LAPORAN PROJECT PENAMBANGAN DATA
SEMESTER GENAP 2024/2025

IDENTITAS PROYEK	
Judul	Klasifikasi Sentimen Berita melalui Reduksi Dimensi PCA
Topik	Analisis Sentimen
Identitas Penyusun	1. Shinta Usaila Farachin (23031554160) 2. Ikhrima Atusifah (23031554181) 3. Arina Tri Yuni Wahyuning Tiyas (23031554203)
Kelas	2023 A

1. PENDAHULUAN

Pendahuluan penelitian tidak lebih dari 1000 kata yang terdiri dari:
A. Latar belakang dan rumusan permasalahan yang akan diteliti
B. Pendekatan pemecahan masalah

1.1. Latar Belakang (min. 250 kata)

Dalam era modern saat ini, data berita yang didapatkan dari pengguna internet, seperti komentar, ulasan produk, dan postingan blog, semakin meningkat. Adanya analisis sentimen berguna untuk memeriksa apakah suatu berita mengandung pendapat yang positif ataudan negatif menjadi sesuatu yang penting dalam berbagai aplikasi, seperti pemantauan merek, analisis ulasan, dan pengambilan keputusan bidang bisnis. Akan tetapi, terdapat tantangan utama dalam menganalisis data berita yaitu volume dan dimensi data yang lumayan besar. Setiap data berita per kata dianggap memiliki dimensi tersendiri, sehingga dalam implementasi data berita seperti metode bag of words atau tf-idf, jumlah dimensi yang didapatkan bisa sangat besar, bisa mencapai ribuan atau bahkan jutaan dimensi [1]. Hal ini tentu akan berpotensi membuat klasifikasi sentimen menjadi lebih rumit dan mudah terkena overfitting [2]. Sehingga pada projek ini peran data mining sangat penting untuk menggali informasi dari data teks yang besar, data yang besar ini akan sulit dianalisis tanpa menerapkan teknik-teknik yang tepat.

Pada projek ini, kita menggunakan Reduksi dimensi yang mengacu pada teknik yang mengurangi jumlah fitur input dalam set data. Teknik pengurangan dimensi dapat menggunakan metode analitik yang diawasi dan tidak diawasi. Namun, karakteristik dari algoritma Reduksi dimensi bervariasi tergantung pada jenis algoritma yang digunakan. Sebagai contoh, dalam kasus algoritma pembelajaran tak terawasi, teknik reduksi reduksi dimensi harus bertujuan untuk meminimalkan kehilangan informasi fitur. Sedangkan dalam kasus pembelajaran yang diawasi, teknik ini harus bekerja

untuk memaksimalkan informasi kelas. Karena sifat kompleks dari proses Reduksi dimensi, tidak ada metode tunggal yang cocok untuk menangani semua situasi [3].

Secara umum, teknik Reduksi dimensi dapat berupa linier dan non linier. Reduksi dimensi linier mengubah data menjadi ruang dimensi rendah sebagai kombinasi linier dari variabel asli. Hal ini dapat diklasifikasikan secara luas ke dalam dua kelompok. Kelompok pertama mengacu pada teknik yang memanfaatkan informasi keanggotaan kelas saat menghitung ruang dimensi rendah. Contoh dari metode tersebut termasuk berbagai skema pemilihan fitur yang mengurangi dimensi dengan memilih subset dari fitur asli dan teknik yang mendapatkan fitur baru dengan kombinasi garis dari istilah-istilah tersebut. Jenis kedua dari teknik pengurangan dimensi adalah algoritma komputasi berdasarkan analisis statistik [4].

Setelah data direduksi dimensinya menggunakan pca, algoritma klasifikasi k-nn, svm dan naive bayes dapat diterapkan untuk mendapatkan klasifikasi sentimen berita. Pengertian salah satu algoritma yang digunakan yaitu knn, KNN merupakan algoritma yang sederhana namun efektif, dengan mengklasifikasikan data berdasarkan kedekatan dengan titik data lain yang ada dalam sebuah ruang fitur [5]. Dalam suatu penelitian menunjukkan bahwa kombinasi PCA dengan knn, svm dan naive bayes dalam analisis sentimen mendapatkan hasil yang berbeda dalam pengaplikasiannya. Seperti contoh, penelitian oleh Supriyanto dkk [6].

