JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10593

# ANALISIS KARAKTERISTIK DATASET TERJEMAHAN AL-QUR'AN BAHASA INDONESIA UNTUK KLASIFIKASI TEKS MULTI-LABEL

**Ismail Akbar[1*], Affi Nizar Suksmawati[2], Firman Nurdiyansyah[3], Devita Putri Anggraini[4]**

[1,2,4]Program Studi S1 Sistem dan Teknologi Informasi; Jl. Borobudur 35, Malang; (0341) 492282
[3]Program Studi S1 Teknik Informatika; Jl. Borobudur 35, Malang; (0341) 492282

**Keywords:**
Analisis dataset;
Multi-label classification;
Karakteristik dataset;
Al-Qur'an Bahasa Indonesia;
Label correlation.

**Corespondent Email:**
ismaelakbar12@gmail.com

**Abstrak.** Karakteristik *dataset* merupakan salah satu faktor yang memengaruhi keberhasilan klasifikasi teks *multi-label*. Meskipun berbagai penelitian telah mengembangkan model klasifikasi pada terjemahan Al-Qur'an Bahasa Indonesia, analisis terhadap karakteristik intrinsik *dataset* masih terbatas. Penelitian ini bertujuan menganalisis karakteristik *dataset* sebagai dasar pengembangan metode klasifikasi *multi-label*. *Dataset* berasal dari terjemahan resmi Al-Qur'an Kementerian Agama Republik Indonesia Tahun 2022 yang terdiri atas 6.236 ayat dari 114 surah, sedangkan analisis *multi-label* dilakukan pada subset berlabel sebanyak 461 ayat dengan empat label tematik, yaitu Tauhid, Ibadah, Akhlaq, dan Sejarah. Analisis meliputi profil *dataset*, karakteristik dokumen, distribusi label, Label *Cardinality*, Label *Density*, dan Label *Correlation* menggunakan matriks *co-occurrence*, *heatmap*, serta *network graph*. Hasil penelitian menunjukkan distribusi label yang relatif seimbang, nilai Label *Cardinality* sebesar 2,02, Label *Density* sebesar 0,506, serta hubungan antartema pada tingkat sedang hingga kuat. Selain itu, variasi panjang dokumen menunjukkan keragaman representasi teks yang memadai. Temuan ini menunjukkan bahwa *dataset* memiliki kompleksitas *multi-label* yang representatif dan layak digunakan sebagai dasar pengembangan serta evaluasi metode klasifikasi *multi-label*.



**Abstract.** *Dataset characteristics are one of the factors influencing the success of multi-label text classification. Although various studies have developed classification models for Indonesian translations of the Quran, analysis of the intrinsic characteristics of datasets is still limited. This study aims to analyze dataset characteristics as a basis for developing a multi-label classification method. The dataset comes from the official translation of the Quran from the Ministry of Religious Affairs of the Republic of Indonesia in 2022, consisting of 6,236 verses from 114 suras. While multi-label analysis was conducted on a labeled subset of 461 verses with four thematic labels: Tauhid, Ibadah, Akhlaq, and Sejarah. The analysis includes dataset profile, document characteristics, label distribution, Label Cardinality, Label Density, and Label Correlation using a co-occurrence matrix, heatmap, and network graph. The results show a relatively balanced label distribution, with a Label Cardinality value of 2.02, Label Density of 0.506, and moderate to strong inter-theme relationships. Furthermore, variations in document length indicate adequate diversity in text representation. These findings indicate that the dataset has representative multi-label complexity and is suitable for use as a basis for developing and evaluating multi-label classification methods.*

## 1. PENDAHULUAN

Perkembangan *Natural Language Processing* (NLP) telah mendorong kemajuan berbagai aplikasi pengolahan bahasa alami, seperti analisis sentimen, klasifikasi dokumen, sistem temu kembali informasi, dan sistem tanya jawab [1], [2]. Salah satu permasalahan yang terus berkembang adalah klasifikasi teks *multi-label*, yaitu proses mengelompokkan sebuah dokumen ke dalam lebih dari satu kategori secara bersamaan [3]. Berbeda dengan klasifikasi *single-label*, pendekatan ini harus mampu mengenali keterkaitan antarlabel yang muncul secara simultan sehingga memiliki tingkat kompleksitas yang lebih tinggi [3]. Oleh karena itu, keberhasilan proses klasifikasi tidak hanya dipengaruhi oleh algoritma yang digunakan, tetapi juga oleh kualitas dan karakteristik *dataset* yang menjadi dasar proses pembelajaran [1], [3].

Perkembangan *machine learning* dan *deep learning* telah menghasilkan berbagai metode yang mampu meningkatkan performa klasifikasi *multi-label*, mulai dari *Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), *Bidirectional Long Short-Term Memory* (BiLSTM), hingga model berbasis *Transformer* [2], [4], [5], [6]. Berbagai penelitian menunjukkan bahwa peningkatan performa model tidak hanya dipengaruhi oleh arsitektur yang digunakan, tetapi juga oleh karakteristik data yang digunakan selama proses pembelajaran [1], [3]. Faktor-faktor seperti distribusi label, ketidakseimbangan data, dan hubungan antarlabel berperan penting dalam menentukan kemampuan model untuk mempelajari pola klasifikasi secara optimal [3]. Meskipun demikian, sebagian besar penelitian masih berorientasi pada pengembangan model klasifikasi, sedangkan analisis terhadap karakteristik dataset sebagai dasar proses pembelajaran masih relatif terbatas [1], [3].

Penerapan *Natural Language Processing* pada teks keagamaan, khususnya Al-Qur'an, juga mengalami perkembangan yang cukup pesat dalam beberapa tahun terakhir. Berbagai penelitian telah dilakukan untuk klasifikasi tema ayat, analisis semantik, pencarian ayat, klasifikasi topik, hingga pengembangan sistem rekomendasi dengan memanfaatkan pendekatan *machine learning* maupun *deep learning* [7], [8], [9]. Pada konteks terjemahan Al-Qur'an Bahasa Indonesia, penelitian sebelumnya telah mengembangkan model klasifikasi *multi-label* menggunakan pendekatan *Long Short-Term Memory* yang menunjukkan kemampuan model dalam mengidentifikasi lebih dari satu tema pada setiap ayat [10]. Penelitian selanjutnya memperluas kajian tersebut melalui perbandingan *Convolutional Neural Network* (CNN), *Bidirectional Long Short-Term Memory* (BiLSTM), dan *FastText* sehingga diperoleh pemahaman mengenai pengaruh arsitektur jaringan serta representasi kata terhadap performa klasifikasi [11]. Meskipun demikian, penelitian-penelitian tersebut masih berfokus pada peningkatan performa model klasifikasi, sedangkan karakteristik *dataset* yang digunakan sebagai dasar proses pembelajaran belum dianalisis secara komprehensif.

