JURNAL ILMIAH FIFO **P-ISSN 2085-4315 / E-ISSN 2502-8332**
DOI: http://dx.doi.org/10.22441/fifo.2023.v15i2.001

Sentimen Analisis Mengenai Polusi Udara Menggunakan Algoritma Support Vector Machine dan Random Forest

Lukman Hakim*[1], Muhammad Variansjah Dalimunthe[2], Chyquitha Danuputri[3], Destriana Widyaningrum[4]

[1,2]Fakultas Ilmu Komputer, Universitas Mercu Buana
Jl. Meruya Selatan No.1, Jakarta Barat 11650
[3,4]Informatika, Universitas Bunda Mulia
Jl. Lodan Raya No. 2, Ancol, Jakarta 14430
[1*]lukman_hakim@mercubuana.ac.id, [2]41519010191@student.mercubuana.ac.id,
[3]chyquitha@gmail.com, [4]des3ana@gmail.com

*) Corresponding author

Abstract

Air pollution is the contamination of indoor or outdoor environments with chemical, physical, or biological substances that change the natural properties of the atmosphere. Domestic incinerators, cars, motorbikes, combustion products from factory processing, waste burning, and forest fires are common sources of air pollution. In Indonesia, there is no doubt that air pollution occurs because of the many forest fires in Indonesia. As a result of this case, many people's opinions differ. Various sentiments occur in cyberspace, one of which is Twitter. Twitter is the social media that accommodates the most various kinds of positive, negative and neutral opinions. Therefore, researchers want to solve the problem by implementing the SVM and Random Forest algorithms. The dataset was obtained from scrapping results using tweet harvest. The data obtained was 5545 tweets. By dividing the dataset model by 80% and 20%, the results showed that the accuracy of the SVM algorithm was better than the Random Forest algorithm. The accuracy of the SVM algorithm is 83% while the Random Forest algorithm is 81%..

***Keyword**: random forest, svm, twitter, sentiment analyst, pollution*

Abstrak

Polusi udara adalah kontaminasi lingkungan dalam atau luar ruangan dengan zat kimia, fisik, atau biologis yang mengubah sifat alami atmosfer. Insinerator domestik, mobil, motor, hasil pembakaran produksi pengolahan pabrik, pembakaran sampah, dan kebakaran hutan merupakan sumber polusi udara yang umum. Di Indonesia, tidak diragukan lagi kalau polusi udara terjadi karena banyaknya kebakaran hutan di Indonesia. Akibat kasus tersebut, banyak opini masyarakat yang berbeda-beda. Berbagai sentiment terjadi di dunia maya, salah satunya Twitter. Twitter adalah social media yang paling banyak menampung berbagai macam opini positif, negatif maupun netral. Oleh karena itu, peneliti ingin memecahkan masalah dengan implementasi algoritma SVM dan Random Forest. Dataset didapatkan dari hasil scrapping menggunakan tweet harvest. Data yang diperoleh didapatkan sebanyak 5545 tweet. Dengan membagi model dataset 80% dan 20%, hasil didapat bahwa akurasi algoritma SVM lebih baik dari algoritma Random Forest. Akurasi dari algoritma SVM sebesar 83% sedangkan algoritma Random Forest sebesar 81%.

Kata Kunci: *random forest, svm, twitter, sentiment analyst, pollution*

I. Pendahuluan

Polusi udara merupakan salah satu fenomena yang disebabkan oleh ulah manusia. Kota besar menyumbang polusi yang besar dari pembakaran sisa produksi industri, kendaraan bermotor terlalu banyak atau banyaknya pabrik yang tidak bisa mengatasi limbahny, Laporan *Air Quality Life Index (AQLI)* menyebut Indonesia sebagai salah satu dari enam negara penyumbang polusi udara global terbesar. Sementara itu, beberapa kelompok masyarakat sipil sedang mempertimbangkan tuntutan hukum *class action* karena masalah polusi udara tampaknya semakin parah. Data IQAir pertanggal 17 februari 2024, Indonesia peringkat 77 dari 111 negara untuk kondisi polusi udara, dengan tingkat polutan PM 2.5 konsentrasi terendah 19.5 mikrogram, ini masih kategori sedang, bukan baik. Standar WHO level polusi udara PM 2.5 untuk kategori baik 0-15,5 mikrogram/meter^3[1]. Kondisi inilah perlu dilakukan penelitian mengenai kondisi polusi udara terhadap masyarakat dengan mengukur tanggapan positif atau negatif, hal tersebut menjadi masukan bagi pemerintah dan masyarakat dalam penanggulangi polutan di udara. Indonesia, bersama dengan Tiongkok, India, Pakistan, Bangladesh, dan Nigeria, menyumbang 75% dari total beban polusi udara global karena tingginya tingkat polusi udara dan jumlah penduduk yang besar [2] .

