KOMPARASI KINERJA ALGORITMA SVM DAN RF DALAM KLASIFIKASI SENTIMEN DENGAN DETEKSI SARKASME PADA KOMENTAR YOUTUBE

Muhammad Hilman Habib Habibi, Eka Dyar Wahyuni, Reisa Permatasari
Sistem Informasi, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur, Indonesia
hilmanhabibi07@gmail.com

ABSTRAK

Youtube merupakan salah satu platform berbagi video yang paling sering diakses di Indonesia, terutama dengan maraknya diskusi mengenai PILKADA 2024. Beragam isu yang muncul menjelang PILKADA 2024 ini memicu pro dan kontra, mendorong masyarakat untuk memberikan tanggapan melalui media sosial. Klasifikasi sentimen bertujuan mengelompokkan opini menjadi positif atau negatif, namun sering menghadapi hambatan akibat keberadaan sarkasme, yaitu bentuk ironi yang menyampaikan makna bertentangan dengan pernyataan eksplisit. Dalam penelitian ini, data diperoleh melalui crawling komentar Youtube. Data tersebut kemudian diproses melalui tahapan cleaning, case folding, dan stemming. Klasifikasi sentimen dengan deteksi sarkame ini akan menggunakan algoritma Support Vector Machine (SVM) dan Random Forest (RF) dengan berbagai skenario, termasuk pembagian data menggunakan metode holdout dengan rasio 80:20 dan 70:30, teknik resampling Random Oversampling (ROS) dan Random Undersampling (RUS), serta pembobotan kata menggunakan TF-IDF dan TF-ABS. Berdasarkan hasil evaluasi, algoritma SVM dengan teknik ROS dan pembobotan kata TF-IDF memberikan hasil terbaik untuk klasifikasi sentimen dengan nilai 0.80, sedangkan algoritma SVM dengan TF-IDF tanpa resampling memberikan hasil terbaik untuk deteksi sarkasme dengan nilai 0.73. Hasil ini menunjukkan keandalan SVM dalam menangkap pola data yang kompleks, terutama dalam klasifikasi sentimen dengan deteksi sarkasme.

Kata kunci : *PILKADA, Sentimen, Sarkasme, Support Vector Machine (SVM), Random Forest (RF)*

1. PENDAHULUAN

Teknologi Informasi dan Komunikasi saat ini telah banyak berkembang seiring berjalannya waktu. Semakin tinggi teknologi informasi dan komunikasi yang ditemukan dan digunakan maka tidak terlepas dari bagaimana perkembangan komunikasi di masyarakat dan bagaimana masyarakat sosial tersebut berinteraksi [1].

Hal tersebut membuat hampir semua orang memiliki perangkat komunikasi yang memungkinkan untuk berkomunikasi dengan semua orang diseluruh dunia melalui media sosial. Menurut data dari WeAreAsocial, Saat ini pengguna media sosial di dunia mencapai lebih dari 5 miliar jiwa dari total 8,08 miliar jiwa populasi di Dunia. Indonesia sendiri memiliki 139 juta pengguna media sosial pada Januari 2024, setara dengan 49,9 persen dari total populasi [2]

Media sosial ada banyak jenisnya, salah satunya adalah media social yang berbasis video dan youtube merupakan salah satu media sosial yang berbasis video yang paling sering dikunjungi di Indonesia. Menurut studi [3].

YouTube menyumbang 20% dari lalu lintas Web dan 10% dari total lalu lintas Internet. Salah satu topik yang sedang ramai diperbincangkan pada youtube yakni topik mengenai Pilkada (Pemilihan Kepala Daerah) 2024. Banyak sekali isu-isu yang muncul menjelang Pilkada 2024 sehingga memunculkan pro kontra dan menyebabkan masyarakat Indonesia beramai-ramai mengawal pilkada ini dengan memberikan tanggapan serta komentar. Komentar-komentar tersebut dapat memberikan gambaran mengenai sentimen atau perasaan masyarakat mengenai pilkada di Indonesia.

Klasifikasi sentimen, yang disebut juga dengan opinion mining, merupakan salah satu cabang ilmu dari data mining yang bertujuan untuk menganalisis, memahami, mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas seperti produk, servis, organisasi, individu, dan topik tertentu[4].

Menurut Rozi dkk, klasifikasi sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung beropini negatif atau positif[5].

Tugas dasar dalam klasifikasi sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, apakah pendapat yang dikemukakan dalam bersifat positif atau negatif. Namun seringkali sentimen pada sebuah komentar tidak dapat dilakukan dengan benar ketika komentar tersebut mengandung sarkasme.

Sarkasme merupakan suatu bentuk ironi khusus yang terjadi ketika seseorang menyampaikan informasi yang tersirat, biasanya memiliki makna yang berlawanan dengan apa yang diucapkan [6].