Karakteristik dataset merupakan faktor penting dalam penelitian klasifikasi *multi-label* karena berpengaruh terhadap proses pelatihan maupun evaluasi model [12], [13]. Distribusi label yang tidak seimbang, hubungan antarlabel, variasi panjang dokumen, serta tingkat kompleksitas *dataset* dapat memengaruhi kemampuan model dalam mempelajari pola klasifikasi [3], [12], [14]. Analisis terhadap karakteristik tersebut tidak hanya memberikan gambaran mengenai tingkat kesulitan *dataset*, tetapi juga membantu peneliti dalam menentukan metode klasifikasi, strategi validasi, maupun teknik penanganan *label imbalance* yang sesuai [12], [13], [14].

Berdasarkan telaah penelitian terdahulu, masih terdapat *research gap* berupa belum adanya kajian yang secara khusus menganalisis karakteristik *dataset* terjemahan Al-Qur'an Bahasa Indonesia untuk klasifikasi teks *multi-label* [10], [11]. Sebagian besar penelitian berfokus pada peningkatan nilai akurasi, *precision*, *recall*, dan *F1-score*, sedangkan informasi mengenai distribusi label, kompleksitas *dataset*, karakteristik dokumen, serta hubungan antarlabel belum disajikan secara sistematis [10], [11], [12], [13], [14]. Padahal, informasi tersebut penting sebagai dasar dalam memahami perilaku *dataset* sekaligus mendukung pengembangan model klasifikasi yang lebih efektif [12], [13], [14].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan menganalisis

karakteristik *dataset* terjemahan Al-Qur'an Bahasa Indonesia untuk klasifikasi teks multi-label melalui analisis profil *dataset*, karakteristik dokumen, distribusi label, kompleksitas *multi-label*, serta hubungan antarlabel. Hasil penelitian diharapkan dapat memberikan gambaran menyeluruh mengenai sifat intrinsik *dataset* serta menjadi referensi bagi penelitian selanjutnya dalam memilih strategi klasifikasi yang sesuai. Kontribusi utama penelitian ini adalah penyajian analisis karakteristik *dataset* secara komprehensif sebagai landasan pengembangan metode klasifikasi teks *multi-label* pada domain Al-Qur'an Bahasa Indonesia [12], [13], [14].

## 2. TINJAUAN PUSTAKA

### 2.1. Klasifikasi Teks *Multi-Label*

Klasifikasi teks *multi-label* merupakan pendekatan klasifikasi yang memungkinkan satu dokumen memiliki lebih dari satu label secara bersamaan sesuai dengan informasi yang terkandung di dalamnya. Berbeda dengan klasifikasi *single-label* yang hanya menetapkan satu kelas untuk setiap dokumen, pendekatan *multi-label* mampu merepresentasikan berbagai karakteristik atau topik yang muncul secara simultan pada sebuah dokumen. Karakteristik tersebut menyebabkan proses klasifikasi menjadi lebih kompleks karena model tidak hanya mempelajari hubungan antara dokumen dan label, tetapi juga keterkaitan antarlabel (*label correlation*) yang dapat memengaruhi hasil prediksi [3].

Selain pemilihan metode klasifikasi, keberhasilan sistem *multi-label* juga dipengaruhi oleh karakteristik *dataset* yang digunakan selama proses pembelajaran. Distribusi label yang tidak seimbang, jumlah label pada setiap dokumen, serta hubungan antarlabel merupakan faktor-faktor yang dapat memengaruhi kemampuan model dalam mempelajari pola klasifikasi secara optimal. Oleh karena itu, analisis karakteristik *dataset* menjadi tahapan yang penting untuk memahami sifat intrinsik data sebelum digunakan dalam pengembangan maupun evaluasi model klasifikasi *multi-label* [12], [13], [14]. Berdasarkan karakteristik tersebut, penelitian ini memfokuskan analisis pada karakteristik dataset terjemahan Al-Qur'an Bahasa Indonesia sebagai dasar untuk memahami kompleksitas klasifikasi teks *multi-label*.

### 2.2. Karakteristik *Dataset Multi-Label*

Karakteristik *dataset* merupakan salah satu aspek yang menentukan kualitas dan tingkat kompleksitas klasifikasi *multi-label*. *Dataset multi-label* memiliki karakteristik khusus, seperti distribusi label, jumlah label pada setiap dokumen, hubungan antarlabel, dan tingkat ketidakseimbangan label yang memengaruhi proses pembelajaran serta evaluasi model. Oleh karena itu, analisis karakteristik *dataset* diperlukan untuk memahami sifat intrinsik data sebagai dasar dalam pengembangan dan interpretasi model klasifikasi *multi-label* [3], [12], [13], [14].

#### 2.2.1. Profil *Dataset*

Profil *dataset* menggambarkan informasi umum mengenai *dataset* yang digunakan dalam penelitian, seperti jumlah dokumen, jumlah label, jumlah penugasan label (*label assignments*), serta karakteristik dasar lainnya [12], [14]. Informasi tersebut memberikan gambaran awal mengenai ukuran dan ruang lingkup *dataset* sehingga dapat menjadi dasar dalam memahami hasil analisis karakteristik berikutnya. Selain itu, penyajian profil *dataset* memudahkan peneliti dalam membandingkan karakteristik dataset dengan penelitian lain yang menggunakan domain maupun skenario klasifikasi yang berbeda [12]. Pada penelitian ini, profil *dataset* digunakan untuk mendeskripsikan karakteristik umum *dataset* terjemahan Al-Qur'an Bahasa Indonesia sebelum dilakukan analisis yang lebih mendalam terhadap distribusi label dan kompleksitas *multi-label*.

#### 2.2.2. Karakteristik Dokumen

Karakteristik dokumen menggambarkan sifat dasar setiap dokumen dalam dataset, seperti panjang dokumen, jumlah kata, jumlah karakter, serta variasi panjang antar dokumen. Karakteristik tersebut memberikan informasi mengenai keragaman representasi teks yang dapat memengaruhi proses ekstraksi fitur, pembentukan representasi dokumen, serta kinerja model klasifikasi. Oleh karena itu, analisis karakteristik dokumen diperlukan untuk memahami variasi data sebelum dilakukan proses pembelajaran model[12], [14]. Pada penelitian ini, karakteristik dokumen dianalisis menggunakan statistik deskriptif panjang ayat untuk menggambarkan variasi dokumen pada dataset terjemahan Al-Qur'an Bahasa Indonesia.

$$LC\ (D) = \frac{1}{N} \sum_{i=1}^{N} |Y_i| \quad (1)$$

$$LD\ (D) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i|}{|L|} \quad (2)$$

### 2.2.3. Distribusi Label

Distribusi label menunjukkan frekuensi kemunculan setiap label pada *dataset multi-label* [3]. Analisis distribusi label digunakan untuk mengidentifikasi tingkat keseimbangan maupun ketidakseimbangan label (*label imbalance*) yang dapat memengaruhi proses pembelajaran model [3], [15]. Distribusi label juga memberikan informasi mengenai dominasi maupun kelangkaan suatu label sehingga menjadi dasar dalam menentukan strategi klasifikasi dan evaluasi yang sesuai [3]. Pada penelitian ini, distribusi label dianalisis berdasarkan frekuensi kemunculan setiap label pada seluruh ayat dalam *dataset*.