Perdebatan mengenai polusi di Indonesia saat ini sedang menarik perhatian masyarakat. Beberapa pihak menyalahkan pembangkit listrik tenaga uap (PLTU) berbahan bakar batu bara sebagai penyebab buruknya kualitas udara di kota Jakarta dan sekitarnya. Namun ada juga yang berpendapat bahwa polusi berasal dari kendaraan bermotor. Sepeda motor menghasilkan beban pencemaran per penumpang tertinggi dibandingkan dengan mobil penumpang berbahan bakar bensin, mobil penumpang *diesel*, mobil penumpang, dan bus [3]. Data menunjukkan sektor transportasi menyumbang 44% konsumsi bahan bakar di Jakarta, disusul sektor energi sebesar 31%, produksi industri sebesar 10%, disusul sektor perumahan sebesar 14%, dan sektor komersial sebesar 1%. Dari sisi penghasil emisi karbon monoksida *(CO)* terbesar, sektor transportasi diperkirakan menyumbang 96,36% atau 28.317 ton per tahun, disusul pembangkit listrik 1,76% (5.252 ton per tahun) dan industri 1,25%. 3.738 ton. setiap tahun. Sepeda motor juga menghasilkan beban pencemaran per penumpang tertinggi dibandingkan dengan mobil penumpang berbahan bakar bensin, mobil penumpang *diesel*, mobil penumpang dan bus. Jumlah penduduk 78 jiwa, jumlah kendaraan bermotor di DKI Jakarta sebanyak 24,5 juta kendaraan atau tumbuh 1.046.837 sepeda motor per tahun. Diperkirakan terdapat lebih dari 20 juta kendaraan bermotor di Jakarta. Belum termasuk kendaraan yang masuk dan keluar Jakarta, termasuk bus dan truk, yang turut menyumbang polusi di ibu kota. Permasalahan terjadinya polusi udara dari hasil sisa pengolahan industri, kendaraan bermotor menjadi respon pedas dari masyarakat terhadap kebijakan pemerintah, maka dilakukan analisis sentimen pencemaran udara pada DKI Jakarta dengan algoritma Support Vector Machine, untuk mengetahui nilai positif, negatif dan netra terhadap masukan masyarakat ke pemerintah.

II. Metodologi Penelitian

2.1 Jenis Penelitian

Metode penelitian kuantitatif digunakan pada penelitian ini. Penelitian kuantitatif merupakan jenis penelitian yang berfokus pada analisis data numerik untuk mencapai prediksi pandangan masyarakat Indonesia terhadap polusi udara di Indonesia berdasarkan keseluruhan data yang dikumpulkan. Penerapan metode kuantitatif ini bersifat objektif karena penelitian ini berfokus pada perhitungan matematis dan perhitungan terkait hipotesa.

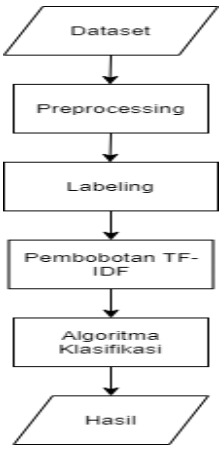
2.2. Metode Pengumpulan Data

Objek data yang akan diteliti adalah berbagai pesan atau *tweet* berbahasa Indonesia di *platform* media sosial *Twitter* dengan topik pandangan masyarakat Indonesia terhadap polusi udara di Indonesia. Analisis sentimen adalah penerapan teknologi pemrosesan bahasa alami yang melatih perangkat lunak komputer untuk memahami teks dengan cara yang sama seperti manusia[16].

2.3. Tahapan Penelitian

Penelitian ini dilakukan dalam beberapa tahap, yaitu mengumpulkan data, *pre-processing*, *labeling*, kemudian *training model*, implementasi algoritma, dan yang terakhir evaluasi perbandingan algoritma. Tahapan penelitian ini disajikan pada Gambar 1. Tahapan proses *sentiment analysis* yang dilakukan dapat di dibagi menjadi beberapa yaitu :

DOI: http://dx.doi.org/10.22441/fifo.2023.v15i2.001

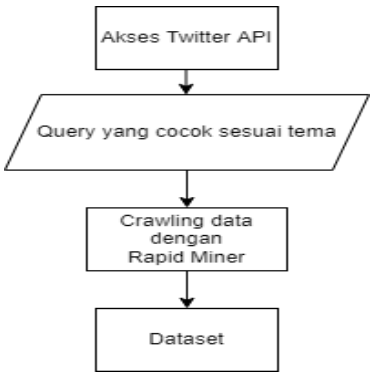


Gambar 1. Tahapan proses sentiment analysis

Tahapan awal pada penelitian ini mengelola hasil pengumpulan data secara kuantitatif, dengan memproses data melalui API Twitter menggunakan tools RapidMiner. Tahap *pre-processing* dan labeling pada Google Colab menggunakan tools Python dan R. Dalam penelitian ini juga dilakukan eksperimen dengan *Support Vector Machine* (SVM) untuk mengetahui algoritma yang memiliki tingkat akurasi terbaik.

a. Pengumpulan Data

Dataset diambil dari Twitter dapat dilihat pada gambar 2. Sebagai berikut :



Gambar 2. Perolehan dataset

Pengumpulan data yang akan digunakan untuk penelitian berasal dari berbagai tempat seperti media sosial, termasuk *Twitter*. Untuk penelitian ini, pengumpulan data mentah melalui *Twitter*, yang akan memudahkan peneliti mendapatkan dataset sebanyak 5545 dari *Tweet* menggunakan API yang disediakan oleh *Twitter* secara resmi.

