acara *Annual Workshopmon Computational Learning Theory* tahun 1992[23].

SVM adalah teknik klasifikasi yang digunakan untuk menganalisa data dan memprediksi kelas berdasarkan pola[1]. Teknik klasifikasi pada SVM dilakukan dengan membentuk *hyperplane* atau garis pembatas (*decision boundary*). Hyperplane ini bertujuan untuk memisahkan satu kelas dengan kelas lain[24]. Prinsip dasar *Support Vector Machine* adalah *linier classifier*, yaitu dapat memisahkan data linear. Semisal terdapat himpunan X = {X_1, X_2, X_3,.....X_n}, maka himpunan akan dinyatakan sebagai kelas positif jika *f*(x) ≥ 0 sedangkan lainnya termasuk ke dalam kelas negatif[25]. Fungsi *linier classifier* bisa didefinisikan sebagai :

$$\boldsymbol{g(x) = sign\ (f(x))} \quad (6)$$

dengan $f(x) = (w^T x + b)$ dan w adalah bidang normal, sedangkan b merupakan poisi relatif terhadap koordinat pusat.

F. Pengukuran Performa Model

Model klasifikasi menggunakan machine learning memerlukan pengukuran performa untuk mengetahui akurasi, precision, recall dan F1-Score dari model. Proses pengukuran performa dengan menggunakan confusion matrix. Confusion matrix atau error matrix memetakan hasil prediksi dari machine learning terhadap data testing dengan label yang sebenarnya dari dataset. Hasil perbandingan ini disimpan dalam variabel True positive, false positive, true negative dan false negative. True Positive (TP) jika data termasuk kelas positif juga diprediksi oleh *machine learning* (ML) sebagai data kelas positif. False positive (FP) jika data kelas negatif namun diprediksi sebagai kelas positif. True Negative (TN) jika data kelas negatif juga diprediksi sebagai data negatif oleh ML. False Negative (FN) jika data kelas positif namun diprediksi sebagai data kelas negatif. Keempat variabel ini akan menentukan nilai *Accuracy, Recall, Precision dan F1-score*.

IV. HASIL PENELITIAN

Bagian ini akan membahas mengenai hasil penelitian guna menunjukkan kontribusi yang disebutkan pada bagian pertama.

A. Tahap preprocessing Dataset

Sebelum melakukan proses analisis, dokumen tweet harus melalui tahap preprocessing terlebih dahulu untuk memperoleh format data yang sesuai. Tahap preprocessing dokumen tweet terdiri dari 5 tahap, yaitu cleaning, case folding, tokenizing, stopwords removing, dan stemming. Pada tabel 3 akan ditunjukkan preprocessing data menggunakan komentar pertama pada tabel 1.

1) Cleaning: perbedaan proses cleaning hanya terlihat pada penghilangan angka, karena tidak ada hastag, username pada komentar.

2) Case Folding: pada *case folding*, dapat dilihat kata "Sedih" berubah menjadi "sedih". Sesuai dengan tujuan dari case folding yaitu merubah huruf besar menjadi huruf kecil.

3) Tokenizing: pada tabel 2 dapat dilihat, proses tokenizing memecah kalimat menjadi sekumpulan kata yang selanjutnya akan dicocokan dengan daftar stopword, jika terdapat kata yang sama, maka akan dihapus pada tahapan *stopword removal*.

4) Stopword Removal: pada tahap stopword removal, ada beberapa kata yang dihapus karena termasuk dalam daftar stopword Indonesia di library NLTK. Kata-kata tersebut adalah [*hari, lalu, belum, paling, kecil, masih, bisa, rasanya, jadi, mereka, ke].* Pada tabel 2 dapat diamati bahwa masih banyak kata tidak baku, juga kata yang umum dan tidak memiliki pengaruh seperti kata "yg", "bgt", "dapet" dan lain-lain yang bisa ditambahkan ke dalam daftar stopword, sehingga kata tersebut dapat dihapus pada proses stopword removal.

5) Stemming: proses stemming pada tabel 2, dengan merubah kata menjadi kata dasar seperti "dirawat" menjadi "rawat" dst.