Opini sarkasme ini sering sekali disampaikan seseorang secara tidak langsung melalui media sosial sehingga merupakan salah satu masalah yang mempengaruhi hasil analisis sentimen karena salah diklasifikasikan.

Beberapa penelitihan mengenai deteksi sarkasme dalam pengklasfikasian sentimen pernah dilakukan. Salah satunya ada penelitian yang ditulis oleh Aisyah.

Penelitian ini menggunakan algortima Naïve Bayes dan Random Forest untuk analisis sentimen lalu Random Forest untuk deteksi sarkasme. Hasil menunjukan bahwa akurasi terbaik pada model analisis sentimen diperoleh dengan menggunakan Random Forest dengan akurasi 71% dan akurasi model sarkasme dengan Random Forest adalah 83%[7].

Selain itu berdasarkan literatur review yang ditulis oleh Samer Muthana Sarsam. Literatur review ini membandingkan 31 penelitian mengenai deteksi sarkasme menggunakan algoritma machine learning dan hasil menunjukan bahwa Support Vector Machine (SVM) merupakan algoritma terbaik dan paling umum digunakan untuk mendeteksi sarkasme di Twitter[8].

Berdasarkan kedua penelitian tersebut, maka dilakukan penelitian untuk mengetahui hasil komparasi kinerja algoritma SVM dan RF dalam mengklasifikasikan sentimen sebelum dan sesudah deteksi sarkasme pada komentar video youtube terkait topik Pilkada 2024.

.

2. TINJAUAN PUSTAKA

Pada tahap ini, akan disajikan berbagai literatur ilmiah yang berkaitan dengan penelitian ini.

2.1. Penelitian Terdahulu

Penelitian yang dilakukan oleh Aisyah Muhaddisi, Bambang Nurcahyo Prastowo, dan Diyah Utami Kusumaning Putri bertujuan untuk membandingkan metode *Naïve Bayes* dan *Random Forest* dalam analisis sentimen dengan deteksi sarkasme pada komentar Instagram politisi Indonesia. Data dikumpulkan menggunakan Selenium dari akun seperti Puan Maharani, Joko Widodo, dan DPR RI. Hasil menunjukkan bahwa *Random Forest* lebih unggul dibandingkan *Naïve Bayes*. Pada analisis sentimen tanpa deteksi sarkasme, *Naïve Bayes* mencapai akurasi 61%, sedangkan *Random Forest* 72%. Dengan deteksi sarkasme, akurasi *Naïve Bayes* sedikit menurun menjadi 60%, sementara *Random Forest* mencapai 71%. Temuan ini menunjukkan bahwa deteksi sarkasme memengaruhi analisis sentimen dan *Random Forest* lebih efektif dalam klasifikasi sentimen.[7].

Penelitian yang dilakukan oleh Yessi Yunitasari, Aina Musdholifah, dan Anny Kartika Sari bertujuan untuk menganalisis sentimen pada ulasan film dengan membandingkan performa algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) guna menentukan model dengan akurasi terbaik. Penelitian ini menggunakan metode eksperimen yang terdiri dari tiga tahapan utama, yaitu pengumpulan data, pengolahan data awal, serta eksperimen dan pengujian model. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki akurasi lebih tinggi dibandingkan *Naïve Bayes*. *Naïve Bayes* mencapai akurasi sebesar 84,50%, sedangkan *SVM* memperoleh akurasi sebesar 90,00%. Dengan demikian, *SVM* terbukti lebih unggul dalam mengklasifikasikan sentimen pada ulasan film dibandingkan *Naïve Bayes[9]*.

Penelitian yang dilakukan oleh Samer Muthana Sarsam, Hosam Al-Sammaraie, Ahmed Ibrahim Alzahrani, dan Bianca Wright bertujuan untuk mengulas berbagai algoritma machine learning yang digunakan dalam deteksi sarkasme di Twitter berdasarkan penelitian sebelumnya. Dari 31 penelitian yang dianalisis, studi ini mengelompokkan algoritma ke dalam dua kategori, yaitu Automatic Machine Learning Approaches (AMLA) dan Customized Machine Learning Approaches (CMLA). Hasil kajian menunjukkan bahwa algoritma Support Vector Machine (SVM) adalah metode yang paling banyak digunakan dalam kelompok AMLA untuk mendeteksi sarkasme. Selain itu, kombinasi Convolutional Neural Network (CNN) dan SVM terbukti memiliki kinerja prediksi yang tinggi. Namun, pendekatan CMLA menunjukkan variasi dalam hasil kinerja karena perbedaan dalam fitur pemrosesan teks dan teknik klasifikasinya. Secara keseluruhan, algoritma SVM dan kombinasi CNN-SVM merupakan metode yang paling efisien untuk memprediksi sarkasme di Twitter[8].

2.2. Youtube

Youtube merupakan salah satu dari social Network yang berkategori Multimedia sharing [10].