Twitter API ini mempunyai banyak fitur yang diantaranya fungsi yang akan di gunakan adalah mengekstrak *Tweet* untuk dataset yang berisi opini orang Indonesia mengenai polusi udara di Indonesia untuk memudahkan proses pengumpulannya. Pengumpulan data menggunakan API. Pengguna *library python Tweet-Harvest* untuk mengekstrak *Tweet* secara massal dengan cepat. Pencarian kueri yang digunakan untuk memfilter kumpulan data cocok dengan topik #Polusi, #PolusiUdara.

untuk menentukan kata kunci pencarian *tweet* yang diinginkan. Pada saat penghapusan, *script* akan memasukkan hasil penghapusan ke dalam daftar dan kemudian dengan menggunakan *library Pandas* dapat mengubah daftar tersebut menjadi kerangka data yang kemudian dapat di ekspor sebagai .csv.

Tabel 1. Dataset tweet

No	Text Tweet
1	@Pemudabaja Kak polusi udara udh sampe ke kotaÂ² kecil Sumatra. Capek bgt hirup kabut asap. ðŸ˜ But anyway thank you kak~ Enjoy your weekend too ðŸŒ¸

No	Text Tweet
2	Polusi Jakarta semakin parah wajib banget punya air purifier penjernih udara buat si kecil https://t.co/rrGorE2Bxi #airpurifier #dysonairpurifier #PolusiJAKARTA #polusiudara https://t.co/UM71rCGnGX
3	Cukup Stay With Alam Hidup Sendirian Tenang Dan Damai Udara Masih Dingin Dan Sejuk Serta Polusi Udara Lebih Minim And Lagian Udah Pernah Ngerasain Hidup Hedon DiKota Metropolitan https://t.co/ixrCZDRgq3
4	Polusi udara menjadi masalah di banyak daerah, terutama di ibukota. Hal ini bisa mengancam kesehatan paru-paru masyarakat. ðŸ˜· Untuk mencegah dan mengurangi dampak polusi, detikers bisa mengkonsumsi makanan sehat untuk paru-paru. Berikut 8 makanan dan minuman untuk membersihkanâ€¦ https://t.co/CyiVnXP4r3
...	Bhabinkamtibmas Kel. Kunciran Indah Aiptu T.Raharjo melaksanakan kegiatan pemantauan titik api diwilayah Kel. Kunciran Indah, dari hasil pemantauan dilokasi tersebut ditemukan tumpukan daun kering dibakar yang menimbulkan asap dan polusi udara
5545	@tanyakanrl Ini suaminya yg nangis itu bukann si Kasiann banget yg di fitnah sm polusi udara saat itu wkwk

b. Pre-Processing

Setelah memperoleh kumpulan data mentah, proses selanjutnya adalah melakukan *pre-processing*, yaitu bagian dari proses data mining untuk memodifikasi data mentah yang diperoleh agar dapat dipahami dan diproses lebih lanjut. untuk memastikan bahwa data yang akan digunakan untuk melatih model akurat dan konsisten sehingga tidak mempengaruhi proses pelatihan model, untuk penelitian sentimen analisis yang terfokus terhadap Analisa opini dari teks proses *pre-processing* yang dilakukan adalah sebagai berikut[16]:

1) Cleaning

Merupakan tahap pertama yang dilakukan yaitu untuk menghapus teks yang tidak penting digunakan pada penelitian seperti *created_at, id_str, quote_count, reply_count, retweet_count, lang, user_id_str, conversation_id_str, username, url*. Hal ini dilakukan untuk meningkatkan kinerja algoritma dan mempermudah proses analisis sentiment. Tabel 2 merupakan proses cleaning datase sebagai berikut:

Contoh :

Hwaaa polusi nya emang parah... Nyoba berangkat nggak pake masker ternyata udara nya nggak sehat Akhir nya berhenti dulu pake masker.

Hasil cleaning : polusinya emang parah berangkat nggak pake masker ternyata udaranya nggak sehat akhirnya berhenti dulu pake masker.

Tabel 2. Proses cleaning dataset

tweet	tweet_preprocessing
kak polusi udara udh sampe ke kota kecil sumatra capek bgt hirup kabut asap *but anyway thank you* kak *enjoy your weekend too*	kak polusi udara udh sampe ke kota kecil sumatra capek bgt hirup kabut asap *but anyway thank you kak enjoy your weekend too*
polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat si kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat si kecil
cukup *stay with* alam hidup sendirian tenang dan damai udara masih dingin dan sejuk serta polusi udara lebih minim and lagian udah pernah ngerasain hidup hedon dikota metropolitan	cukup *stay with* alam hidup sendirian tenang dan damai udara masih dingin dan sejuk serta polusi udara lebih minim and lagian udah pernah ngerasain hidup hedon dikota metropolitan
Banyak ruang hijau diperlukan untuk mengurangi polusi udara di Jakarta.	Banyak ruang hijau diperlukan untuk mengurangi polusi udara di Jakarta .
lbh ga seru lg ke jakarta kmn mn bukanny menghirup udara mlh menghirup polusi	lebih enggak seru lagi ke jakarta kemana mana bukanny menghirup udara malah menghirup polusi
jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt	jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt
ayo kurangi polusi udara dengan menggunakan transportasi publik	ayo kurangi polusi udara dengan menggunakan transportasi publik
wfh menjadi langkah tepat untuk kurangi polusi udara di jakarta	wfh menjadi langkah tepat untuk kurangi polusi udara di jakarta