TABEL III
TEKNIK PELABELAN DATA MANUAL

Proses	Hasil
Cleaning	Sedih bgt dapet kabar om adi meninggal dunia stlh hari dirawat grgr covid Istrinya meninggal setahun yg lalu Skrg anaknya yatim piatu Ketiganya belum nikah Yg paling kecil masih SMA Ngga bisa bayangin rasanya jadi mereka Gabisa nganter ayah ke peristirahatan trkhir.
Case Folding	sedih bgt dapet kabar om adi meninggal dunia stlh hari dirawat grgr covid istrinya meninggal setahun yg lalu skrg anaknya yatim piatu ketiganya belum nikah yg paling kecil masih sma ngga bisa bayangin rasanya jadi mereka gabisa nganter ayah ke peristirahatan trkhir
Tokenizing	['sedih', 'bgt', 'dapet', 'kabar', 'om', 'adi', 'meninggal', 'dunia', 'stlh', 'hari', 'dirawat', 'grgr', 'covid', 'istrinya', 'meninggal', 'setahun', 'yg', 'lalu', 'skrg', 'anaknya', 'yatim', 'piatu', 'ketiganya', 'belum', 'nikah', 'yg', 'paling', 'kecil', 'masih', 'sma', 'ngga', 'bisa', 'bayangin', 'rasanya', 'jadi', 'mereka', 'gabisa', 'nganter', 'ayah', 'ke', 'peristirahatan', 'trkhir']
Stopword Removal	['sedih', 'bgt', 'dapet', 'kabar', 'om', 'adi', 'meninggal', 'dunia', 'stlh', 'dirawat', 'grgr', 'covid', 'istrinya', 'meninggal', 'setahun', 'yg', 'skrg', 'anaknya', 'yatim', 'piatu', 'ketiganya', 'nikah', 'yg', 'sma', 'ngga', 'bayangin', 'gabisa', 'nganter', 'ayah', 'peristirahatan', 'trkhir']
Stemming	sedih bgt dapet kabar om adi tinggal dunia stlh rawat grgr covid istri tinggal tahun yg skrg anak yatim piatu tiga nikah yg sma ng gabayangin gabisa nganter ayah istirahat trkhir.

B. Hasil Pengukuran Empati Masyarakat

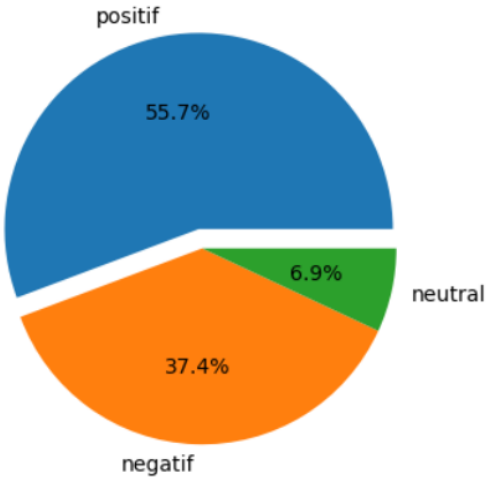
Berdasarkan penelitian yang telah dilakukan, dataset yang didapatkan sebanyak 2386 komentar, namun komentar yang mengandung empati hanya sebanyak 984 data. Hasil pengukuran dataset empati ditunjukkan pada gambar 2. Jumlah kelas positif sebanyak 548 atau sebesar 55,7% komentar masyarakat indonesia yang mengandung empati. Kelas negatif sebanyak 368 data, yaitu 37,4% tidak berempati, dan 6,9% atau sebanyak 68 komentar netral.

C. Model Prediksi Sentimen dan Pengujian Performa..

Pemodelan digunakan untuk membuat metode machine learning SVM belajar menentukan jenis komentar apakah berempati atau tidak berempati. Tahapan membangun model yang pertama adalah merubah dataset menjadi *vector space model* dengan metode ekstraksi fitur TF-IDF. Hasil ekstraksi fitur ditunjukan pada tabel 2.

Tahap kedua adalah membagi dataset menjadi data training dan data testing. Data training mengambil porsi sebanyak 80% dan data testing sebanyak 20% sehingga menghasilkan data training sebanyak 320 komentar dan data testing sebanyak 80 komentar.

```
positif    548
negatif    368
neutral     68
Name: Sentiment, dtype: int64
```



Gambar 2. Hasil pengukuran empati masyarakat

TABEL IIIII
EKSTRAKSI FITUR TF-IDF PADA DATASET EMPATI

abis	ada	adek	adik	ajar	amal	ampun	anak
0.16	0	0	0	0	0	0	0
0	0	0.572	0	0	0	0	0
0	0	0.249	0	0	0	0	0
0	0	0	0.323	0	0	0	0
0	0	0	0	0	0.227	0.201	0
0	0	0	0	0.191	0	0	0
0	0.221	0	0	0.18	0	0	0.29

Tahap ketiga adalah proses seleksi fitur untuk mengidentifikasi fitur yang berpengaruh dengan mengesampingkan fitur yang kurang relevan. Pada penelitian ini, jumlah fitur yang akan dibuang sebanyak 5%, 10%, 15% dan 20% dari total fitur. Nilai-nilai ini dipilih dengan harapan bahwa tidak terlalu banyak fitur yang dibuang dan tetap mempertahankan informasi penting dalam data.