### 2.2.4. Label *Cardinality*

Label *Cardinality* merupakan metrik yang menunjukkan rata-rata jumlah label yang dimiliki oleh setiap dokumen dalam dataset *multi-label*. Nilai ini memberikan gambaran mengenai banyaknya label yang secara umum melekat pada setiap dokumen sehingga dapat digunakan untuk menilai tingkat kompleksitas klasifikasi. Semakin tinggi nilai Label *Cardinality*, semakin banyak label yang harus diprediksi oleh model pada setiap dokumen [16], [17], [18], [19]. Pada penelitian ini, Label *Cardinality* digunakan untuk mengukur rata-rata jumlah label pada setiap ayat dalam *dataset* terjemahan Al-Qur'an Bahasa Indonesia.

Persamaan (1) menunjukkan bahwa nilai Label *Cardinality* diperoleh dengan menghitung rata-rata jumlah label yang dimiliki oleh seluruh dokumen dalam *dataset*. Pada persamaan tersebut, $LC\ (D)$ menyatakan nilai Label *Cardinality dataset*, $N$ merupakan jumlah dokumen, $Y_i$ adalah himpunan label pada dokumen ke- $i$, sedangkan $|Y_i|$ menunjukkan jumlah label yang dimiliki oleh dokumen ke- $i$. Dengan demikian, nilai Label *Cardinality* merepresentasikan rata-rata banyaknya label yang melekat pada setiap dokumen dalam dataset.

Nilai Label *Cardinality* yang tinggi menunjukkan bahwa setiap dokumen cenderung memiliki lebih banyak label, sehingga tingkat kompleksitas klasifikasi menjadi lebih tinggi. Sebaliknya, nilai yang rendah menunjukkan bahwa sebagian besar dokumen hanya memiliki sedikit label sehingga proses klasifikasi relatif lebih sederhana.

### 2.2.5. Label *Density*

Label *Density* merupakan normalisasi dari Label *Cardinality* terhadap jumlah label yang tersedia pada *dataset*. Metrik ini menggambarkan proporsi label yang dimiliki setiap dokumen dibandingkan dengan keseluruhan label yang mungkin muncul. Nilai Label *Density* digunakan untuk mengetahui tingkat kepadatan label dalam *dataset* sehingga memberikan gambaran mengenai kompleksitas distribusi label. Semakin tinggi nilai Label *Density*, semakin besar proporsi label yang dimiliki setiap dokumen terhadap keseluruhan label yang tersedia [16], [17], [18]. Pada penelitian ini, Label *Density* digunakan untuk mengevaluasi kepadatan label pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia.

Persamaan (2) menunjukkan bahwa nilai Label *Density* diperoleh dengan menormalisasi jumlah label pada setiap dokumen terhadap jumlah keseluruhan label yang tersedia dalam *dataset*. Pada persamaan tersebut, $LD\ (D)$ menyatakan nilai Label *Density dataset*, $N$ merupakan jumlah dokumen, $Y_i$ adalah himpunan label pada dokumen ke- $i$, sedangkan $|L|$ menunjukkan jumlah seluruh label yang tersedia dalam *dataset*. Dengan demikian, nilai Label *Density* merepresentasikan proporsi rata-rata label yang dimiliki setiap dokumen terhadap keseluruhan label yang mungkin muncul.

Nilai Label *Density* yang tinggi menunjukkan bahwa setiap dokumen cenderung memiliki proporsi label yang lebih besar terhadap jumlah label yang tersedia, sehingga hubungan antarlabel dalam *dataset* relatif lebih padat. Sebaliknya, nilai yang rendah menunjukkan bahwa setiap dokumen hanya memiliki sebagian kecil dari keseluruhan label yang tersedia sehingga distribusi label cenderung lebih jarang (*sparse*). Oleh karena itu, Label *Density* digunakan sebagai salah satu indikator untuk mengevaluasi tingkat kepadatan label pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia.

### 2.2.6. Label *Correlation*

Label *Correlation* menggambarkan hubungan atau keterkaitan antarlabel yang muncul secara bersamaan pada dokumen dalam *dataset multi-label*. Analisis ini bertujuan untuk

$$C = Y^T Y \quad (3)$$

mengidentifikasi pasangan label yang memiliki kecenderungan muncul secara simultan sehingga dapat memberikan informasi mengenai pola keterkaitan antarlabel. Informasi tersebut bermanfaat dalam memahami struktur dataset serta mendukung pemilihan pendekatan klasifikasi yang mampu memanfaatkan ketergantungan antarlabel [3], [16], [17], [18]. Pada penelitian ini, Label *Correlation* dianalisis menggunakan matriks *co-occurrence* dan divisualisasikan dalam bentuk jaringan hubungan antarlabel.

Matriks *co-occurrence* digunakan untuk merepresentasikan frekuensi kemunculan dua label secara bersamaan pada seluruh dokumen dalam *dataset*. Semakin tinggi nilai *co-occurrence* suatu pasangan label, semakin kuat hubungan antarlabel yang terbentuk sehingga pola keterkaitan antarlabel dapat diidentifikasi secara lebih jelas [3], [16], [17], [18]. Hubungan tersebut dihitung menggunakan Persamaan (3).

Persamaan (3) menunjukkan bahwa matriks *co-occurrence* ($C$) diperoleh melalui perkalian transpos matriks label biner ($Y^T$) dengan matriks label biner ($Y$). Pada persamaan tersebut, $Y$ merupakan matriks biner berukuran $N \times L$, dengan $N$ menyatakan jumlah dokumen dan $L$ menyatakan jumlah label. Setiap elemen pada matriks $C$ menunjukkan frekuensi kemunculan bersama (*co-occurrence frequency*) antara dua label dalam seluruh *dataset*. Semakin besar nilai *co-occurrence*, semakin kuat hubungan antarlabel sehingga pola keterkaitan antarlabel dapat diidentifikasi secara lebih jelas.

### 2.2.7. Kompleksitas *Multi-Label*

Kompleksitas *multi-label* menggambarkan tingkat kesulitan suatu *dataset* berdasarkan kombinasi berbagai karakteristik, seperti distribusi label, Label *Cardinality*, Label *Density*, karakteristik dokumen, dan hubungan antarlabel. Semakin kompleks karakteristik *dataset*, semakin besar tantangan yang dihadapi model dalam mempelajari pola klasifikasi secara akurat [3], [16], [17], [18]. Oleh karena itu, analisis kompleksitas *multi-label* tidak hanya memberikan gambaran mengenai tingkat kesulitan *dataset*, tetapi juga menjadi dasar dalam menentukan strategi pengembangan model yang sesuai.