Pada umumnya, file-file Multimedia yang dibagikan berupa file gambar, audio, dan video melalui internet. Fitur-fitur yang disediakan multimediasharing antara lain mengirim dan menanggapi komentar, pemberian rate (peringkat), berbagi (share) dalam bentuk alamat (URL). Proses yang dilakukan meliputi unggah (upload) dan unduh (download). Di Youtube, para pengguna dapat melakukan interaksi dengan mengetahui jumlah orang yang menonton video yang dibagikan, memberikan dukungan melalui tombol like dan memberikan komentar pada video tersebut.

2.3. Text Mining

Text mining merupakan proses untuk mengekstrak informasi yang bermanfaat dari teks yang tidak terstruktur. Ini meliputi berbagai teknik, seperti 8 pengambilan informasi dari sumber dokumen atau situs web, mengelompokkan teks dan mengklasifikasikannya, hingga menjadi sebuah entity, relation, dan event extraction [11].

Text Mining sering digunakan untuk mengkategorikan, dan untuk menganalisis sentimen yang berguna mengidentifikasi wawasan dan pola trend dalam volume besar dari suatu data. Beberapa perusahaan menggunakannya untuk menganalisis pendapat, emosi dan sentimen dari masyarakat terhadap merek dan produk mereka [12]

2.4. Pilkada

Penyelenggaraan Pemilihan Kepala Daerah (Pilkada) sejatinya merupakan bagian penting kehidupan bernegara Indonesia di era Reformasi. Pemilihan Kepala Daerah (Pilkada) merupakan proses

demokrasi di Indonesia yang memberikan kesempatan kepada masyarakat untuk memilih pemimpin daerah secara langsung, termasuk gubernur, bupati, dan wali kota [13].

Meskipun Pilkada membawa berbagai manfaat bagi demokrasi lokal, proses ini juga dihadapkan pada tantangan, seperti politik uang, hoaks, dan rendahnya partisipasi pemilih di beberapa daerah. Selain itu, berbagai bentuk kritik atau sentimen publik terhadap calon kepala daerah sering kali muncul di platform digital seperti YouTube dan media sosial lainnya, di mana masyarakat menyuarakan pendapat mereka tentang calon atau kebijakan politik. Salah satu fenomena yang sering ditemui di media sosial adalah penggunaan sarkasme dalam komentar, terutama terkait dengan Pilkada. YouTube, sebagai platform yang banyak digunakan untuk mengunggah konten kampanye dan debat politik, sering kali menjadi tempat di mana masyarakat mengekspresikan kekecewaan atau ketidakpuasan mereka terhadap calon kepala daerah melalui komentar bernada sarkastik.

2.5. Klasifikasi Sentimen

Informasi tekstual secara umum dapat dibagi menjadi informasi fakta dan opini [14].

Klasifikasi sentimen, yang disebut juga dengan opinion mining, merupakan salah satu cabang ilmu dari data mining yang bertujuan untuk mengklasifikasikan, memahami, mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas seperti produk, servis, organisasi, individu, dan topik tertentu [4].

Klasifikasi sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif [5].

Tugas dasar dalam klasifikasi sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, apakah pendapat yang dikemukakan dalam bersifat positif atau negatif.

2.6. Deteksi Sarkasme

Sarkasme merupakan suatu bentuk ironi khusus yang terjadi ketika seseorang menyampaikan informasi yang tersirat, biasanya memiliki makna yang berlawanan dengan apa yang diucapkan [6].

Kalimat sindiran atau sarkasme masih sering digunakan oleh kalangan publik untuk mengungkapkan maksud isi hati dan pikiran baik itu yang disampaikan secara langsung maupun tidak langsung [15].

Salah satu jenis sarkasme yang digunakan adalah jenis sarkasme yang berlainan makna yaitu kalimat sarkasme berbentuk kalimat positif yang memiliki makna negatif. Opini sarkasme ini sering sekali disampaikan seseorang secara tidak langsung melalui media sosial sehingga merupakan salah satu masalah yang mempengaruhi hasil analisis sentimen karena salah diklasifikasikan [14].

2.7. *Preprocessing Text*

Preprocessing Text mining adalah salah satu bagian pada algoritma *Text Mining*, proses mengolah teks yang difungsukan untuk pengubahan bentuk dokumen menjadi data yang terstruktur sesuai dengan kepentingannya supaya dapat diolah lebih lanjut pada proses text mining [16].

Beberapa tahap *pre-processing text* yaitu *case folding, cleaning, tokenizing, stopword removal*, dan *stemming.*

2.8. *Support Vector Machine*

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi [17].

Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada problem non-linear [18], SVM telah dikembangkan agar dapat bekerja pada problem *non-linier* dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak (*margin*) antara kelas data [16].

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah class pada input space [17].

Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan pattern terdekat dari masing-masing *class* [17].