Pada penggunaan ReliefF untuk mendapatkan bobot dari setiap fitur, perlu ditentukan terlebih dahulu satu fitur untuk dijadikan nilai tengah (R), fitur ini dipilih secara acak. Setelah nilai R sudah ditentukan, ReliefF akan mengindentifikasi tetangga terdekat dari R yang disebut near hit dan near miss. Relevansi fitur dengan nilai R dapat dilihat berdasarkan skor yang diperoleh. Fitur yang sangat relevan adalah fitur yang memiliki skor tertinggi,

sedangkan fitur yang memiliki nilai rendah adalah fitur yang tidak relevan. Tahap pertama yang dilakukan adalah mengimport library yang dibutuhkan. Library yang digunakan dalam proses seleksi fitur ReliefF adalah *skrebate*.

Tahap kedua pembuatan model dengan menggunakan *machine learning* SVM untuk mengklasifikasikan data apakah termasuk ke dalam kelas positif atau negatif serta melihat tingkat akurasi yang dihasilkan. Kernel yang digunakan pada SVM ini adalah kernel *Radial Basis Function* (RBF). Sebelum melakukan pengujian dengan seleksi fitur, dilakukan uji coba untuk melihat hasil performa model menggunakan SVM tanpa seleksi fitur. Hasil penelitian dapat diamati pada penjelasan berikut.

1. SVM Tanpa Seleksi Fitur

Skenario uji coba 1 adalah melakukan klasifikasi pada data dengan metode Support Vector Machine tanpa menggunakan seleksi fitur. Tujuan dari pengujian ini untuk mengetahui nilai akurasi, precision, recall, dan f1-score tanpa penggunaan seleksi fitur. *Confusion matrix* dari klasifikasi analisis sentimen menggunakan data testing sebanyak 80 data, menghasilkan 31 *true positive* (TP), 0 *false positive* (FP), 14 *false negative* (FN), dan 35 *true negative* (TN).

```
svc = SVC(kernel='rbf', random_state = 42)
svc.fit(X_train, y_train)
predicted = svc.predict(X_test)
expected = y_test
print(metrics.confusion_matrix(expected, predicted))
```

```
[[31  0]
 [14 35]]
```

Gambar 3. Hasil running subprogram *Confusion matrix*

Berdasarkan matrik konfusi yang diperoleh sesuai Gambar 3, maka bisa dilakukan perhitungan akurasi, *precision*, *recall*, dan *f1-score*.

$$\text{Akurasi} = \frac{(TP+TN)}{(TP+TN+FP+FN} \times 100\% \quad (7)$$

$$Precision = = \frac{(TP)}{(TP+FP)} \times 100\% \quad (8)$$

$$Recall = = \frac{(TP)}{(TP+FN)} \times 100\% \quad (9)$$

$$\text{F1} - \text{score} = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (10)$$

Nilai Akurasi yang diperoleh adalah 82.5% ≈ 83%, nilai Precision 100%, nilai Recall 68.8% ≈ 69% dan nilai F1-score adalah 81,6% ≈ 82%.

2. SVM Dengan Seleksi Fitur ReliefF

Pada penelitian ini, seleksi fitur ReliefF akan mengeleminasi 5%, 10%, 15% dan 20% dari total keseluruhan fitur. Hasil ekstraksi fitur menggunakan TF-IDF pada dataset menghasilkan 2218 fitur, sehingga 5% dari total fitur adalah 2107 fitur. 1996 fitur untuk 10%, 1885 fitur untuk 15% dan 1774 fitur untuk 20%. Hasil confusion matrix dan performa model menggunaan seleksi fitur dapat dilihat pada tabel 4 dan 5. Pada tabel 4 dapat dilihat bahwa semakin banyak fitur dibuang, tidak menjadi jaminan akurasi semakin bagus.

