**LAPORAN TUGAS AKHIR**

**MATA KULIAH TEXT MINING & NATURAL LANGUAGE PROCESSING**

Logo

Description automatically generated

Tim Penyusun:

1. <5231811022> <Lathif Ramadhan>
2. <5231811029> <Andini Angel Meivita>
3. <5231811033> <Rama Panji Nararendra>
4. <5231811036> <Giffari Riyanda Pradithya>

**PROGRAM STUDI SAINS DATA PROGRAM SARJANA**

**FAKULTAS SAINS & TEKNOLOGI**

**UNIVERSITAS TEKNOLOGI YOGYAKARTA**

**2025**

**SUMMARY PERBANDINGAN KLASIFIKASI ALGORITMA DECISION TREE**

**PENGATURAN AWAL DAN TUNING PERBAIKAN, SERTA ALGORITMA KLASIFIKASI LAINNYA**

**Nama Dataset: ramadan\_labeled\_sentiment.csv (Dataset 2 label: Positive, Negative)**

|  | **Decision Tree awal** | **Decision Tree Perbaikan** | **Algoritma Random Forest** | **Algoritma Neural Network** | **Algoritma SVM** | **Algoritma Logistic Regression** |
| --- | --- | --- | --- | --- | --- | --- |
| **Split Data** (cross atau split validation), berapa perbandingannya | cross dan split validation, perbandingan (75/25) | cross dan split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) |
| **Komposisi ngram** | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) |
| **Model stemming** | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | if token.isalpha() and token not in english\_stopwords:  # Lakukan lemmatization  # 'v' untuk verb, 'n' untuk noun, dll. Defaultnya 'n'. Mungkin perlu eksplorasi POS tagging  lemmatized\_token = lemmatizer.lemmatize(token)  processed\_tokens.append(lemmatized\_token) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) |
| **Max\_depth** | 40 | 80 | 80 | - | - | - |
| **Min\_samples\_split** | 20 | 25 | 25 | - | - | - |
| **Criterion** | 'entropy' | 'entropy' | ‘gini’ | - | - | - |
| **SMOTE** (oversampling atau Undersampling) | Belum menggunakan SMOTE | Tidak menggunakan SMOTE karena menyebabkan akurasi turun | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling |
| **Prosentase Akurasi** | 70% | 80% | 87.15% | 89% | 84.52% | 83.92% |
| **Tuliskan komposisi True positif, true negative, false positif, false negatif** | True Negatives (TN): 79  False Positives (FP): 18  False Negatives (FN): 44  True Positives (TP): 68 | True Negatives (TN): 63  False Positives (FP): 13  False Negatives (FN): 21  True Positives (TP): 71 | True Negatives (TN): 72  False Positives (FP): 4  False Negatives (FN): 18  True Positives (TP): 74 | True Negatives (TN): 69  False Positives (FP): 7  False Negatives (FN): 17  True Positives (TP): 75 | True Negatives (TN): 64  False Positives (FP): 12  False Negatives (FN): 15  True Positives (TP): 77 | True Negatives (TN): 66  False Positives (FP): 10  False Negatives (FN): 17  True Positives (TP): 75 |

**Nama Dataset: ramadan\_labeled\_sentiment.csv (Dataset 3 label: Positive, Negative, Neutral)**

|  | **Decision Tree awal** | **Decision Tree Perbaikan** | **Algoritma Random Forest** | **Algoritma Neural Network** | **Algoritma SVM** | **Algoritma Logistic Regression** |
| --- | --- | --- | --- | --- | --- | --- |
| **Split Data** (cross atau split validation), berapa perbandingannya | ross dan split validation, perbandingan (75/25) | cross dan split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) | split validation, perbandingan (80/20) |
| **Komposisi ngram** | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) | ngram\_range=(1, 2) |
| **Model stemming** | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | if token.isalpha() and token not in english\_stopwords:  # Lakukan lemmatization  # 'v' untuk verb, 'n' untuk noun, dll. Defaultnya 'n'. Mungkin perlu eksplorasi POS tagging  lemmatized\_token = lemmatizer.lemmatize(token)  processed\_tokens.append(lemmatized\_token) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) | # Inisialisasi Lemmatizer dan Stop Words  lemmatizer = WordNetLemmatizer()  english\_stopwords = set(nltk\_stopwords.words('english')) |
| **Max\_depth** | None | None | None | - | - | - |
| **Min\_samples\_split** | 2 | 2 | 3 | - | - | - |
| **Criterion** | 'gini' | 'entropy' | ‘gini’ | - | - | - |
| **SMOTE** (oversampling atau Undersampling) | Belum menggunakan SMOTE | Tidak menggunakan SMOTE karena menyebabkan akurasi turun | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling | Menggunakan SMOTE oversampling |
| **Prosentase Akurasi** | 80% | 87% | 87.15% | 95,91% | 8725 | 94,23% |
| **Tuliskan komposisi True positif, true negative, false positif, false negatif** | True Negatives (TN): 2  False Positives (FP): 2  False Negatives (FN): 1  True Positives (TP): 31 | True Negatives (TN): 0  False Positives (FP): 5  False Neutrals (FN\_Neg): 0  True Neutrals (TN): 203  False Positives (FP\_Neg): 5  False Negatives (FN\_Neg): 22  True Positives (TP): 16 | True Negatives (TN): 269  False Positives (FP): 3  False Neutrals (FN\_Neg): 2  True Neutrals (TN): 250  False Positives (FP\_Neg): 0  False Negatives (FN\_Neg): 1  True Positives (TP): 118 | True Negatives (TN): 264  False Positives (FP): 8  False Neutrals (FN\_Neg): 7  True Neutrals (TN): 262  False Positives (FP\_Neg): 4  False Negatives (FN\_Neg): 1  True Positives (TP): 126 | True Negatives (TN): 268  False Positives (FP): 4  False Neutrals (FN\_Neg): 4  True Neutrals (TN): 230  False Positives (FP\_Neg): 3  False Negatives (FN\_Neg): 7  True Positives (TP): 102 | True Negatives (TN): 272  False Positives (FP): 0  False Neutrals (FN\_Neg): 0  True Neutrals (TN): 254  False Positives (FP\_Neg): 7  False Negatives (FN\_Neg): 1  True Positives (TP): 116 |