1.2. Rumusan Masalah dan Tujuan

Rumusan masalah

1. Bagaimana implementasi metode PCA dapat mereduksi dimensi data berita dalam analisis sentimen?
2. Seberapa efektif kombinasi PCA dengan knn, svm dan naive bayes dalam meningkatkan akurasi klasifikasi sentimen?
3. Bagaimana peran data mining dalam mengoptimalkan proses analisis sentimen berbasis data berita?

Tujuan

1. Mengimplementasi metode PCA dapat mereduksi dimensi data berita dalam analisis sentimen
2. Mengkaji efektifitas kombinasi PCA dengan knn, svm dan naive bayes dalam meningkatkan akurasi klasifikasi sentimen

3. Menganalisis peran data mining dalam proses analisis sentimen terhadap data berita

Manfaat

1. Memberikan solusi untuk mengatasi dimensi data berita dalam analisis sentimen menggunakan metode reduksi dimensi PCA
2. Menunjukkan efektivitas kombinasi PCA dengan knn,svm dan naive bayes dalam meningkatkan akurasi dan efisiensi klasifikasi data berita
3. Memperkuat pemanfaatan data mining dalam proses ekstraksi informasi dan pengambilan keputusan

2. Metodologi

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

2.1. Eksplorasi Dataset

Pemahaman dataset yang dimiliki

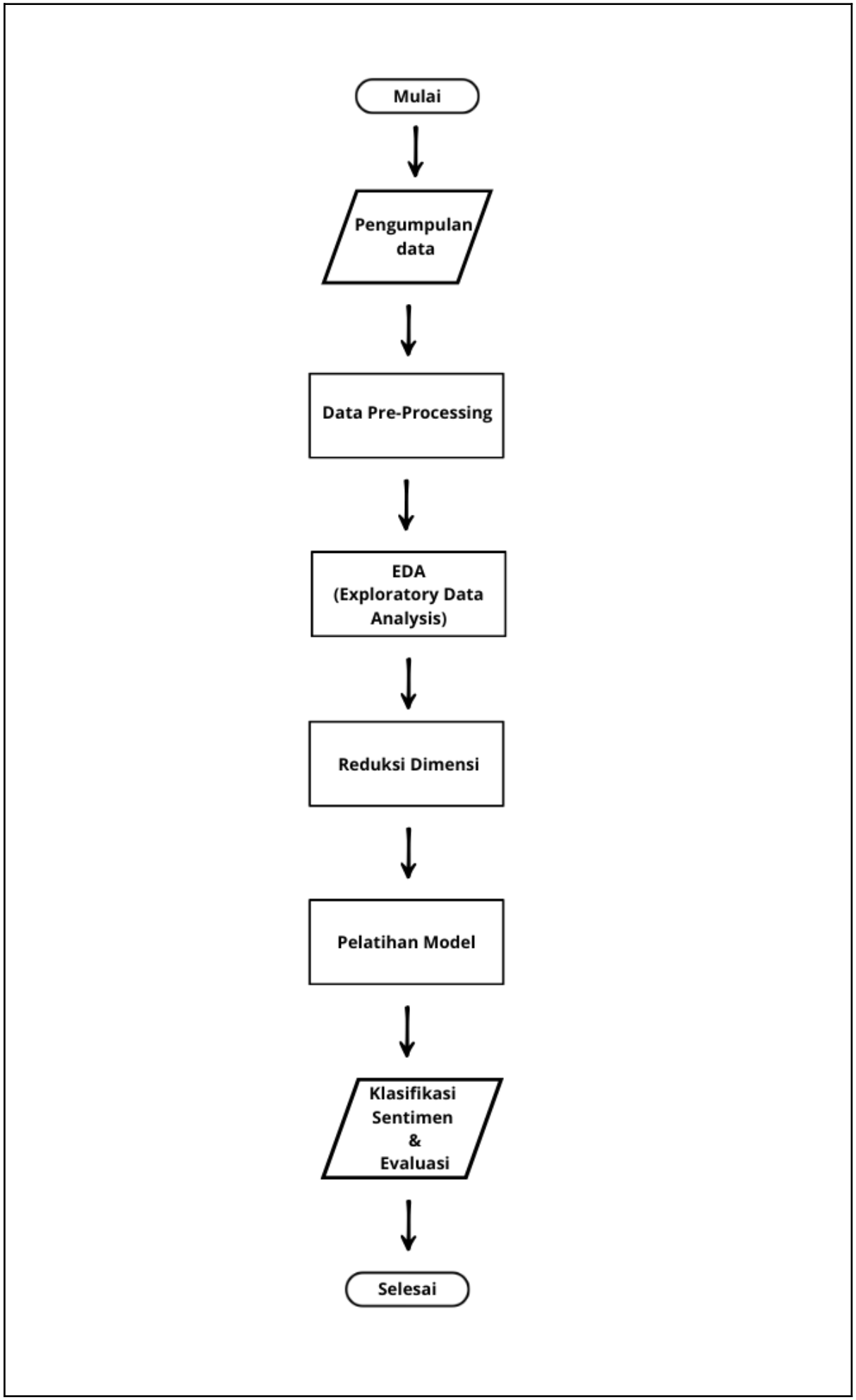
Dataset diambil dari kaggle dengan judul datasetnya yaitu tweet analisis sentiment berita. Dalam dataset ini berisi data sentiment tentang berita. Data ini akan digunakan dalam analisis sentimen dengan kategori sentimen positif, sentimen negatif dan sentiment netral. Pada dataset ini terdiri dari 3500 baris dan 8 kolom yaitu kolom source, author, title, description, url, published at, sentiment dan type. Dataset ini diambil per 1 Juli 2024. Dataset ini diambil menggunakan Media Stack API. Dataset ini masih mengandung tanda baca, emoticon dan kata kata yang tidak baku sehingga perlu diproses lebih lanjut. Data berita yang akan kita analisis yaitu pada description dan sentiment. Untuk yang direduksi yaitu data berita pada kolom description.