**Tabel 1.** Contoh Data pada Dataset Terjemahan Al-Qur'an Bahasa Indonesia.

| Surah | Ayat | Terjemahan | Label |
|---|---|---|---|
| Al-Ma'idah | 26 | Allah berfirman: "(Jika demikian), maka sesungguhnya negeri itu diharamkan atas mereka selama empat puluh tahun, (selama itu) mereka akan berputar-putar kebingungan di bumi (padang Tiih) itu. Maka janganlah kamu bersedih hati (memikirkan nasib) orang-orang yang fasik itu". | Akhlaq, Sejarah |
| Al-An'am | 159 | Sesungguhnya orang-orang yang memecah belah agama-Nya dan mereka menjadi bergolongan, tidak ada sedikitpun tanggung jawabmu kepada mereka. Sesungguhnya urusan mereka hanyalah terserah kepada Allah, kemudian Allah akan memberitahukan kepada mereka apa yang telah mereka perbuat. | Tauhid, Akhlaq |

Pada penelitian ini, kompleksitas *multi-label* diinterpretasikan berdasarkan hasil keseluruhan analisis karakteristik *dataset* yang diperoleh sehingga memberikan gambaran menyeluruh mengenai sifat intrinsik *dataset* terjemahan Al-Qur'an Bahasa Indonesia.

## 3. METODE PENELITIAN

### 3.1. *Dataset*

Pada penelitian ini digunakan *dataset* berupa teks terjemahan Al-Qur'an Bahasa Indonesia yang bersumber dari terjemahan resmi

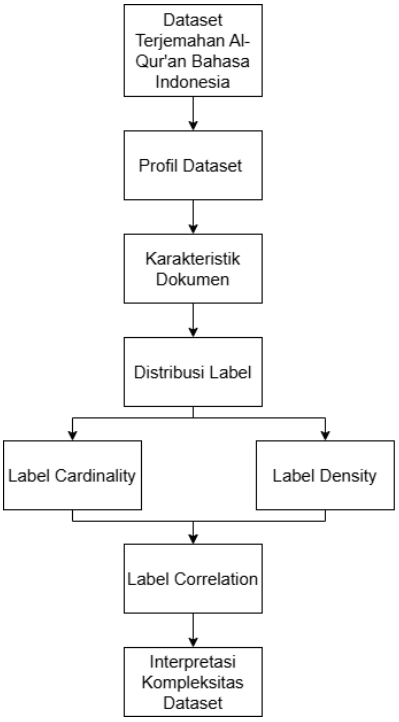
**Tabel 2.** Label dalam klasifikasi terjemahan Al-Quran.

| Label | Keterangan |
|---|---|
| Tauhid | Menggambarkan teks tersebut menyatakan tentang keesaan Allah Subhanahu Wa Ta'ala |
| Ibadah | Menggambarkan teks yang mempunyai makna tentang tunduk dan patuh dalam melaksanakan dan menajuhi segala larangan Allah Subhanahu Wa Ta'ala. |
| Akhlaq | Menggambarkan teks yang menyatakan budi pekerti dan sifat yang dimiliki seseorang yang melahirkan perbuatan baik dan buruk. |
| Sejarah / Tarikh | Menggambarkan teks yang mengandung kisah ataupun peristiwa masa lampau yang menceritakan sejarah umat terdahulu untuk diambil pelajaran bagi umat sesudahnya. |

Kementerian Agama Republik Indonesia Tahun 2022 dan dapat diakses melalui situs https://quran.kemenag.go.id. *Dataset* berlabel tersebut sebelumnya digunakan dalam pengembangan model klasifikasi *multi-label* menggunakan *Long Short-Term Memory* (LSTM) serta dikembangkan melalui perbandingan model *Convolutional Neural Network* (CNN), *Bidirectional Long Short-Term Memory* (BiLSTM), dan *FastText* [10], [11]. Tabel 1 menyajikan sebagian data sebagai ilustrasi struktur *dataset*.

Secara keseluruhan, *dataset* terdiri atas 6.236 ayat yang berasal dari 114 surah. Analisis *multi-label* pada penelitian ini menggunakan 461 ayat berlabel hasil anotasi manual oleh pakar yang terdiri atas empat label tematik, yaitu Tauhid, Ibadah, Akhlaq, dan Sejarah/Tarikh, sebagaimana ditunjukkan pada Tabel 2. Setiap ayat dapat memiliki satu atau lebih label sehingga membentuk karakteristik *multi-label*

*Dataset* pada penelitian ini tidak digunakan untuk membangun ataupun mengevaluasi model klasifikasi, melainkan untuk menganalisis karakteristik intrinsiknya melalui analisis profil *dataset*, karakteristik dokumen, distribusi label, Label *Cardinality*, Label *Density*, Label *Correlation*, serta kompleksitas




**Gambar 1.** Tahapan analisis karakteristik *dataset multi-label*.

*multi-label*. Hasil analisis tersebut diharapkan menjadi landasan bagi pengembangan penelitian klasifikasi teks *multi-label* pada masa mendatang.

### 3.2. Tahapan Penelitian

Tahapan analisis *dataset* pada penelitian ini disajikan pada Gambar 1. Analisis diawali dengan penyusunan profil *dataset*, dilanjutkan dengan analisis karakteristik dokumen melalui statistik deskriptif panjang ayat serta distribusi label untuk mengetahui frekuensi kemunculan setiap label [3], [12], [19].

Selanjutnya dilakukan perhitungan Label *Cardinality* dan Label *Density* sebagai indikator karakteristik *multi-label*, kemudian dianalisis Label *Correlation* menggunakan matriks *co-occurrence* berdasarkan Persamaan (3) dan divisualisasikan dalam bentuk *heatmap* serta *network graph*. Tahap terakhir adalah interpretasi kompleksitas *dataset* berdasarkan keseluruhan hasil analisis untuk memperoleh gambaran karakteristik intrinsik *dataset* [16], [17], [18], [19].

Seluruh proses analisis dilakukan menggunakan bahasa pemrograman Python untuk mendukung pengolahan data, perhitungan metrik, dan visualisasi hasil analisis.

### 3.3. Analisis Profil *Dataset*

Analisis profil *dataset* dilakukan untuk memperoleh gambaran umum mengenai karakteristik dasar *dataset* yang digunakan dalam penelitian. Analisis ini mencakup identifikasi jumlah dokumen, jumlah label, jumlah penugasan label (*label assignments*), serta informasi umum lainnya yang berkaitan dengan struktur *dataset*. Informasi tersebut menjadi dasar dalam memahami ruang lingkup *dataset* sebelum dilakukan analisis karakteristik yang lebih spesifik [12], [13], [14].

Pada penelitian ini, analisis profil *dataset* dilakukan secara deskriptif dengan menghitung karakteristik umum dataset terjemahan Al-Qur'an Bahasa Indonesia yang telah melalui proses pelabelan *multi-label*. Hasil analisis profil *dataset* selanjutnya disajikan dalam bentuk tabel statistik deskriptif sebagai dasar untuk analisis karakteristik dokumen, distribusi label, serta kompleksitas *dataset* pada tahap berikutnya.

### 3.4. Analisis Karakteristik Dokumen

Analisis karakteristik dokumen dilakukan untuk menggambarkan variasi representasi teks pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia berdasarkan jumlah kata dan jumlah karakter pada setiap ayat [12], [13], [14].

Analisis menggunakan statistik deskriptif yang meliputi nilai minimum, maksimum, rata-rata (*mean*), median, simpangan baku (*standard deviation*), serta kuartil. Selain itu, distribusi panjang dokumen divisualisasikan menggunakan *histogram* dan *boxplot* untuk mengidentifikasi pola sebaran data serta keberadaan nilai ekstrem (*outlier*).