TABEL IVV
CONFUSION MATRIX SVM DAN SELEKSI FITUR RELIEFF

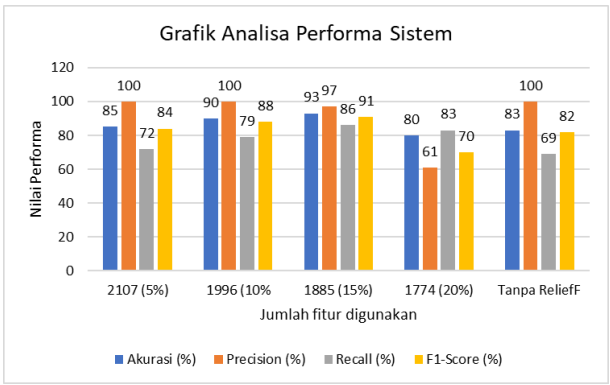
Jumlah Fitur	TP	FP	FN	TN
2107 (5%)	31	0	12	37
1996 (10%	31	0	8	41
1885 (15%)	30	1	5	44
1774 (20%)	19	12	4	45

Pada tabel 4 dapat dilihat confusion matrix yang dihasilkan dengan menerapkan seleksi fitur. Semakin tinggi nilai TP dan TN maka akan menghasilkan nilai akurasi yang semakin tinggi. Precision akan semakin tinggi jika nilai FP semakin kecil. Recall dipengaruhi oleh nilai FN yang semakin kecil dan nilai TP yang semakin besar. F1-Score dipengaruhi oleh nilai *precision* dan *recall*.

TABEL V
PERFORMA MODEL SVM DAN SELEKSI FITUR RELIEFF

Jumlah Fitur	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
2107 (5%)	85	100	72	84
1996 (10%	90	100	79	88
1885 (15%)	**93**	97	86	91
1774 (20%)	80	61	83	70

Pada tabel 5 dapat diamati bahwa akurasi yang tertinggi adalah saat dilakukan klasifikasi dengan menerapkan seleksi fitur ReliefF 15% yaitu sebesar 93%. Sedangkan akurasi terendah adalah saat dilakukan klasifikasi dengan menerapkan seleksi fitur ReliefF 20% yaitu sebesar 80%. Hal itu menunjukkan bahwa saat fitur dieleminasi sebanyak 15%, fitur yang tidak relevan berhasil dieleminasi sehingga *machine learning* lebih mudah melakukan klasifikasi. Sedangkan saat melakukan klasifikasi dengan mengeleminasi fitur sebanyak 20%, selain fitur yang tidak relevan, fitur yang relevan juga tereleminasi sehingga *machine learning* kekurangan fitur yang penting selama proses klasifikasi.



Gambar 4. Grafik analisa performa sistem

Pada Gambar 4, dapat diketahui bahwa nilai akurasi tertinggi adalah akurasi yang diperoleh dari SVM dengan seleksi fitur ReliefF 15% yaitu sebesar 93%. Nilai precision tertinggi diperoleh dari SVM tanpa seleksi fitur, SVM dengan ReliefF 5%, dan SVM dengan ReliefF 10% yaitu sebesar 100%, nilai recall tertinggi adalah akurasi yang diperoleh dari SVM dengan seleksi fitur ReliefF 15% yaitu sebesar 86%.

Dari hasil analisa, kesalahan prediksi kemungkinan disebabkan karena belum adanya penanganan untuk makna kalimat, sehingga ekstraksi fitur masih bersifat leksikon belum bersifat semantik. Contohnya dalam suatu kalimat yang ada kata 'belum' dan 'nikah', akan berbeda sentimennya jika fitur yang diambil adalah 1 term ('belum' dan 'nikah') dengan 2 term ('belum nikah').

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa perkembangan teknologi mungkin telah membentuk mental individualis, namun bukan berarti telah membentuk masyarakat tanpa empati seluruhnya, masih banyak masyarakat yang berempati dengan penderitaan orang lain karena pandemi covid. Hal ini berdasarkan jumlah komentar berempati (positif) sebanyak 548 atau sebesar 55,7% dari toal komentar masyarakat indonesia yang dicrawling dari bulan November-Desember 2020.

Penelitian ini juga membuktikan bahwa seleksi fitur sangat berpengaruh terhadap peningkatan performa model machine learning SVM. Akurasi SVM meningkat sampai 10% dari penggunaan SVM tanpa seleksi fitur. Akurasi terbaik didapatkan pada saat fitur dikurangi sebesar 15% dari total fitur. Semakin banyak fitur yang dikurangi tidak berbanding lurus dengan peningkatan akurasi. Hal ini dikarenakan pengurangan fitur yang terlalu banyak dapat menyebabkan dataset kehilangan fitur yang penting sehingga tidak dapat mengklasifikan dengan baik. Pada saat fitur ditingkatkan menjadi 20%, akurasi mengalami penurunan sebesar 13%.

REFERENSI

[1] I. Imamah, H. Husni, E. M. Rohman, I. O. Suzanti, and F. A. Mufarroha, "Text mining and Support Vector Machine for Sentiment Analysis of tourist Reviews in Bangkalan Regency," *Journal of Physics: Conference Series*, vol. 1477, no. 2, pp. 0–6, 2020.