Berikut link dataset dari kaggle :

https://www.kaggle.com/datasets/clovisdalmolinvieira/news-sentiment-analysis/data

2.2. Langkah Penelitian

Berikut diagram alur dalam proses analisis sentimen


Mulai
Pengumpulan data
Data Pre-Processing
EDA (Exploratory Data Analysis)
Reduksi Dimensi
Pelatihan Model
Klasifikasi Sentimen & Evaluasi
Selesai

Dalam melakukan analisis sentimen terhadap data berita seperti tweet, diperlukan beberapa tahapan sistematis untuk mendapat hasil yang baik dan efisien. Berdasarkan studi Omuya dkk [7] serta berbagai referensi lain, berikut langkah-langkah yang dilakukan:

1. Pengumpulan data (Data Collection)

   Tahap pertama dalam projek ini yaitu megumpulkan data yang akan dianalisis. Dalam projek ini data didapat dari kaggle dengan judul analisis sentimen berita.

2. Data Preprocessing

   Pada tahap ini akan dicek missing value dan duplikat, setelah dilihat terdapat missing value di kolom author dan ada sebanyak 737 baris duplikat, setelah dicek kita menghapus baris duplikat sedangkan bagian missing value kami biarkan karena tidak mempengaruhi hasil akhir. Pada tahapan ini juga data berita dibersihkan sebelum dianalisis, proses ini mencangkup:

   1. Menghapus karakter khusus, angka, URL dan emoticon yang tidak sesuai
   2. Tokenisasi teks menjadi kata-kata individual
   3. Mengubah huruf menjadi lowercase agar konsisten
   4. Menghapus stopwords
   5. Memberi label (kata kerja, kata sifat dll) dan hanya memilih kata yang penting. Tahap ini cukup penting agar model bekerja pada data yang lebih terstruktur dan bersih [8].
   6. Visualisasi kategori sentimen

3. Exploratory Data Analysis (EDA)

   Pada tahapan ini mencangkup:

   1. Menghitung panjang berita : melihat kecenderungan panjang kata terhadap suatu sentiment
   2. Word cloud dan Frequent Terms : Menunjukkan kata yang paling sering muncul pada setiap sentiment
   3. POS (Part Of Speech) : mengukur proporsi kata kerja, kata sifat dan kata keterangan

   Untuk hasil EDA akan dijelaskan pada bagian hasil EDA

4. Reduksi Dimensi

   Dataset memiliki ratusan fitur kata kata sehingga perlu dilakukan reduksi, pada tahapan ini akan dilakukan TF-IDF dan PCA (Principal Component Analysis). PCA merupakan metode yang digunakan untuk mengurangi

dimensionalitas kumpulan data tertentu [9]. Pada tahap ini juga didapatkan hasil jumlah kata asli sebanyak 19269 kemudian dilakukan pca dengan persentase 95% didapatkan jumlah kata menjadi 1807

5. Pelatihan Model

   Data yang sudah diproses kemudian kan dibagi menjadi 70% data latih dan 30% data uji untuk kemudian dilatih menggunakan algoritma KNN, SVM dan Naive Bayes. Pemilihan salah satu algoritma seperti algoritma knn didasarkan karena alasan algoritma ini memiliki tingkat kesalahan cenderung optimal bayes dalam kondisi ringan, hal ini karena ukuran sampel cenderung besar [10].