### 3.5. *Analisis Distribusi Label*

Analisis distribusi label dilakukan untuk mengetahui frekuensi kemunculan setiap label pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia. Analisis ini bertujuan untuk mengidentifikasi pola penyebaran label, termasuk label yang memiliki frekuensi tinggi maupun rendah, sehingga dapat memberikan gambaran mengenai tingkat keseimbangan (*label balance*) atau ketidakseimbangan (*label imbalance*) pada *dataset* [15], [16], [18], [19].

Pada penelitian ini, distribusi label dihitung berdasarkan jumlah kemunculan masing-masing label pada seluruh ayat dalam *dataset*. Frekuensi setiap label kemudian dikonversi ke

$$P(l_i) = \frac{f_i}{\sum_{j=1}^{L} f_j} \, X \, 100\% \quad (4)$$

dalam bentuk persentase terhadap total penugasan label (*label assignments*) untuk mempermudah interpretasi distribusi label. Hasil analisis disajikan dalam bentuk tabel frekuensi dan diagram batang (*bar chart*) sehingga pola distribusi setiap label dapat diamati secara lebih jelas.

Persamaan (4) menunjukkan bahwa persentase distribusi suatu label diperoleh dengan membandingkan frekuensi kemunculan label tersebut terhadap total frekuensi seluruh label dalam *dataset*. Pada persamaan tersebut, $P(l_i)$ menyatakan persentase label ke- $i$, $f_i$ merupakan frekuensi kemunculan label ke- $i$, sedangkan $\sum_{j=1}^{L} f_j$ merupakan total frekuensi seluruh label pada *dataset* dengan $L$ sebagai jumlah label yang dianalisis. Nilai persentase tersebut digunakan untuk menggambarkan proporsi kemunculan setiap label sehingga tingkat keseimbangan maupun ketidakseimbangan distribusi label dapat diinterpretasikan secara lebih jelas.

### 3.6. *Analisis Label Cardinality*

Analisis Label *Cardinality* dilakukan untuk mengetahui rata-rata jumlah label yang dimiliki oleh setiap ayat dalam *dataset* terjemahan Al-Qur'an Bahasa Indonesia. Nilai Label *Cardinality* digunakan sebagai salah satu indikator untuk menggambarkan tingkat kompleksitas *dataset multi-label*, karena menunjukkan banyaknya label yang secara umum melekat pada setiap dokumen [16], [17], [18], [19].

Pada penelitian ini, nilai Label *Cardinality* dihitung menggunakan Persamaan (1) pada seluruh *dataset*. Perhitungan dilakukan berdasarkan jumlah label yang dimiliki setiap ayat, kemudian dirata-ratakan terhadap seluruh dokumen dalam *dataset*. Nilai yang diperoleh selanjutnya digunakan sebagai dasar untuk menginterpretasikan tingkat kompleksitas *dataset* dan dibandingkan dengan karakteristik *multi-label* lainnya, seperti Label *Density* dan Label *Correlation*.

### 3.7. *Analisis Label Density*

Analisis Label *Density* dilakukan untuk mengetahui tingkat kepadatan label pada dataset terjemahan Al-Qur'an Bahasa

Indonesia. Nilai Label *Density* digunakan untuk menggambarkan proporsi rata-rata jumlah label yang dimiliki setiap ayat terhadap jumlah keseluruhan label yang tersedia, sehingga dapat memberikan informasi mengenai tingkat kepadatan distribusi label pada *dataset* multi-*label* [16], [17], [18], [19].

Pada penelitian ini, nilai Label *Density* dihitung menggunakan Persamaan (2) pada seluruh *dataset*. Perhitungan dilakukan dengan menormalisasi jumlah label yang dimiliki setiap ayat terhadap jumlah keseluruhan label, kemudian dirata-ratakan pada seluruh dokumen dalam *dataset*. Nilai yang diperoleh selanjutnya digunakan untuk melengkapi interpretasi karakteristik multi-label bersama dengan hasil analisis Label *Cardinality* dan Label *Correlation*.

### 3.8. *Analisis Label Correlation*

Analisis Label *Correlation* dilakukan untuk mengidentifikasi hubungan antarlabel yang muncul secara bersamaan pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia. Analisis ini bertujuan untuk mengetahui pola keterkaitan antarlabel sehingga dapat memberikan gambaran mengenai struktur hubungan label dalam *dataset multi-label* [3], [16], [17], [18], [19].

Pada penelitian ini, Label *Correlation* dianalisis menggunakan matriks *co-occurrence* yang dibentuk berdasarkan Persamaan (3) dengan menghitung frekuensi kemunculan setiap pasangan label pada seluruh ayat dalam *dataset*. Selanjutnya, hasil perhitungan divisualisasikan dalam bentuk *heatmap* dan jaringan hubungan antarlabel (*network graph*) untuk mempermudah interpretasi pola keterkaitan antarlabel. Visualisasi tersebut digunakan untuk mengidentifikasi pasangan label yang memiliki hubungan kuat maupun lemah sehingga dapat mendukung interpretasi karakteristik intrinsik *dataset*.

### 3.9. *Analisis Kompleksitas Dataset Multi-Label*

Analisis kompleksitas *dataset multi-label* dilakukan untuk memperoleh gambaran menyeluruh mengenai tingkat kompleksitas dataset terjemahan Al-Qur'an Bahasa Indonesia berdasarkan hasil analisis pada tahap sebelumnya. Interpretasi kompleksitas tidak ditentukan oleh satu metrik tertentu, melainkan melalui integrasi karakteristik *dataset* yang meliputi profil *dataset*, karakteristik dokumen, distribusi label, Label *Cardinality*, Label *Density*, serta Label *Correlation* [3], [16], [17], [18], [19].

**Tabel 3.** Profil *Dataset* Terjemahan Al-Qur'an Bahasa Indonesia.

| Karakteristik | Nilai |
|---|---|
| Jumlah Surah | 114 |
| Jumlah Ayat (Dokumen) | 6.236 |
| Jumlah Label | 4 |
| Jenis Dataset | Multi-label |
| Bahasa | Indonesia |
| Sumber *Dataset* | Kementerian Agama RI (2022) |

Pada penelitian ini, hasil dari setiap tahapan analisis diinterpretasikan secara komprehensif untuk mengidentifikasi karakteristik intrinsik dataset yang berpotensi memengaruhi proses klasifikasi *multi-label*. Interpretasi tersebut digunakan untuk memberikan pemahaman mengenai tingkat kompleksitas *dataset* serta menjadi dasar dalam memberikan rekomendasi terhadap pengembangan metode klasifikasi teks *multi-label* pada penelitian selanjutnya.

## 4. HASIL DAN PEMBAHASAN

### 4.1. *Profil Dataset*

Profil *dataset* pada Tabel 3 dianalisis untuk memberikan gambaran umum mengenai karakteristik dasar dataset terjemahan Al-Qur'an Bahasa Indonesia yang digunakan pada penelitian ini. Analisis meliputi jumlah dokumen, jumlah surah, jumlah label, jumlah penugasan label (*label assignments*), serta karakteristik umum lainnya yang menjadi dasar analisis pada tahap berikutnya.