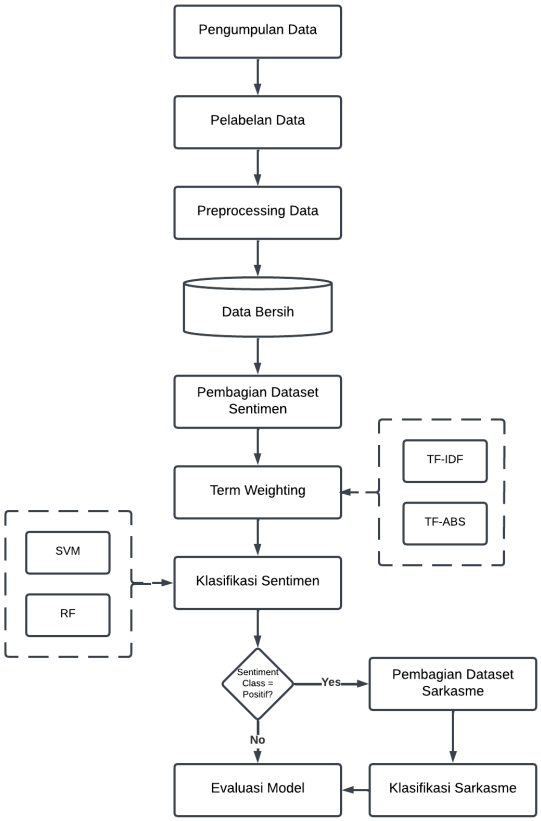
2.9. *Random Forest*

Random Forest adalah sebuah algoritma yang menggunakan teknik *ensemble learning* dengan menggabungkan beberapa pohon keputusan yang dibuat secara acak untuk meningkatkan akurasi klasifikasi pada data yang kompleks. Algoritma ini berupa kombinasi dari beberapa *tree predictors* atau bisa disebut *decision trees* dimana setiap *tree* bergantung pada nilai *random vector* yang dijadikan sampel secara bebas dan merata pada semua *tree* dalam *forest* tersebut [19].

Hasil prediksi dari *Random Forest* didapatkan melalui hasil terbanyak dari setiap *individual decision tree* (voting untuk klasifikasi dan rata-rata untuk regresi).

3. METODE PENELITIAN

Alur penelitian yang akan dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian

3.1. Pengumpulan Data

Data yang dikumpulkan berasal dari komentar video Youtube yang berhubungan dengan Pilkada (Pemilihan Kepala Daerah) 2024. Alasan diambilnya data dari komentar video Youtube, sebab banyaknya ketersediaan data di Youtube. yang berasal dari beberapa channel youtube yang terpercaya di antaranya, Kompas Tv, Metro Tv, Liputan 6, iNews, TvOne, Kumparan. Selanjutnya data diambil melalui proses crawling menggunakan algoritma yang dikembangkan peneliti dalam bahasa pemrograman Python dan disimpan dalam bentuk csv.

3.2. Pelabelan Data

Pelabelan data adalah proses memberikan kelas atau label pada data yang terkumpul, sehingga data tersebut dapat digunakan untuk analisis dan pembelajaran model. Untuk memberikan label pada data menggunakan *library* pada *python* yaitu TextBlob dan VADER serta library transformers untuk memanggil model dari *Huggingface*. Setelah dilakukan pelabelan menggunakan ketiga *library* tersebut dan didapatkan hasil pelabelan yang berbeda. Maka akan dilakukan pengambilan modus nilai dari ketiga library tersebut untuk menentukan label sentimen yang akan digunakan. Adapun bantuan dari *library* AI untuk melakukan pelabelan sarkasme dengan menggunakan model GPT 3.5 Turbo. Hasil dari tahap ini akan digunakan untuk tahap selanjutnya.

3.3. Preprocessing Data

Data mentah seringkali tidak lengkap, tidak konsisten, dan mungkin mengandung banyak kesalahan. *Pre-processing* merupakan teknik data mining yang melibatkan perubahan data mentah menjadi sebuah format yang terstruktur dan dimengerti. Tahap data preprocessing berfungsi untuk menghilangkan noise pada data, karena sebelum diklasifikasikan data harus benar-benar bersih dan terstruktur. Pada tahap ini terdapat beberapa aktivitas yang akan dilakukan, yakni sebagai berikut:

3.4. C*leaning*

Tahap ini dilakukan cleaning atau dibersihkan dari kata-kata atau karakter yang tidak penting. Karena data dari berbagai macam karakter pengguna sehingga tidak semua pengguna menggunakan karakter kata yang penting. Proses cleaning atau pembersihan bertujuan menghilangkan tanda baca, menghilangkan angka, menghilangkan URL, menghilangkan nama akun atau mention, menghilangkan tagar atau hashtag, menghilangkan huruf tunggal.

3.5. *Case Folding*

Tahap ini dilakukan karena format huruf dalam dokumen teks yang dikumpulkan tidak selalu sama, ada yang huruf besar dan ada yang huruf kecil. Oleh karena itu, semua kata dalam data diubah menjadi huruf kecil menggunakan fungsi lower().