6. Klasifikasi Sentimen dan Evaluasi

   Pada tahap ini model kemudian digunakan untuk mengklasifikasikan tweet baru menjadi kategori positif atau negatif. Untuk mengukur performa model digunakan metrik: accuracy (tingkat prediksi yang benar), precision (ketepatan dalam klasifikasi positif), recall (kemampuan dalam menangkap semua data positif) dan F1-Score (keseimbangan antara precision dan recall). Evaluasi ini penting untuk memastikan generelisasi model terhadap data baru.

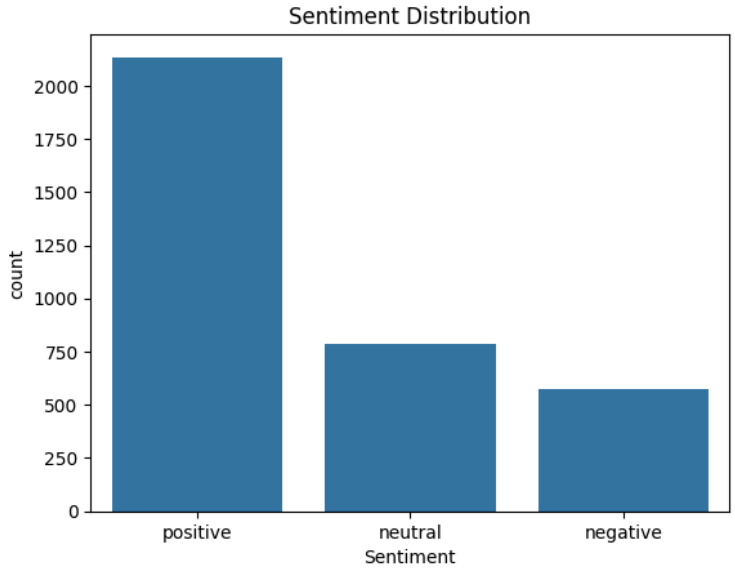
3. Hasil dan Analisis

3.1 Exploratory Data Analysis

Berikut adalah EDA dari projek ini:

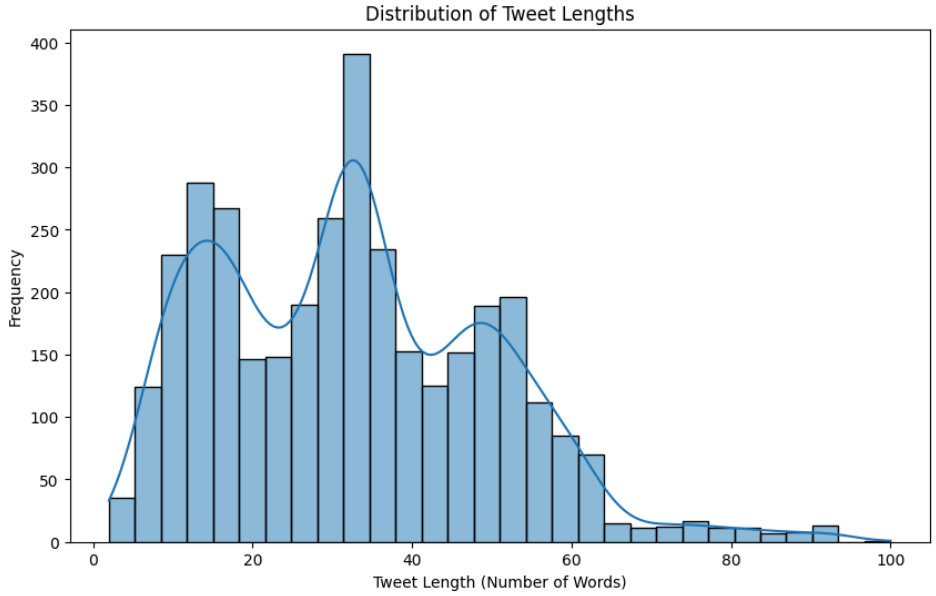
1. Distribusi sentiment

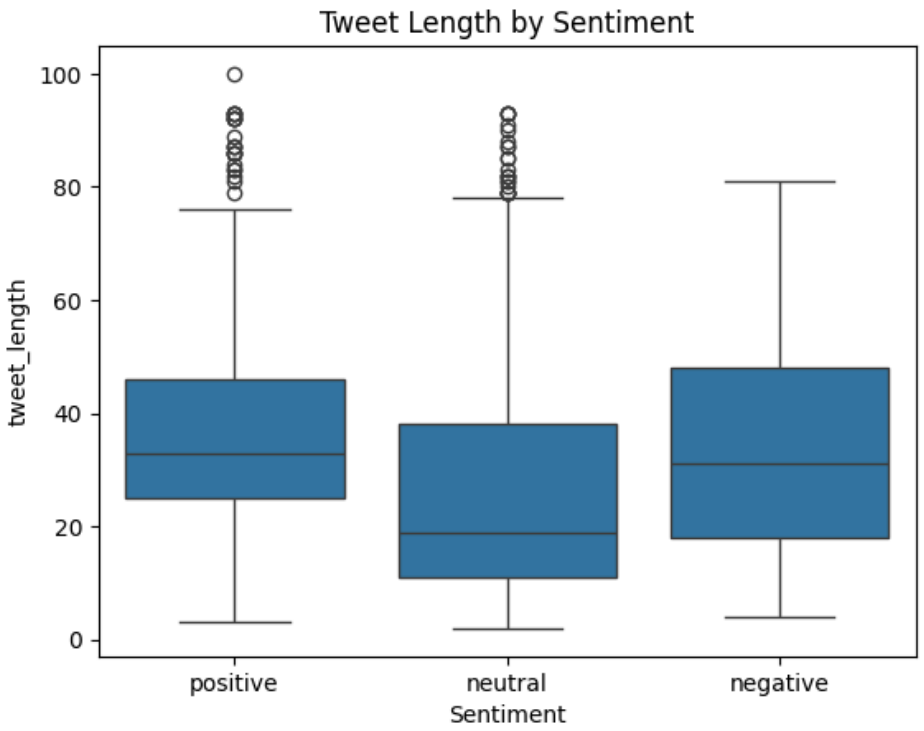
   Dataset yang digunakan berisi kolom description untuk teks berita atau opini, serta kolom sentiment yang menjadi label target. Visual awal memperlihatkan sebaran data berdasarkan sentimen. Hasilnya menunjukkan data bersifat tidak seimbang dengan data sentimen positif lebih banyak, kemudian dilanjutkan data sentiment netral dan terakhir data sentimen negatif.



2. Distribusi panjang teks

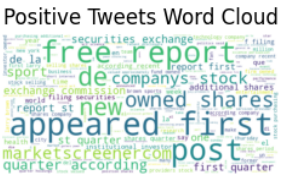
Berdasarkan visualisasi yang dilakukan, menunjukkan hasil bahwa teks dengan sentimen positif dan negatif lebih banyak daripada teks netral. Pola ini menunjukkan bahwa ekspresi yang bersifat positif dan negatif biasanya dituangkan dalam teks yang lebih panjang dan rinci, sementara yang netral lebih singkat dan langsung ke intinya.

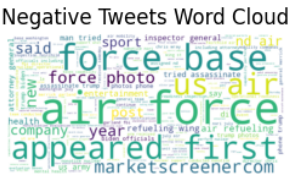


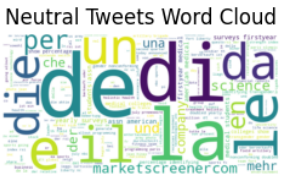


3. Word cloud dan Frequent Terms

   Visualisasi ini menunjukkan bahwa kata dominan pada sentiment positif yaitu kata free, report, appeared first, dimana kata kata ini menunjukkan kecenderungan pada informasi atau promosi yang bersifat positif, sementara kata dominan pada sentimen negatif didominasi oleh kata kata seperti air, force dan base yang mengartikan tentang isu militer atau kejadian tertentu yang lebih berbahaya. Sementara sentimen negatif didominasi kata seperti de dan dail yang sesuai dengan karakteristik netral dan tidak terlalu spesifik. Pola ini menunjukkan bahwa setiap sentimen memiliki ciri khas kosakata sendiri.