2) Case Folding

Tahapan ini semua huruf kapital akan diubah menjadi huruf kecil sehingga kata seperti "Kotor" menjadi "kotor" proses ini digunakan untuk menstandarkan format teks agar analisis sentiment tidak membedakan huruf "A" dengan "a".
Contoh:
Hwaaa polusi nya emang parah... Nyoba berangkat nggak pake masker ternyata udara nya nggak sehat Akhir nya berhenti dulu pake masker.
Hasil folding
polusi nya emang parah nyoba berangkat nggak pake masker ternayata udaranya nggak sehat akhirnya berhenti dulu pake masker.

3) *Tokenizing*

Tokenizing adalah proses pertukaran data sensitif dengan data non-sensitif yang disebut "token" yang dapat digunakan dalam database atau sistem internal tanpa memasuki ruang lingkup [19].
Contoh:
polusi nya emang parah nyoba berangkat nggak pake masker ternayata udaranya nggak sehat akhirnya berhenti dulu pake masker.
Hasil Tokenizing: polusi,nya, emang, parah, nyoba, berangkat, nggak, pake, masker, ternayata, udaranya, nggak, sehat, akhirnya, berhenti, dulu, pake, masker.

Tabel 3. Tokenisasi dataset

tweet	tweet_normalisasi	stopwords	token
kak polusi udara udh sampe ke kota kecil sumatra capek bgt hirup kabut asap *but anyway thank you kak enjoy your weekend too*	kak polusi udara sudah sampai ke kota kecil sumatra capek banget hirup kabut asap *but anyway thank you kak enjoy your weekend too*	kak polusi udara sampai kota kecil sumatra capek banget hirup kabut asap *but anyway thank you kak enjoy your weekend too*	['kak', 'polusi', 'udara', 'sampai', 'kota', 'kecil', 'sumatra', 'capek', 'banget', 'hirup', 'kabut', 'asap', 'but', *'anyway', 'thank', 'you', 'kak', 'enjoy', 'your', 'weekend', 'too'*]
polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat si kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil	['polusi', 'jakarta', 'semakin', 'parah', 'wajib', 'banget', 'punya', *'air', 'purifier'*, 'penjernih', 'udara', 'buat', 'sih', 'kecil']
atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	['atasi', 'polusi', 'udara', 'dki', 'jakarta', 'butuh', 'banyak', 'ruang', 'terbuka', 'hijau']

4) *Stopword*

Langkah ini meliputi menghilangkan kata-kata yang tidak mengandung informasi emosional seperti kata aku, aku, dia. Langkah ini didukung dengan kosakata yang berisi daftar *stopword*, yang kemudian dikeluarkan dari kumpulan data.

5) *Stemming*

Tahap stemming adalah tahap mencari root (bentuk dasar) dari tiap kata. Pada tahap ini, dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama [20].

6) Remove Duplicate

Tahapan ini dilakukan penghapusan data yang sama, sehingga tidak adanya duplikasi data pada saat di olah ke tahap selanjutnya.

Tabel 4, merupakan hasil dari normalisasi, stopword, token, labeling dari tahapan datamining untuk menghasilkan dataset yang sudah bersih dan mengurangi tingkat dupikasi data.

Tabel 4. Stopword Removal

tweet	tweet_normalisasi	stopwords
kak polusi udara udh sampe ke kota kecil sumatra capek bgt hirup kabut asap but *anyway thank you* kak *enjoy your weekend too*	kak polusi udara sudah sampai ke kota kecil sumatra capek banget hirup kabut asap *but anyway thank you kak enjoy your weekend too*	kak polusi udara sampai kota kecil sumatra capek banget hirup kabut asap *but anyway thank you kak enjoy your weekend too*

polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat si kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil
lbh ga seru lg ke jakarta kmn mn bukanny menghirup udara mlh menghirup polusi	lebih enggak seru lagi ke jakarta kemana mana bukanny menghirup udara malah menghirup polusi	lebih enggak seru ke jakarta mana bukanny menghirup udara malah menghirup polusi
jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt	jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt	jakarta dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan naik mrt