[2] B. S. Rintyarna, H. Kuswanto, R. Sarno, and E. K. Rachmaningsih, "Modelling Service Quality of Internet Service Providers during COVID-19 : The Customer Perspective Based on Twitter Dataset," *Informatics*, vol. 9, pp. 1–12, 2022.

[3] F. H. Rachman, I. Imamah, and B. S. Rintyarna, "Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning," in *International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 2022, pp. 23–27.

[4] H. Junawan and N. Laugu, "Eksistensi Media Sosial,Youtube, Instagram dan Whatsapp Ditengah Pandemi Covid-19 Dikalangan Masyarakat Virtual Indonesia," *Baitul 'Ulum: Jurnal Ilmu Perpustakaan dan Informasi*, vol. 4, no. 1, pp. 41–57, 2020.

[5] Q. Fadhillah, "GAMBARAN EMPATI GENERASI MILLENIAL DI PEKANBARU," *Journal of Islamic and Contemporary Psychology (JICOP)*, vol. 1, no. 2, pp. 61–66, 2021.

[6] I. Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion," in *The 6th Information Technology International Seminar (ITIS)*, 2020, pp. 238–242.

[7] F. H. Rachman, R. Sarno, and C. Fatichah, "Hybrid Approach of Structural Lyric and Audio Segments for Detecting Song Emotion," *International Journal of Intelligent Engineering & Systems*, vol. 13, no. 1, 2020.

[8] A. N. Rohman, E. Utami, and S. Raharjo, "Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing," *Eksplora Informatika*, vol. 9, no. 1, pp. 70–76, 2019.

[9] D. J. Haryanto, L. Muflikhah, and M. A. Fauzi, "Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, vol. 2, no. 9, pp. 2909–2916, 2018.

[10] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, "Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine," *Komputek*, vol. 3, no. 2, p. 52, 2019.

[11] M. Ichwan, I. A. Dewi, and Z. M. S, "Klasifikasi Support Vector Machine (SVM) Untuk Menentukan TingkatKemanisan Mangga Berdasarkan Fitur Warna," *MIND Journal*, vol. 3, no. 2, pp. 16–23, 2019.

[12] Y. Alapati and K. Sindhu, "Combining Clustering with Classification: A Technique to Improve Classification Accuracy," *International Journal of Computer Science Engineering*, vol. 5, no. 06, pp. 336–338, 2016.

[13] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-square Statistics Based Feature Selection," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 160–163, 2018.

[14] O. Somantri and M. Khambali, "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, vol. 6, no. 3, 2017.

[15] A. Elsayya, M. Al-Dhaifallah, and N. A, "Features Selection for Arrhythmia Diagnosis using Relief-F Algorithm and Support Vector Machine," pp. 461–468, 2017.

[16] B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 10, no. 2, pp. 118–123, 2021.

[17] I. Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion," in *Information Technology International Seminar (ITIS)*, 2020, pp. 238–242.

[18] Z. M. Kesuma, "Feature Selection Data Indeks Kesehatan Masyarakat Menggunakan Algoritma Relief-F," *Statistika*, vol. 11, no. 1, pp. 61–66, 2011.

[19] A. Elsayya, M. Al-Dhaifallah, and A. M. Nassef, "Features Selection for Arrhythmia Diagnosis using Relief-F Algorithm and Support Vector Machine," pp. 461–468, 2017.

[20] L. Silva, B. Bispo, J. Paulo, L. Silva, B. Bispo, and J. Paulo, "Features Selection Algorithms for Classification of Voice Signals," *Procedia Computer Science*, vol. 181, no. 2020, pp. 948–956, 2021.

[21] D. Jain and V. Singh, "An Efficient Hybrid Selection model for Dimensionality Reduction," *Procedia Computer Science*, vol. 132, no. Iccids, pp. 333–341, 2018.

[22] F. P. B. Muhamad, D. O. Siahaan, and C. Fatichah, *Perbaikan Prediksi Kesalahan Perangkat Lunak Menggunakan Seleksi Fitur dan Cluster-Based Classification*, vol. 6, no. 3. 2017.

[23] N. M. G. D. Purnamasari, M. A. Fauzi, Indriarti, and L. S. Dewi, "Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, vol. 2, no. 11, pp. 5326–5332, 2018.

[24] W. Wijanarto and R. Puspitasari, "Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus," *Eksplora Informatika*, vol. 9, no. 1, pp. 50–59, 2019.

[25] F. A. Novianti and S. W. Purnami, "Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi," *Jurnal SAINS dan Seni ITS*, vol. 1, no. 1, pp. D147–D152, 2012.