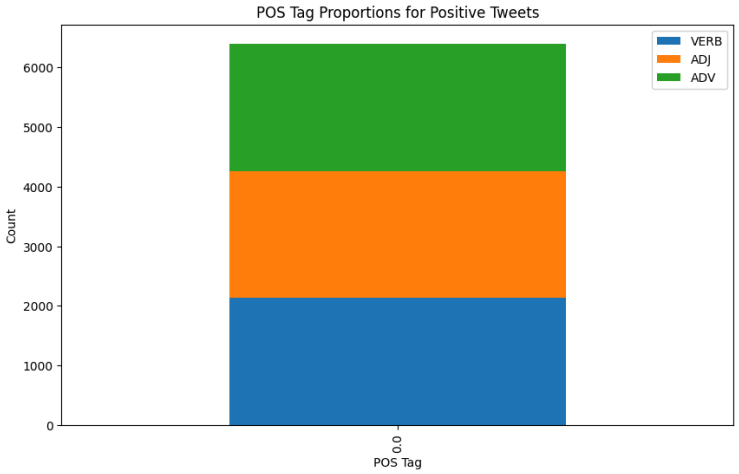


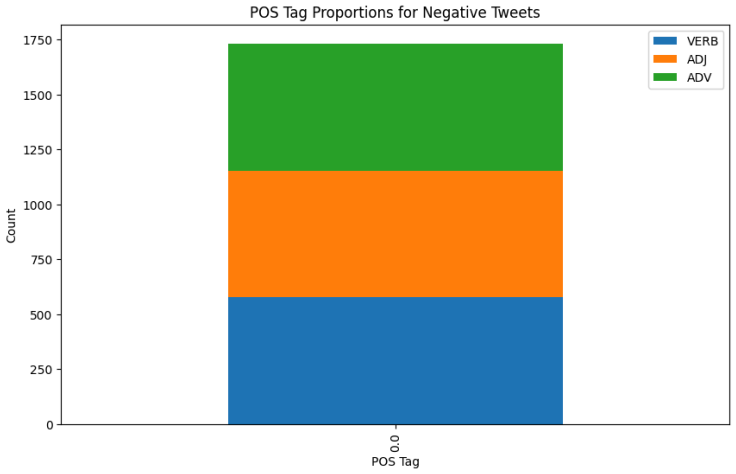


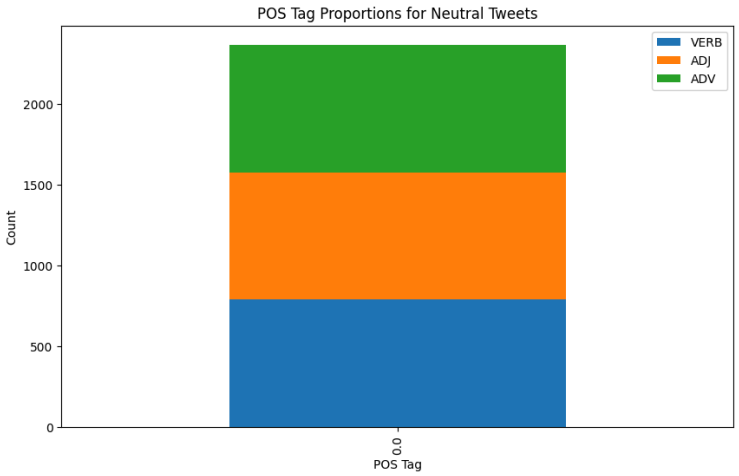
4. POS (Part Of Speech)

   Berdasarkan visualisasi distribusi pos menunjukkan bahwa tag pada sentimen positif, negatif dan netral memiliki pola distribusi yang seimbang antara kata kerja, kata sifat dan kata keterangan, akan tetapi berbeda secara jumlah. Berita dengan sentimen positif memiliki jumlah pos tertinggi, hal ini menunjukkan berita dengan sentimen positif cenderung lebih panjang dan ekspresif, dengan jumlah sekitar 2.100 kata kerja, kata sifat dan kata keterangan. Sementara

berita dengan sentimen netral memiliki jumlah menengah (780 per jenis pos), dan berita negatif memiliki pos cenderung lebih singkat atau kurang ekspresif. Pola ini menunjukkan bahwa jenis sentiment berkaitan dengan seberapa banyak dan beragam pos tag yang digunakan.