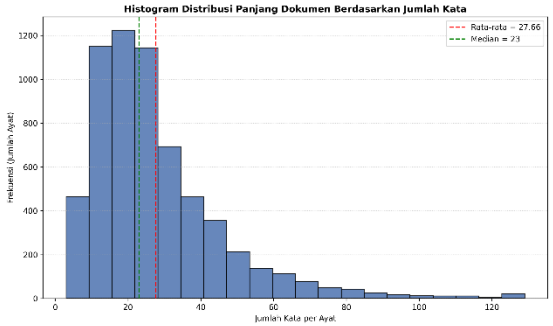
*Dataset* penelitian terdiri atas 6.236 ayat yang berasal dari 114 surah. Analisis karakteristik *multi-label* dilakukan pada subset *dataset* berlabel yang terdiri atas 461 ayat dengan empat label tematik, yaitu Tauhid, Ibadah, Akhlaq, dan Sejarah. Setiap ayat pada subset tersebut dapat memiliki satu atau lebih label sehingga membentuk karakteristik *multi-label*.

### 4.2. *Karakteristik Dokumen*

Karakteristik dokumen dianalisis berdasarkan panjang dokumen yang diukur

**Tabel 4.** Statistik Karakteristik Dokumen.

| Statistik | Jumlah Kata | Jumlah Karakter |
|---|---|---|
| Minimum | 3 | 18 |
| Maksimum | 129 | 798 |
| Rata-rata | 27,84 | 167,35 |
| Median | 23 | 141 |
| Standar Deviasi | 18,92 | 111,47 |




***Gambar 2.*** *Histogram Distribusi Panjang Dokumen Berdasarkan Jumlah Kata.*

melalui jumlah kata dan jumlah karakter pada setiap ayat. Berdasarkan Tabel 4, panjang dokumen pada *dataset* bervariasi antara 3–129 kata dan 18–798 karakter, dengan rata-rata 27,84 kata (median 23 kata) serta 167,35 karakter (median 141 karakter). Perbedaan antara nilai rata-rata dan median, serta nilai standar deviasi yang relatif tinggi, menunjukkan adanya variasi panjang dokumen dalam *dataset*.

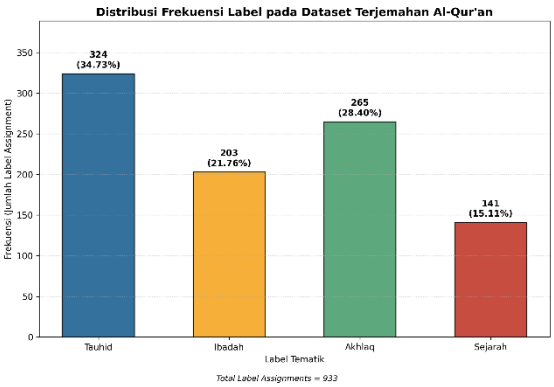
Berdasarkan Gambar 2, distribusi panjang dokumen cenderung membentuk pola yang menceng ke kanan (*positively skewed*), ditunjukkan oleh nilai rata-rata yang lebih besar dibandingkan median. Sebagian besar ayat memiliki panjang relatif pendek hingga sedang, sedangkan hanya sebagian kecil ayat yang memiliki jumlah kata lebih banyak. Karakteristik tersebut menunjukkan bahwa *dataset* memiliki variasi representasi teks yang cukup baik sehingga dapat merepresentasikan kompleksitas dokumen dalam analisis klasifikasi *multi-label*.

### 4.3. Distribusi *Label*

Distribusi label dianalisis untuk mengetahui frekuensi kemunculan setiap label pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia. Analisis ini bertujuan mengidentifikasi

**Tabel 5**. Hasil Perhitungan Label Cardinality.

| Parameter | Nilai |
|---|---|
| Jumlah Dokumen | 461 |
| Jumlah Label Assignments | 933 |
| Label Cardinality | **2,02** |




**Gambar 3**. Distribusi Frekuensi Label pada Dataset.

karakteristik distribusi label serta tingkat keseimbangan label pada *dataset multi-label*. Hasil analisis ini menjadi dasar dalam perhitungan Label *Cardinality*, Label *Density*, dan Label *Correlation*.

Berdasarkan Gambar 3, label Tauhid memiliki frekuensi kemunculan tertinggi sebanyak 324 (34,73%), diikuti Akhlaq sebanyak 265 (28,40%), Ibadah sebanyak 203 (21,76%), dan Sejarah sebanyak 141 (15,11%). Perbedaan frekuensi tersebut menunjukkan bahwa distribusi label tidak bersifat seragam, namun masih berada pada tingkat yang relatif seimbang sehingga tidak menunjukkan kondisi label *imbalance* yang ekstrem. Karakteristik ini menunjukkan bahwa setiap label memiliki representasi yang memadai untuk dianalisis serta menjadi dasar dalam analisis Label *Cardinality*, Label *Density*, dan Label *Correlation* pada subbab berikutnya.

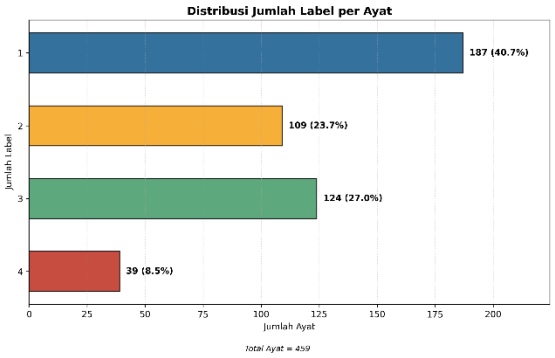
### 4.4. Label *Cardinality*

Label *Cardinality* dianalisis untuk mengetahui rata-rata jumlah label yang dimiliki oleh setiap ayat pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia. Metrik ini digunakan untuk menggambarkan tingkat kompleksitas dataset *multi-label* berdasarkan banyaknya label yang melekat pada setiap dokumen.

Berdasarkan Tabel 5, diperoleh nilai Label *Cardinality* sebesar 2,02, yang menunjukkan bahwa setiap ayat memiliki rata-rata sekitar dua label. Hasil tersebut mengindikasikan bahwa

**Tabel 6**. Hasil Perhitungan Label *Density*.

| Parameter | Nilai |
|---|---|
| Jumlah Dokumen | 461 |
| Jumlah Label | 4 |
| Label *Cardinality* | 2,02 |
| Label *Density* | **0,506** |




**Gambar 4.** Distribusi Jumlah Label per Ayat.

sebagian besar dokumen merepresentasikan lebih dari satu tema sehingga mencerminkan karakteristik *multi-label*.

Berdasarkan Gambar 4, sebagian besar ayat memiliki satu hingga tiga label, sedangkan hanya sebagian kecil ayat memiliki empat label secara bersamaan. Distribusi tersebut mendukung nilai Label *Cardinality* sebesar 2,02, yang menunjukkan bahwa *dataset* memiliki keterkaitan antartema pada tingkat sedang. Karakteristik ini menjadi dasar dalam analisis Label *Density* pada subbab berikutnya.