3.6. *Stemming*

Langkah terakhir yang dilakukan adalah tahap *Stemming*. *Stemming* mengubah kata-kata yang memiliki imbuhan, baik awalan, sisipan, akhiran, atau gabungan dari ketiganya (imbuhan) menjadi kata dasarnya. Dalam penelitian ini, kita menggunakan stemmer dari Sastrawi untuk memproses kata-kata dalam bahasa Indonesia.

3.7. Pembagian Dataset Sentimen

Pada tahap ini, proses pembagian dataset sarkasme dilakukan untuk keperluan training dan testing model klasifikasi. Pembagian dataset ini menggunakan metode hold-out, yaitu dengan memisahkan data menjadi dua subset, di mana sebagian besar data digunakan untuk melatih model, sementara sisanya digunakan untuk mengevaluasi kinerja model yang telah dilatih. Dalam implementasi ini, komposisi pembagian dataset dilakukan dalam dua skenario, yaitu 80:20 dan 70:30 dengan komposisi dataset training lebih besar.

3.8. Pembobotan Data dengan Term Weighting

Pada Tahap ini, akan dilakukan teknik pembobotan dengan TF-IDF dan TF-ABS. TF-IDF mengukur pentingnya sebuah kata dalam sebuah dokumen dengan mempertimbangkan frekuensi kemunculan kata tersebut dalam dokumen (TF) dan kebalikannya di seluruh koleksi dokumen (IDF). Dalam penelitian ini, dilakukan pemrograman dengan

Python menggunakan library scikit-learn dan gensim. Sedangkan TF-ABS menggunakan frekuensi kemunculan kata dalam dokumen (TF) tanpa mempertimbangkan kebalikannya seperti pada IDF.

3.9. Klasifikasi Sentimen

Pada tahap ini, dilakukan klasifikasi sentimen pada data train menggunakan model SVM dan Random Forest (RF). Data yang digunakan telah melalui proses pelabelan sentimen dan pembobotan, dengan tujuan mengoptimalkan kinerja model agar mampu mengelompokkan data secara akurat sesuai dengan karakteristik yang telah ditentukan.

3.10. Pembagian Dataset Sarkasme

Pada tahap ini, proses pembagian dataset sarkasme dilakukan untuk keperluan training dan testing model klasifikasi. Pembagian dataset ini menggunakan metode hold-out, yaitu dengan memisahkan data menjadi dua subset, di mana sebagian besar data digunakan untuk melatih model, sementara sisanya digunakan untuk mengevaluasi kinerja model yang telah dilatih. Dalam implementasi ini, komposisi pembagian dataset dilakukan dalam dua skenario, yaitu 80:20 dan 70:30 dengan komposisi dataset training lebih besar.

3.11. Klasifikasi Sarkasme

Pada tahap ini, dilakukan klasifikasi sarkasme pada data train menggunakan model SVM dan Random Forest (RF). Data yang digunakan telah melalui proses pelabelan sarkasme dan pembobotan, dengan tujuan mengoptimalkan kinerja model agar mampu mengelompokkan data secara akurat sesuai dengan karakteristik yang telah ditentukan.

3.12. Evaluasi Model

Evaluasi model merupakan tahap pengujian data uji (*testing*) yang dilakukan dalam penelitian ini untuk mengukur kinerja model dalam melakukan klasifikasi sentimen dan klasifikasi sarkasme. Pada tahap ini, hasil pengujian data uji akan ditampilkan dalam bentuk bobot atau nilai tertentu yang merepresentasikan kinerja model terhadap atribut yang telah dipersiapkan sebelumnya. Proses evaluasi ini juga melibatkan penggunaan *Confusion Matrix* untuk menghitung metrik evaluasi, seperti akurasi, *precision*, *recall*, dan *F1-score*, yang menjadi indikator utama dalam menilai efektivitas model.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Pengumpulan data ini dilakukan melalui proses *crawling* komentar YouTube dengan memanfaatkan *video ID* dari konten yang membahas Pilkada 2024. Data yang dikumpulkan mencakup nama akun sebagai identitas pemberi komentar dan isi komentar dari akun tersebut.

	authorName	text
0	@radenmasplenggong1731	Semoga bakal ndlosor semua
1	@XhiaLee-y6u	Ngaur ngomong ga jelas nenek wati
2	@maifat8416	Guru yg hbt..bu mega..❤❤
3	@adiarmanidris9929	Iihh kok si nenek lampir gak malu ya, sekolah ...
4	@RofifRofif-jw6kt	I love you Bu Mega semangat terus
...	...	...

Gambar 2. Hasil Pengumpulan Data

Hasil pengumpulan data pada Gambar 2 merupakan data yang diperoleh berdasarkan *video ID*. Untuk mendapatkan dataset yang relevan, proses ini melibatkan pengumpulan *video ID* dari konten yang membahas topik pilkada 2024. Sebanyak 4,256 data komentar berhasil dikumpulkan dan siap digunakan dalam penelitian ini.