3.2 Insight dan hasil mining dari project

Berdasarkan hasil analisis terhadap tiga algoritma klasifikasi, yaitu K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Naive Bayes, baik dengan maupun tanpa penerapan Principal Component Analysis (PCA), dapat disimpulkan

bahwa PCA memberikan efek positif yang sangat signifikan terhadap peningkatan performa model. Dalam kondisi menggunakan PCA, model SVM dengan pembagian 70/30 menunjukkan hasil terbaik dengan akurasi mencapai 78,17%, precision 85,83%, recall 62,71%, dan F1-score tertinggi adalah 68,43%. Dalam validasi silang 10-fold, SVM masih menunjukkan kinerja yang baik dengan akurasi 77,42%, presisi 86,34%, recall 61,95%, serta F1-score 67,19%. Model KNN dengan PCA yang menerapkan pembagian data 70/30 memperoleh akurasi 69,36%, precision 62,10%, recall 59,33%, dan F1-score 60,48%, sementara pada 10-fold cross-validation mendapatkan akurasi 70,07%, precision 64,48%, recall 60,97%, serta F1-score 62,15%. Sementara itu, Naive Bayes yang menggunakan PCA menunjukkan performa terendah, meskipun masih lebih baik dibandingkan tanpa PCA, yaitu akurasi 64,41%, precision 57,13%, recall 55,82%, dan F1-score 56,21% pada pembagian 70/30, serta akurasi 61,85%, precision 54,13%, recall 53,71%, dan F1-score 53,71% pada validasi silang.

Sebaliknya, evaluasi tanpa penerapan PCA menunjukkan bahwa terdapat penurunan performa yang signifikan pada ketiga model. Model SVM tanpa PCA hanya mencatat akurasi 62,97%, presisi 55,03%, recall 36,51%, dan F1-score 32,76% pada pembagian 70/30, serta akurasi 61,96%, presisi 49,35%, recall 35,85%, dan F1-score 30,66% pada validasi silang. Model KNN yang tidak menggunakan PCA mencatat akurasi 62,36%, precision 52,83%, recall 51,25%, dan F1-score 51,83% dalam pembagian data, serta akurasi 61,92%, precision 52,45%, recall 49,87%, dan F1-score 50,72% pada 10-fold CV. Penurunan paling signifikan terjadi pada Naive Bayes, dengan akurasi 62,12%, precision 20,71%, recall 33,33%, dan F1-score 25,55% pada pembagian data, serta akurasi 60,91%, precision 20,30%, recall 33,33%, dan F1-score 25,24% pada validasi silang. Temuan ini menunjukkan bahwa tanpa adanya proses reduksi dimensi, semua model mengalami kesulitan dalam mengelola data teks dengan dimensi tinggi, yang mengakibatkan penurunan akurasi, konsistensi, dan kestabilan klasifikasi.

Perbedaan nilai akurasi, precision, recall, dan F1-score yang dihasilkan oleh setiap model dapat dijelaskan melalui karakteristik algoritma dan kepekaannya terhadap dimensi data. SVM sangat baik dalam mengklasifikasikan data berdimensi tinggi karena kemampuannya untuk membentuk hyperplane optimal yang memisahkan kelas dengan jelas. Saat PCA diterapkan, fitur-fitur yang tidak penting dihapus agar SVM dapat lebih memfokuskan diri pada dimensi yang berinformasi, menghasilkan

akurasi dan presisi yang tinggi. Sebaliknya, tanpa PCA, adanya banyak fitur yang tidak relevan membuat SVM kesulitan dalam melakukan pemisahan yang efisien, terlihat dari penurunan nilai recall dan F1-score. KNN, sebagai algoritma berbasis instance dan jarak, menunjukkan kinerja yang cukup stabil. PCA berperan dalam mempertahankan jarak antar data agar tetap relevan dengan hanya menyimpan dimensi yang signifikan, yang meningkatkan konsistensi prediksi. Namun tanpa PCA, KNN mulai terpengaruh oleh curse of dimensionality, di mana jarak antar titik menjadi kurang mencerminkan. Sementara itu, Naive Bayes sangat tergantung pada asumsi kemandirian antara fitur, yang dalam data teks sering kali tidak dipenuhi. Jumlah fitur yang saling terkait menyebabkan model menghasilkan estimasi probabilitas yang tidak tepat. Itulah mengapa precision dan F1-score model ini sangat rendah, terutama tanpa PCA. Meskipun PCA diterapkan, peningkatan kinerja Naive Bayes tetap tidak signifikan sebab model ini tidak dapat menangkap hubungan kompleks antara fitur seperti kedua algoritma lainnya