### 4.5. *Label Density*

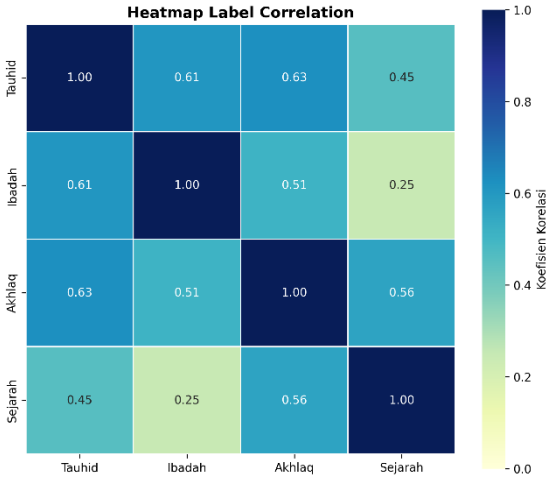
Label *Density* dianalisis untuk mengetahui tingkat kepadatan label pada *dataset multi-label*. Metrik ini menunjukkan proporsi rata-rata jumlah label yang dimiliki setiap dokumen terhadap jumlah keseluruhan label yang tersedia.

Berdasarkan Tabel 6, diperoleh nilai Label *Density* sebesar 0,506, yang menunjukkan bahwa setiap ayat merepresentasikan sekitar 50,6% dari keseluruhan label yang tersedia. Nilai tersebut mengindikasikan bahwa sebagian besar dokumen memiliki lebih dari satu label, sehingga mencerminkan adanya keterkaitan antartema dalam *dataset*.

Nilai Label *Density* tersebut memperkuat hasil analisis Label *Cardinality* dan menunjukkan bahwa *dataset* memiliki kepadatan label pada tingkat sedang. Karakteristik ini menjadi dasar dalam analisis

**Tabel 7.** Matriks *Co-occurrence* Antarlabel.

| | Tauhid | Ibadah | Akhlaq | Sejarah |
|---|---|---|---|---|
| **Tauhid** | 324 | 156 | 184 | 97 |
| **Ibadah** | 156 | 203 | 118 | 42 |
| **Akhlaq** | 184 | 118 | 265 | 109 |
| **Sejarah** | 97 | 42 | 109 | 141 |




**Gambar 5.** Heatmap Label Correlation.

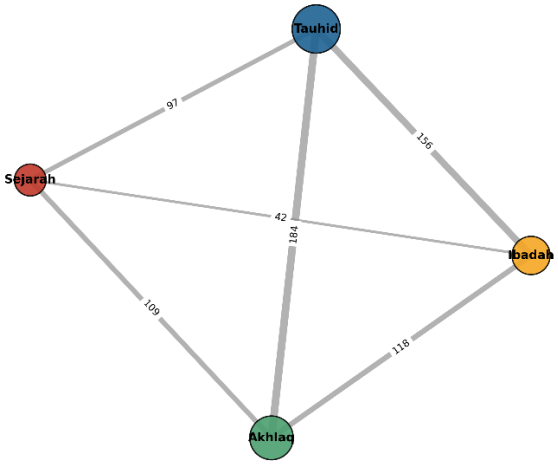
hubungan antarlabel (Label *Correlation*) pada subbab berikutnya.

### 4.6. Label *Correlation*

Label *Correlation* dianalisis untuk mengidentifikasi hubungan kemunculan antarlabel pada *dataset* terjemahan Al-Qur'an Bahasa Indonesia menggunakan matriks *co-occurrence*. Hasil analisis kemudian divisualisasikan dalam bentuk *heatmap* dan *network graph* untuk mempermudah interpretasi pola hubungan antarlabel.

Berdasarkan Tabel 7 dan Gambar 5, setiap pasangan label memiliki tingkat hubungan yang berbeda. Nilai *co-occurrence* yang lebih tinggi menunjukkan kecenderungan dua label lebih sering muncul secara bersamaan, sedangkan nilai yang lebih rendah menunjukkan hubungan yang relatif lemah. Visualisasi heatmap memperjelas intensitas hubungan tersebut melalui perbedaan gradasi warna sehingga pola keterkaitan antarlabel dapat diamati dengan lebih mudah.

Gambar 6 menyajikan hubungan antarlabel dalam bentuk *network graph*, di mana ukuran node menunjukkan frekuensi kemunculan label dan ketebalan *edge* merepresentasikan kekuatan hubungan berdasarkan nilai *co-occurrence*. Visualisasi ini menunjukkan bahwa beberapa




**Gambar 6.** *Network Graph* Hubungan Antarlabel.

pasangan label memiliki hubungan yang lebih kuat dibandingkan pasangan lainnya, sehingga menggambarkan adanya keterkaitan semantik antartema pada dataset. Hasil analisis ini menjadi dasar dalam menginterpretasikan kompleksitas *multi-label* pada subbab berikutnya.

### 4.7. Kompleksitas *Multi-Label*

Kompleksitas *multi-label* dianalisis berdasarkan karakteristik *dataset* yang meliputi distribusi label, karakteristik dokumen, Label *Cardinality*, Label *Density*, dan Label *Correlation*. Analisis ini bertujuan memberikan gambaran menyeluruh mengenai tingkat kompleksitas dataset terjemahan Al-Qur'an Bahasa Indonesia.

Berdasarkan Tabel 8, *dataset* menunjukkan karakteristik *multi-label* yang representatif, ditandai dengan distribusi label yang relatif seimbang, variasi panjang dokumen, nilai Label *Cardinality* sebesar 2,02, Label *Density* sebesar 0,506, serta hubungan antarlabel yang berada pada tingkat sedang hingga kuat. Karakteristik tersebut menunjukkan bahwa dataset memiliki tingkat kompleksitas *multi-label* yang sedang dan layak digunakan sebagai objek penelitian dalam pengembangan maupun evaluasi metode klasifikasi teks *multi-label*.

### 4.8. Implikasi terhadap Pengembangan Model

Hasil analisis karakteristik *dataset* menunjukkan bahwa *dataset* terjemahan Al-Qur'an Bahasa Indonesia memiliki karakteristik yang mendukung pengembangan model klasifikasi teks *multi-label*. Variasi panjang dokumen, distribusi label yang relatif seimbang, nilai Label *Cardinality* dan Label *Density* pada tingkat sedang, serta adanya hubungan antarlabel memberikan gambaran bahwa *dataset* memiliki struktur yang representatif untuk mengevaluasi kemampuan model dalam mengenali lebih dari satu label pada setiap dokumen.