4.2. Pelabelan Data

Setelah itu, data yang sudah dikumpulkan akan diterjemahkan ke dalam bahasa Inggris guna memastikan akurasi oleh model yang digunakan. Hasil dari terjemahaannya akan disimpan kedalam kolom baru bernama translated_text. Setelah data diterjemahkan, dilakukan proses pelabelan dengan memanfaatkan beberapa library seperti library vadersentiment untuk pelabelan Vader, library textblob untuk pelabelan TextBlob, library AutoTokenizer dan AutoModelForSequence Classification untuk pelabelan cardiffnlp, serta library openai untuk pelabelan sarkasme GPT 3.5 Turbo.

4.3. *Preprocessing* Data

Setelah data dilabeli, tahap selanjutnya adalah preprocessing data. Proses preprocessing bertujuan untuk membersihkan data dari noise serta memastikan data dalam kondisi yang bersih dan terstruktur sebelum dilakukan pemodelan. Langkah ini sangat penting untuk meningkatkan kualitas data sehingga hasil analisis atau model yang dihasilkan dapat lebih akurat dan andal.

4.3.1. *Cleaning*

Tahap ini dilakukan untuk memastikan data bersih dari elemen yang menganggu seperti tag HTML, *emoticon*, tanda baca, dan karakter lain yang tidak diperlukan. Hasil *cleaning* dapat dilihat pada Tabel 1.

Tabel 1. Hasil *Cleaning* Data

Komentar	*Cleaning*
Nggak nyangka janji-janji sehebat ini, bisa bikin kita mimpi indah nih! ✨	Nggak nyangka janji-janji sehebat ini bisa bikin kita mimpi indah nih

4.3.2. *Case Folding*

Tahap ini dilakukan untuk standarisasi format huruf dalam seluruh komentar, menghilangkan perbedaan besar-kecil huruf pada kata yang sama. Hal tersebut dilakukan karena tidak semua dokumen teks konsisten menggunakan huruf kapital atau huruf kecil. Hasil *case folding* dapat dilihat pada *Tabel 2.*

Tabel 2. Hasil *Case Folding* Data

Komentar	*Case folding*
BOCAH KOSONG, istilah GEN Z	bocah kosong, istilah gen z

4.3.3. *Stemming*

Tahap ini dilakukan untuk mengubah kata-kata yang memiliki imbuhan (prefiks, sufiks, infiks, atau gabungan) menjadi bentuk dasarnya (kata dasar). Hasil *stemming* dapat diluhat pada Tabel 3.

Tabel 3. Hasil *Stemming* Data

Komentar	*Stemming*
Sedang mempersiapkan koruptor untuk ditempatkan di berbagai wilayah sangat menjijikkan	sedang siap koruptor untuk tempat bagai wilayah sangat jijik

4.4. Pembagian Dataset Sentimen

Pada tahap ini, dataset yang telah melewati tahap *preprocessing* akan dilkukan pembagian dataset untuk klasifikasi sentimen. Pembagian dataset dilakukan untuk *training* dan *testing* model dengan rasio 80:20 dan 70:30, serta *resampling* data sesuai dengan rasio tersebut.

4.5. Proses *Term Weighting*

Setelah melakukan pembagian dataset menjadi data latih dan data uji, langkah selanjutnya adalah melakukan pembobotan data menggunakan teknik term weighting. Dalam penelitian ini, digunakan dua jenis *term weighting*, yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Term Frequency Absolute* (TF-ABS).

4.6. Klasifikasi Sentimen

Pada tahap ini, model akan dibangun menggunakan data *training* dengan algoritma SVM dan *Random Forest* (RF), penerapan teknik resampling, serta pembagian dataset dengan rasio 80:20 dan 70:30. Terdapat 24 skenario yang dilakukan pada tahap klasifikasi sentimen yang melibatkan resampling data yaitu random oversampling (ROS) dan Random Undersampling (RUS).

4.7. Pembagian Dataset Sarkasme

Setelah dilakukan klasifikasi sentimen, kemudia akan dilkukan pembagian dataset untuk klasifikasi sarkasme dengan menggunakan data sentimen yang berlabel positif. Hasil dari proses ini disimpan dalam dataframe baru, yang hanya berisi data dengan nilai sentimen 'Positif' untuk dilakukan *training* dan *testing* model sarkasme dengan pembagian dataset 80:20 dan 70:30, serta *resampling* data sesuai dengan rasio tersebut.

4.8. Klasifikasi Sarkasme

Pada tahap ini, model juga akan dibangun menggunakan data *training* dengan algoritma SVM dan *Random Forest* (RF), penerapan teknik resampling, serta pembagian dataset dengan rasio 80:20 dan 70:30. Terdapat 24 skenario yang dilakukan pada tahap klasifikasi sentimen yang melibatkan resampling data yaitu random oversampling (ROS) dan Random Undersampling (RUS).