c. Labeling

Tahapan *pre-processing* dilakukan, selanjutnya melakukan labeling terhadap 5545 dataset. Labeling sentimen bertujuan untuk mengetahui opini masyarakat terhadap suatu isu topik tersebut. Labeling dibagi menjadi 2 yaitu positif dan negatif. Data yang diolah dengan Bahasa R di google collaboratory, yaitu data yang sudah bersih dan hanya memiliki kata yang bermakna. Kemudian data tersebut dicocokan ke dalam kamus korpus bahasa Indonesia sentiment negatif dan positif yang sudah disiapkan sebelumnya dalam format file csv. Kalimat yang terdeteksi akan diberi nilai, untuk kata positif diberi nilai 1, untuk kata negatif di beri nilai -1, dan jika tidak ditemukan kata tersebut di korpus negatif dan positif maka diberi nilai 0. Selanjutnya melakukan penghitungan skor dengan menjumlahkan semua nilai dari kata. Jika skor bernilai < 0 maka tweet dilabeli negatif dan jika jumlah skor >=0 maka tweet dilabeli positif. Ada berbagai contoh dari kelas positif yang berisi pujian, saran, masukan, dan cerminan emosi positif seperti puas, senang, dan bahagia. Sedangkan kelas negatif berisi keluhan, kalimat sindiran, kritik, dan cerminan emosi negatif seperti amarah, kesal, dan kecewa. Labeling manual dilakukan dengan memberikan label pada setiap tweet yang sesuai dengan kata tersebut apakah memiliki kata positif ataupun negatif pada satiap tweet. Pada kamus lexicon yang terdapat 3609 sentiment positif dan 6609 negatif. Disajikan pada Tabel 5 hasil dari labeling dari pegabungan tweet dan kamus lexicol :

Tabel 5. Hasil Procesing dengan Labeling

tweet	tweet_normalisasi	stopwords	token	polarity_score	polarity
kak polusi udara udh sampe ke kota kecil sumatra capek bgt hirup kabut asap *but anyway thank you* kak *enjoy your weekend too*	kak polusi udara sudah sampai ke kota kecil sumatra capek banget hirup kabut asap *but anyway thank you* kak *enjoy your weekend too*	kak polusi udara sampai kota kecil sumatra capek banget hirup kabut asap *but anyway thank you* kak *enjoy your weekend too*	['kak', 'polusi', 'udara', 'sampai', 'kota', 'kecil', 'sumatra', 'capek', 'banget', 'hirup', 'kabut', 'asap', *'but', 'anyway', 'thank', 'you',* 'kak', *'enjoy', 'your', 'weekend', 'too'*]	-7	*negative*
polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat si kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil	polusi jakarta semakin parah wajib banget punya *air purifier* penjernih udara buat sih kecil	['polusi', 'jakarta', 'semakin', 'parah', 'wajib', 'banget', 'punya', *'air', 'purifier'*, 'penjernih', 'udara', 'buat', 'sih', 'kecil']	-8	*negative*
cukup *stay with* alam hidup sendirian tenang dan damai udara masih dingin dan sejuk serta polusi udara lebih minim and lagian udah pernah ngerasain hidup hedon dikota metropolitan	cukup *stay with* alam hidup sendirian tenang dan damai udara masih dingin dan sejuk serta polusi udara lebih minim and lagian sudah pernah merasai hidup hedon dikota metropolitan	cukup *stay with* alam hidup sendirian tenang damai udara dingin sejuk polusi udara lebih minim and lagian pernah merasai hidup hedon dikota metropolitan	['cukup', *'stay', 'with'*, 'alam', 'hidup', 'sendirian', 'tenang', 'damai', 'udara', 'dingin', 'sejuk', 'polusi', 'udara', 'lebih', 'minim', 'and', 'lagian', 'pernah', 'merasai', 'hidup', 'hedon', 'dikota', 'metropolitan']	-17	*negative*

tweet	tweet_normalisasi	stopwords	token	polarity_score	polarity
atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	atasi polusi udara dki jakarta butuh banyak ruang terbuka hijau	['atasi', 'polusi', 'udara', 'dki', 'jakarta', 'butuh', 'banyak', 'ruang', 'terbuka', 'hijau']	-9	*negative*
lbh ga seru lg ke jakarta kmn mn bukanny menghirup udara mlh menghirup polusi	lebih enggak seru lagi ke jakarta kemana mana bukanny menghirup udara malah menghirup polusi	lebih enggak seru ke jakarta mana bukanny menghirup udara malah menghirup polusi	['lebih', 'enggak', 'seru', 'ke', 'jakarta', 'mana', 'bukannya', 'menghirup', 'udara', 'malah', 'menghirup', 'polusi']	-9	*negative*
jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt	jakarta masih dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan dengan naik mrt	jakarta dihantui masalah kualitas udara ayo kurangi polusi amp kemacetan naik mrt	['jakarta', 'dihantui', 'masalah', 'kualitas', 'udara', 'ayo', 'kurangi', 'polusi', 'amp', 'kemacetan', 'naik', 'mrt']	1	*positive*
wfh menjadi langkah tepat untuk kurangi polusi udara di jakarta	wfh menjadi langkah tepat untuk kurangi polusi udara di jakarta	wfh menjadi langkah tepat kurangi polusi udara jakarta	"['wfh', 'menjadi', 'langkah', 'tepat', 'kurangi', 'polusi', 'udara', 'jakarta']"	0	*neutral*