Dalam konteks ini, penambangan data memiliki peran krusial dalam meningkatkan proses analisis sentimen yang berbasis teks berita. Proses penambangan data tidak hanya terbatas pada pemodelan, tetapi juga melibatkan seluruh rangkaian seperti pengumpulan data, pembersihan, dan preprocessing data (melalui tokenisasi, penghapusan stopword, pelabelan POS), serta seleksi fitur dan pengurangan dimensi. Metode seperti PCA, yang merupakan elemen dari proses penggalian data, terbukti berhasil dalam mengekstrak informasi yang signifikan dan menghapus fitur yang tidak berkaitan atau berlebihan. Ini memungkinkan model beroperasi dengan lebih cepat, efisien, dan tepat. Selain itu, data mining mendukung penerapan berbagai algoritma machine learning seperti SVM, KNN, dan Naive Bayes, serta metode evaluasi seperti cross-validation dan metrik kinerja (akurasi, presisi, recall, F1-score).

Evaluasi KNN, SVM, dan Naive Bayes - Split 70/30 dan 10-Fold Cross Validation

Metode	Akurasi	Presisi	Recall	F1-Score
KNN - Split 70/30	62.364294330518696	52.82654008759508	51.25387088291015	51.82856752428591
KNN - 10-Fold CV	61.924867890964265	52.44748560262125	49.8666503414306	50.71558437694219
SVM - Split 70/30	62.96743063932448	55.032745591939545	36.506786456069165	32.75820823525271
SVM - 10-Fold CV	61.961361377073196	49.35020164581106	35.849590533586984	30.664636708486135
Naive Bayes - Split 70/30	62.12303980699638	20.70767993566546	33.33333333333333	25.545634920634924
Naive Bayes - 10-Fold CV	60.91194474964684	20.30398158321561	33.333333333333336	25.236132993980533

Evaluasi KNN, SVM, dan Naive Bayes - Split 70/30, 10-Fold Cross Validation, dan PCA

Metode	Akurasi	Presisi	Recall	F1-Score
KNN - Split 70/30	69.36067551266586	62.10116139619536	59.327687345365945	60.477725359385346
KNN - 10-Fold CV	70.06893214042798	64.48309493074109	60.96781487417607	62.14683075131189
SVM - Split 70/30	78.1664656212304	85.82638920978184	62.71417990951392	68.43407445218261
SVM - 10-Fold CV	77.4168105477947	86.33934682218927	61.95334871301404	67.18771355392012
Naive Bayes - Split 70/30	64.41495778045838	57.1318744533477	55.82406199852946	56.206614012465074
Naive Bayes - 10-Fold CV	61.85083451054256	54.12585841503592	53.70635794938708	53.7060985241439

3. Daftar Pustaka

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

[1]. Murali Krishna, M., & Lavanya Devi, G. (2021). Method of optimizing the dimensional features in sentiment analysis. *International Journal of Computers and Applications*, *43*(7), 643-652.

[2]. Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, *174*, 114765.

[3]. Srivastava, S., Chakraborty, C., & Sarkar, M. K. (2024). Leveraging machine learning and dimensionality reduction for sports and exercise sentiment analysis. *Measurement: Sensors*, *33*, 101182.

[4]. Singh, K. N., Devi, S. D., Devi, H. M., & Mahanta, A. K. (2022). A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights*, *2*(1), 100061.

[5]. Sun, B., & Chen, H. (2021). A survey of k nearest neighbor algorithms for solving the class imbalanced problem. *Wireless Communications and Mobile Computing*, *2021*(1), 5520990.

[6]. Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) algorithm for public sentiment analysis of online learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(2), 121-130.

[7]. Omuya, E. O., Okeyo, G., & Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. *Engineering Reports*, *5*(3), e12579.

[8]. Distante, D., Faralli, S., Rittinghaus, S., Rosso, P., & Samsami, N. (2022). Domainsenticnet: An ontology and a methodology enabling domain-aware sentic computing. *Cognitive computation*, *14*(1), 62-77.

[9]. Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100.

[10]. Zhang, S. (2021). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4663-4675.