4.9. Evaluasi Model

Setelah proses klasifikasi dilakukan, tahap ini dilakukan utnuk mengevaluasi model dengan data *testing*. Tahap ini bertujuan mengetahui kinerja model yang dibangun. Dan hasil dari evaluasi model ini dapat dilihat pada Tabel 4.

Tabel 4. Hasil Evaluasi Model

SVM			Akurasi (Train)	Akurasi (Test)	Overfit / Underfit / Good	Precision	Recall	F1-Score
SENTIMEN TF-IDF	80:20	Normal	0.87	0.80	Good	0.79	0.79	0.79
		ROS	**0.87**	**0.80**	**Good**	**0.80**	**0.81**	**0.80**
		RUS	0.86	0.79	Good	0.79	0.80	0.79
	70:30	Normal	0.87	0.80	Good	0.80	0.79	0.79
		ROS	0.87	0.78	Good	0.78	0.79	0.78
		RUS	0.86	0.79	Good	0.78	0.79	0.78
SENTIMEN TF-ABS	80:20	Normal	0.90	0.79	Good	0.79	0.79	0.79
		ROS	0.91	0.80	Good	0.79	0.80	0.79
		RUS	0.91	0.79	Good	0.78	0.79	0.78
	70:30	Normal	0.91	0.80	Good	0.79	0.78	0.79
		ROS	0.91	0.78	Good	0.78	0.78	0.78
		RUS	0.92	0.78	Good	0.77	0.78	0.77
SARKASME TF-IDF\	80:20	**Normal**	**0.80**	**0.73**	**Good**	**0.73**	**0.66**	**0.67**
		ROS	0.85	0.70	Good	0.68	0.69	0.68
		RUS	0.84	0.67	Overfit	0.66	0.68	0.66
	70:30	Normal	0.81	0.72	Good	0.71	0.64	0.64
		ROS	0.84	0.70	Good	0.67	0.68	0.67
		RUS	0.85	0.66	Overfit	0.64	0.65	0.64
SARKASME TF-ABS	80:20	Normal	0.88	0.67	Overfit	0.64	0.63	0.64
		ROS	0.91	0.66	Overfit	0.64	0.65	0.64
		RUS	0.91	0.61	Overfit	0.62	0.63	0.61

SVM			Akurasi (Train)	Akurasi (Test)	Overfit / Underfit / Good	Precision	Recall	F1-Score
	70:30	Normal	0.89	0.69	Overfit	0.65	0.64	0.64
		ROS	0.92	0.67	Overfit	0.65	0.66	0.65
		RUS	0.91	0.61	Overfit	0.61	0.62	0.60
RF			**Akurasi (Train)**	**Akurasi (Test)**	**Overfit / Underfit / Good**	**Precision**	**Recall**	**F1-Score**
SENTIMEN TF-IDF	80:20	Normal	0.72	0.67	Good	0.75	0.61	0.58
		ROS	0.81	0.73	Good	0.74	0.74	0.73
		RUS	0.79	0.71	Good	0.73	0.73	0.71
	70:30	Normal	0.72	0.65	Good	0.73	0.59	0.55
		ROS	0.82	0.72	Good	0.73	0.74	0.72
		RUS	0.82	0.71	Good	0.73	0.71	0.71
SENTIMEN TF-ABS	80:20	Normal	0.72	0.66	Good	0.72	0.60	0.57
		ROS	0.79	0.72	Good	0.74	0.74	0.72
		RUS	0.78	0.71	Good	0.74	0.74	0.71
	70:30	Normal	0.71	0.65	Good	0.74	0.59	0.55
		ROS	0.80	0.73	Good	0.75	0.75	0.73
		RUS	0.78	0.73	Good	0.75	0.75	0.73
SARKASME TF-IDF	80:20	Normal	0.71	0.68	Good	0.79	0.56	0.50
		ROS	0.82	0.68	Good	0.65	0.65	0.65
		RUS	0.83	0.66	Overfit	0.65	0.65	0.65
	70:30	Normal	0.73	0.68	Good	0.72	0.56	0.52
		ROS	0.83	0.69	Good	0.66	0.66	0.66
		RUS	0.84	0.67	Overfit	0.65	0.65	0.65
SARKASME TF-ABS	80:20	Normal	0.72	0.68	Good	0.77	0.56	0.51
		ROS	0.78	0.67	Good	0.64	0.62	0.62
		RUS	0.79	0.66	Good	0.63	0.63	0.63
	70:30	Normal	0.72	0.68	Good	0.72	0.56	0.51
		ROS	0.76	0.67	Good	0.64	0.63	0.63
		RUS	0.81	0.67	Good	0.64	0.64	0.64