d. Implementasi Algoritma

Setelah model pelatihan dibuat dan prapemrosesan selesai, langkah selanjutnya adalah mengimplementasikan algoritma *Support Vector Machine* dan *Random Forest* dengan mengimpor perpustakaan ke *Python*. *Support Vector Machine*? Tujuan dari algoritma mesin vektor pendukung adalah untuk menemukan *hyperplane* dalam ruang berdimensi N (N - jumlah objek) yang mengklasifikasikan titik data dengan jelas[17]. *Random Forest* adalah salah satu algoritma yang paling populer dan umum digunakan oleh para ilmuwan data. *Random forest* adalah algoritma pembelajaran mesin terawasi yang banyak digunakan dalam masalah klasifikasi dan regresi. Ini membangun pohon keputusan pada sampel yang berbeda dan mengambil suara mayoritas untuk klasifikasi dan rata-rata jika terjadi regresi[18].

e. Evaluasi dan Perbandingan Algoritma

Pada tahap ini, menjelaskan hasil yang diperoleh yaitu tingkat algoritma SVM dan *Random Forest*. Dalam memprediksi dengan topik polusi untuk menentukan tingkat sentimen positif,negatif,netral dari masyarakat Indonesia. Setelah itu membandingkan akurasi dari algoritma SVM dengan *Random Forest* untuk menentukan algoritma mana yang terbaik pada penelitian ini.

III. Hasil dan Pembahasan

3.1. Pembuatan Model

Penentuan model ada 2 yaitu model training dan model test, parameter yang digunakan test size 0.2, agar model mudah melakukan prediksi. Gambar 3 pembagian model uji dan model training.

```
model uji dan model latih dengan test size 0.2

[(5436,), (1359,), (5436,), (1359,)]
```

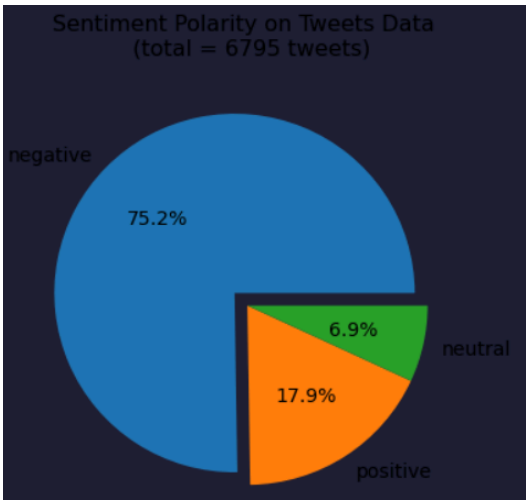
Gambar 3. Model uji dan model training

Gambar diatas merupakan model uji dan model latih dengan pembagian model *training* 80% dan model uji 20%, model *training* sebanyak 5436 baris dan model uji sebanyak 1359 baris.

3.2. Visualisasi

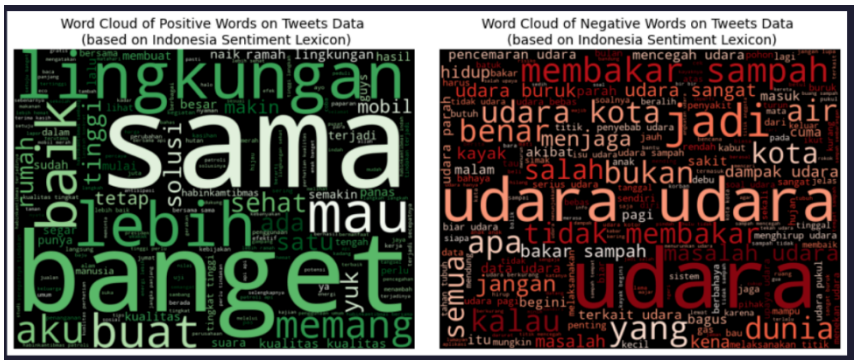
Gambar 4. Merupakan proses dari visualisasi *dataset* penelitian ini berupa *pychart* dan *wordcloud*. proses visualisasi dari *dataset* penelitian ini membahas mengenai polusi udara di Indonesia, visualisasi dilakukan

pertama adalah untuk melihat komparasi sentimen masyarakat Indonesia di *Twitter* akan polusi udara. Visualisasi data dapat dilihat tiga jenis sentiment yaitu positif, netral dan negatif.



Gambar 4. Visualiasi dataset positif dan negatif

Pychart dibawah ini menjabarkan bahwa terdapat 6,9% *tweet* netral; 17,9% *tweet* positif; 75,2% tweet negatif. Didapat kesimpulan bahwa pada *dataset* yang saya *scrapping* terdapat mayoritas *sentiment* negatif dari pada sentiment positif.
visualisasi untuk *wordcloud*. *Worldcloud* adalah teknik visualisasi yang digunakan oleh *library python* untuk menampilkan kata kata yang sering muncul pada *dataset.* Gambar 5 merupakan visualisasi worldcloud positif dan negatif sebagai berikut :