**Tabel 8.** Ringkasan Karakteristik Kompleksitas *Dataset*.

| Karakteristik | Hasil | Interpretasi |
|---|---|---|
| Distribusi Label | Relatif seimbang | Tidak menunjukkan *label imbalance* yang ekstrem |
| Label *Cardinality* | 2,02 | Rata-rata setiap ayat memiliki sekitar dua label |
| Label *Density* | 0,506 | Kepadatan label berada pada tingkat sedang |
| Label *Correlation* | Sedang–Kuat | Terdapat hubungan antartema pada beberapa pasangan label |
| Panjang Dokumen | Bervariasi | Representasi teks cukup beragam |

Selain memberikan informasi mengenai tingkat kompleksitas dataset, hasil analisis ini juga dapat menjadi dasar dalam pemilihan strategi pengembangan model. Informasi mengenai distribusi label dapat digunakan untuk menentukan teknik penanganan label imbalance, sedangkan karakteristik hubungan antarlabel dapat dimanfaatkan dalam pengembangan model yang mempertimbangkan ketergantungan antarlabel (label *dependency*). Dengan demikian, analisis karakteristik *dataset* tidak hanya membantu memahami sifat intrinsik data, tetapi juga mendukung perancangan model klasifikasi *multi-label* yang lebih efektif.

## 5. KESIMPULAN

Penelitian ini berhasil menganalisis karakteristik dataset terjemahan Al-Qur'an Bahasa Indonesia sebagai dasar pengembangan klasifikasi teks *multi-label*. Hasil analisis menunjukkan bahwa *dataset* memiliki distribusi label yang relatif seimbang, nilai Label *Cardinality* sebesar 2,02, Label *Density* sebesar 0,506, serta hubungan antarlabel pada tingkat sedang hingga kuat yang ditunjukkan melalui analisis Label *Correlation*. Selain itu, karakteristik dokumen memperlihatkan variasi panjang teks yang cukup beragam sehingga memberikan representasi data yang memadai untuk analisis *multi-label*. Berdasarkan keseluruhan hasil tersebut, *dataset* memiliki tingkat kompleksitas *multi-label* yang representatif dan layak digunakan sebagai dasar pengembangan maupun evaluasi metode klasifikasi teks *multi-label*.

Kelebihan penelitian ini terletak pada penyajian analisis karakteristik *dataset* secara komprehensif melalui kombinasi analisis profil *dataset*, karakteristik dokumen, distribusi label, Label *Cardinality*, Label *Density*, dan Label *Correlation*. Analisis tersebut memberikan pemahaman yang lebih mendalam mengenai karakteristik intrinsik *dataset* yang selama ini belum banyak dibahas pada penelitian klasifikasi terjemahan Al-Qur'an Bahasa Indonesia. Namun demikian, penelitian ini masih terbatas pada subset *dataset* berlabel yang terdiri atas empat kategori tematik dan belum mengkaji pengaruh karakteristik *dataset* terhadap kinerja berbagai algoritma klasifikasi. Oleh karena itu, penelitian selanjutnya dapat memperluas cakupan *dataset* berlabel, menambah jumlah kategori tematik, serta mengevaluasi hubungan antara karakteristik *dataset* dengan performa model klasifikasi modern, seperti model berbasis *Transformer* maupun *Large Language Models* (LLMs).

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Widya Gama atas dukungan yang diberikan sehingga penelitian ini dapat terlaksana dengan baik.

## DAFTAR PUSTAKA

[1] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," Nov. 01, 2024, *Elsevier Ireland Ltd*. doi: 10.1016/j.cosrev.2024.100664.

[2] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review," *Algorithms*, vol. 16, no. 5, May 2023, doi: 10.3390/a16050236.

[3] S. Huang *et al.*, "Application of Label Correlation in Multi-Label Classification: A Survey," Oct. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app14199034.

[4] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021, doi: 10.1145/3439726.

[5] J. Fields, K. Chovanec, and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," *IEEE Access*, vol. 12, pp. 6518–6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

[6] D. R. Putri, E. Y. Puspaningrum, and H. Maulana, "Klasifikasi Sentimen Tentang Pemindahan Ibu Kota Negara Indonesia Dengan Convolutional Neural Network Menggunakan Glove Dan Fasttext," *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024, doi: https://doi.org/10.23960/jitet.v12i3.4882.

[7] M. H. Bashir *et al.*, "Arabic natural language processing for Qur'anic research: a systematic review," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6801–6854, Jul. 2023, doi: 10.1007/s10462-022-10313-2.

[8] A. Adeleke, N. Samsudin, A. Mustapha, and S. Ahmad Khalid, "Automating quranic verses labeling using machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 925–931, 2019, doi: 10.11591/ijeecs.v16.i2.pp925-931.

[9] A. Tubaishat, S. Razzaq, F. Maqbool, M. Ilyas, and M. S. Khan, "Quran Wordnet: A Framework for Semantic and Sentiment Search," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 2023, pp. 24–32. doi: 10.1109/CiSt56084.2023.10409881.

[10] I. Akbar, M. Faisal, and T. Chamidy, "Multi-Label Classification of Indonesian Qur'an Translation using Long Short-Term Memory Model," *Computer Network, Computing, Electronics, and Control Journal*, vol. 4, no. 3, pp. 119–128, 2019, [Online]. Available:

https://kinetik.umm.ac.id/index.php/kinetik/article/view/1901

[11] A. Rofiqul Muslikh, I. Akbar, D. Rosal Ignatius Moses Setiadi, and H. Md Mehedul Islam, "Multi-label Classification of Indonesian Al-Quran Translation based CNN, BiLSTM, and FastText," 2024. doi: https://doi.org/10.62411/tc.v23i1.9925.

[12] Y. Gong, G. Liu, Y. Xue, R. Li, and L. Meng, "A survey on dataset quality in machine learning," *Inf. Softw. Technol.*, vol. 162, p. 107268, 2023, doi: https://doi.org/10.1016/j.infsof.2023.107268.

[13] P. Singh, "Systematic review of data-centric approaches in artificial intelligence and machine learning," *Data Science and Management*, vol. 6, no. 3, pp. 144–157, 2023, doi: https://doi.org/10.1016/j.dsm.2023.06.001.

[14] M. Priestley, F. O'donnell, and E. Simperl, "A Survey of Data Quality Requirements That Matter in ML Development Pipelines," *J. Data and Information Quality*, vol. 15, no. 2, Jun. 2023, doi: 10.1145/3592616.

[15] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 118, p. 107965, 2021, doi: https://doi.org/10.1016/j.patcog.2021.107965.

[16] N. E. García-Pedrajas, J. M. Cuevas-Muñoz, G. Cerruela-García, and A. de Haro-García, "A thorough experimental comparison of multilabel methods for classification performance," Jul. 01, 2024, *Elsevier Ltd*. doi: 10.1016/j.patcog.2024.110342.

[17] W. Hu, Q. Fan, H. Yan, X. Xu, S. Huang, and K. Zhang, "A Survey of Multi-Label Text Classification Under Few-Shot Scenarios," Aug. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15168872.

[18] N. Li, B. Kang, and T. De Bie, "Your Next State-of-the-Art Could Come from Another Domain: A Cross-Domain Analysis of Hierarchical Text Classification," *Mach. Learn.*, vol. 115, no. 4, Apr. 2026, doi: 10.1007/s10994-025-06993-w.

[19] Y. Zou, X. Hu, P. Li, Y. Wu, and J. Hu, "Online multi-label classification under noisy and changing label distribution," *Pattern Recogn.*, vol. 173, no. C, Jun. 2026, doi: 10.1016/j.patcog.2025.112892.