Hasil evaluasi model pada table evaluasi model skenario terbaik menunjukkan bahwa kombinasi tersebut model svm memberikan akurasi yang baik dalam klasifikasi sentiment dan sarkasme, dimana klasifikasi sentimen mencapai akurasi 0.87 pada data train dan 0.80 pada data test serta klasifikasi sarkasme mencapai akurasi 0.80 pada data train dan 0.73 pada data test, tanpa ada masalah overfit dan undefit, yang berarti model bekerja secara optimal dan konsisten.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, penelitian ini mengembangkan 48 skenario berbasis algoritma Support Vector Machine (SVM) dan Random Forest (RF) dengan pembagian dataset Holdout (80:20 dan 70:30), teknik pembobotan kata berbasis Term Weighting (TF-IDF dan TF-ABS), serta teknik resampling Random Oversampling (ROS) dan Random Undersampling (RUS). Hasil implementasi menunjukkan bahwa algoritma SVM memiliki kinerja lebih unggul dibandingkan RF dalam klasifikasi sentimen dengan deteksi sarkasme pada komentar video YouTube terkait PILKADA 2024.

Pada klasifikasi sentimen, SVM dengan TF-IDF, ROS, dan rasio data 80:20 mencapai akurasi terbaik sebesar 0.87 pada data train dan 0.80 pada data test, sementara pada klasifikasi sarkasme, SVM dengan TF-IDF tanpa resampling menghasilkan akurasi tertinggi sebesar 0.80 pada data train dan 0.73 pada data test. Secara keseluruhan, algoritma SVM terbukti lebih efektif dalam menangkap pola klasifikasi sentimen dengan deteksi sarkasme, menjadikannya pilihan yang lebih optimal untuk analisis komentar YouTube terkait PILKADA 2024. Untuk pengembangan lebih lanjut, disarankan untuk menerapkan langkah preprocessing tambahan seperti normalisasi teks dan tokenizing, mengeksplorasi algoritma deep learning seperti LSTM, GRU, atau CNN guna meningkatkan performa model, serta menambah dataset agar model dapat mempelajari pola yang lebih luas dan menghasilkan hasil yang lebih optimal.

DAFTAR PUSTAKA

[1] J. Kristiyono, "BUDAYA INTERNET: PERKEMBANGAN TEKNOLOGI INFORMASI DAN KOMUNIKASI DALAM MENDUKUNG PENGGUNAAN MEDIA DI MASYARAKAT," *Scriptura*, vol. 5, no. 1, Oct. 2015, doi: 10.9744/scriptura.5.1.23-30.

[2] Datareportal, "Data Reportal," https://datareportal.com/reports/digital-2024-indonesia.

[3] X. Cheng, C. Dale, and J. Liu, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study," Jul. 2007, [Online]. Available: http://arxiv.org/abs/0707.3670

[4] 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," 2015.

[5] Imam Fahrur Rozi, Sholeh Hadi Pramono, and Erfan Achmad Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi," *jurnaleeccis*, vol. 6, no. https://jurnaleeccis.ub.ac.id/index.php/eeccis/issue/view/21, pp. 37–43, Jun. 2013.

[6] M. Bouazizi, *2015 IEEE Global Communications Conference (GLOBECOM) : proceedings : San Diego, California, 6-10 December 2015*. IEEE, 2015.

[7] A. Muhaddisi, B. N. Prastowo, and D. U. Kusumaning Putri, "Sentiment Analysis With Sarcasm Detection On Politician's Instagram," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 4, p. 349, Oct. 2021, doi: 10.22146/ijccs.66375.

[8] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, "Sarcasm detection using machine learning algorithms in Twitter: A systematic review," *International Journal of Market Research*, vol. 62, no. 5, pp. 578–598, Sep. 2020, doi: 10.1177/1470785320921779.

[9] Y. Yunitasari, A. Musdholifah, and A. K. Sari, "Sarcasm Detection For Sentiment Analysis in Indonesian Tweets," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, p. 53, Jan. 2019, doi: 10.22146/ijccs.41136.

[10] H. Al Rasyid Harpizon *et al.*, "Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 1, 2022.

[11] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*. Springer Science & Bussines Media, 2007.

[12] Balya, "Analisis Sentimen Pengguna Youtube di Indonesia pada Review Smartphone Menggunakan Naïve Bayes.," 2019.

[13] A. N. Hidayat, "ANALISIS SENTIMEN TERHADAP WACANA POLITIK PADA MEDIA MASA ONLINE MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES," 2015.

[14] X. Cheng, C. Dale, and J. Liu, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study," Jul. 2007, [Online]. Available: http://arxiv.org/abs/0707.3670

[15] D. Alita and A. Rahman, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," 2020.

[16] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 2, no. 5, pp. 178–185, Apr. 2022, doi: 10.30865/klik.v2i5.362.

[17] Santosa B, "Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.," *Graha Ilmu : Yogyakarta*, 2007.

[18] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1," *Teori dan Aplikasinya dalam Bioinformatika 1*, 2003, [Online]. Available: http://asnugroho.net

[19] L. Breiman, *Random Forest*, vol. 45. Springer, 2001.