Gambar 5. Visualisasi worldcloud positif dan negatif

3.3. Implementasi Algoritma

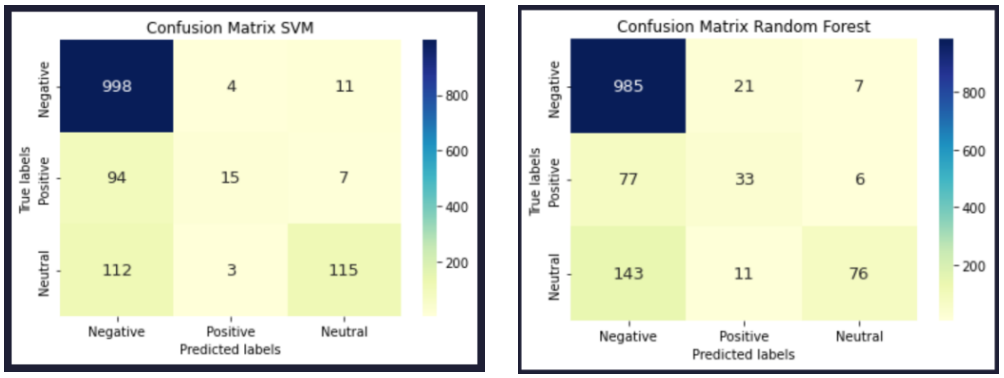
Tahap mengimplementasikan algoritma dari model yang sudah dibuat, untuk mencari akurasi dari algoritma SVM dan *Random Forest*, Import *library* SVM dan *Random Forest* pada teks *editor*, parameter yang dipanggil pada algoritma SVM adalah *kernel linear* yang berarti algoritma SVM akan menggunakan fungsi kernel linear untuk melakukan transformasi pada sebuah data fungsinya adalah menemukan garis pemisah yang paling baik dari kelas data, sedangkan pada algoritma *Random Forest* menggunakan parameter n_*estimator* yang bernilai 100 berarti algoritma ini mengatur pohon keputusan sebanyak 100 kali, semakin banyak pohon keputusan yang dibuat maka semakin kompleks juga penilaian model tersebut lalu ada parameter *max_feature* yang bernilai sqrt merupakan *value* umum yang digunakan pada algoritma *Random Forest* yang mengacu pada akar kuadrat dari total fitur yang tersedia.
Tahap selanjutnya adalah menginisialisasi algoritma maka kedua algoritma setiap proses *fitting* dari model yang sudah dibuat perbandingan *test size* yaitu 80% model *training* dan 20% model *test* menghitung model X yang sudah di ubah menjadi vektorisasi oleh TFIDF dan menargetkan Y yaitu paremeter *polarity*.

```
Implementasi algoritma SVM
Accuracy model: 0.83
Recall model: 0.83
Precision model: 0.82
F1 score model: 0.8
```

```
Implementasi algoritma Random Forest
Accuracy model: 0.81
Recall model: 0.81
Precision model: 0.8
F1 score model: 0.77
```

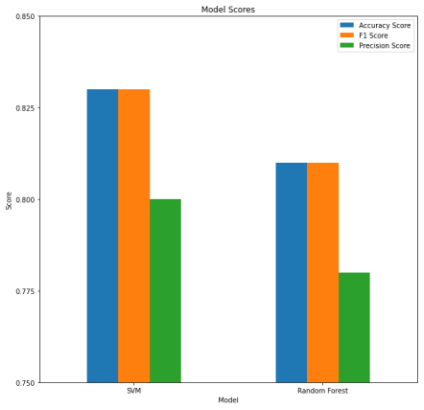
Gambar 6. Hasil kinerja algoritma SVM dan Random Forest

Pada gambar 6 diatas merupakan perhitungan dari akurasi, *recall*, presisi, dan f1 *score* dari algoritma SVM dan *Random Forest* peneliti menghitung kedua algoritma tersebut untuk mengetehaui perbandingan dari keduanya, terdapat akurasi bahwa algoritma SVM mempunyai akurasi sebesar 83% dan Random Forest memiliki akurasi sebesar 81%, lalu selanjutnya adalah menghitung *confusion matrix* dari kedua algoritma tersebut.
Gambar 7 merupakan hasil confusion matrix dari kedua algoritma, pada gambar berikut.



Gambar 7. Hasil confusion matrix SVM dan random Forest

Pada gambar 7 merupakan *confusion matrix* dari algoritma SVM dan Random Forest, yang merupakan performa model dalam memprediksi 3 *value* yaitu netral, positif, negatif pada bagian kiri merupakan label hasil dan bagian bawah merupakan label prediksi.



Gambar 8. Visualisasi bar plot SVM dan Random Forest

Pada gambar 8. diatas, menggunakan visualisasi *bar plot* bahwa algoritma SVM dan *Random Forest*, diketahui algoritma SVM lebih unggul dari segi *perform* jika dibandingkan dengan algoritma *Random Forest*.

	ML Model	Accuracy	f1_score	Recall	Precision
0	SVM	0.83	0.83	0.82	0.80
1	Random Forest	0.81	0.81	0.80	0.78

Gambar 9. Hasil perbandingan algoritma SVM dan Random Forest

Gambar 9 adalah angka tepatnya dari penghitungan performa diketahui bahwa algoritma SVM mendapatkan akurasi sebesar 83% sedangkan *Random Forest* 81% berbeda 2%, dan untuk f1 *score*, *recall*, dan *precision* algoritma SVM lebih unggul untuk uji coba sentimen dibandingkan algoritma *Random Forest.*

IV. Kesimpulan

Hasil penelitian yang berdasar kepada hasil implementasi dan uji analisis hasil sebelumnya, *dataset* yang didapatkan dari *twitter* dengan topik polusi udara di Indonesia memiliki sentiment negatif sebesar 75,2% ,lalu diimplementasi oleh algoritma SVM dan *Random Forest*, hasil menunjukan bahwa algoritma SVM memiliki hasil yang lebih baik dibandingkan algoritma *Random Forest*, dengan akurasi dan presisi sebesar 83% dan 80% untuk algoritma SVM, sedangkan algoritma *Random Forest* sebesar 81% dan 78%. Menunjukkan bahwa algoritma SVM lebih baik untuk diimplementasikan pada analisis *sentiment* dibandingkan algoritma *Random Forest* untuk topik polusi udara. Berdasarkan hasil visualisasi sebanyak 5545 dataset dengan nilai positif 17,9%, negatif 75,2% dan netral 6,9%.

Berdasarkan analisa dan implementasi saran dari peneliti untuk penelitian selanjutnya yaitu untuk mencari jumlah *dataset* yang lebih banyak dan menggunakan metode *upsampling* atau *downsampling* untuk mengatasi *data imbalance*, serta dapat menggunakan algoritma lain seperti XGBOOST untuk membantu memecahkan masalah

V. Daftar Pustaka

[1] BBC Indonesia, “Indonesia masuk ‘enam negara paling berkontribusi terhadap polusi udara global’, warga akan gugat pemerintah dan industri,” https://www.bbc.com/indonesia/articles/c72enp76622o.

[2] CNBC Indonesia, “Bukan PLTU, Ternyata Ini Penyebab Utama Polusi di Jakarta,” https://www.cnbcindonesia.com/news/20230903164954-4-468635/bukan-pltu-ternyata-ini-penyebab-utama-polusi-di-jakarta#:~:text=Siti%20menjelaskan%20penyebab%20pencemaran%20kualitas,juta%20di%20antaranya%20sepeda%20motor.

[3] CNBC Indonesia, “Polusi Makin Parah! Rakyat Menderita, Solusi Malah Amburadul,” https://www.cnbcindonesia.com/research/20230828123040-128-466679/polusi-makin-parah-rakyat-menderita-solusi-malah-amburadul.

[4] N. Hendrastuty, A. Rahman Isnain, and A. Yanti Rahmadhani, “Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine,” vol. 6, no. 3, 2021, [Online]. Available: http://situs.com

[5] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, “ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI,” *Jurnal Teknoinfo*, vol. 14, no. 2, p. 115, Jul. 2020, doi: 10.33365/jti.v14i2.679.

[6] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” 2017. [Online]. Available: https://t.co/jrvaMsgBdH

[7] A. Fathan Hidayatullah, A. Sn, J. Teknik, I. Fakultas, and T. Industri, “ISSN: 1979-2328 UPN "Veteran,” 2014. [Online]. Available: http://www.situs.com

[8] M. Rangga, A. Nasution, and M. Hayaty, “Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter,” *JURNAL INFORMATIKA*, vol. 6, no. 2, pp. 212–218, 2019, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ji

[9] M. I. Fikri, T. S. Sabrila, Y. Azhar, and U. M. Malang, “Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter”.

[10] A. Novantirani, M. S. Kania Sabariah, and V. Effendy, “Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine.”

[11] R. Tineges, A. Triayudi, and I. D. Sholihati, “Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM),” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 3, p. 650, Jul. 2020, doi: 10.30865/mib.v4i3.2181.

[12] D. Darwis, E. Shintya Pratiwi, A. Ferico, and O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," 2020.
[13] N. Dwi Putranti and E. Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine," *IJCCS*, vol. 8, no. 1, pp. 91–100, 2014.
[14] Y. Tao, F. Zhang, C. Shi, and Y. Chen, "Social media data-based sentiment analysis of tourists' air quality perceptions," *Sustainability (Switzerland)*, vol. 11, no. 18, Sep. 2019, doi: 10.3390/su11185070.
[15] M. R. Huq, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," 2017. [Online]. Available: www.ijacsa.thesai.org
[16] AWC, "What is Sentiment Analysis?," https://aws.amazon.com/what-is/sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20the%20process,social%20media%20comments%2C%20and%20reviews.
[17] Rohith Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.
[18] Sruthi E R and Analytics Vidyha, "Understand Random Forest Algorithms With Examples (Updated 2023)," https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.
[19] Tokenex, "What is Tokenization?," Tokenex. Accessed: Oct. 24, 2023. [Online]. Available: https://www.tokenex.com/blog/what-is-tokenization/
[20]Alexander S Gillis, "Definition Lemmatization," 2023. Accessed: Oct. 24, 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/lemmatization