Optimasi Prediksi Vonis Pidana Melalui Strategi Ekstraksi Fitur Hibrida

Rizky Fadhilah
Vincentius Jacob Gunawan
Josh Edward Sutanto

LATAR BELAKANG

Memprediksi durasi hukuman hanya dari narasi teks putusan.

Bukan untuk menggantikan peran hakim, melainkan untuk menyediakan referensi data-driven demi mendukung terciptanya **keadilan yang konsisten** di seluruh Indonesia.

DATASET

Dataset berisikan folder berisi keputusan hakim beserta label hukuman yang dijatuhkan ke tersangka sebanyak berapa bulan

id	lama hukuman (bulan)
doc_13590	10
doc_14914	60
doc_21900	18
doc_14859	72

* **Data Latih**
  16572 Baris

* **Data Uji**
  6666 Baris

* **Fitur**
  Teks / dokumen

* **Target**
  Lama hukuman (bulan)

Tantangan sebenarnya bukan pada volume data yang besar

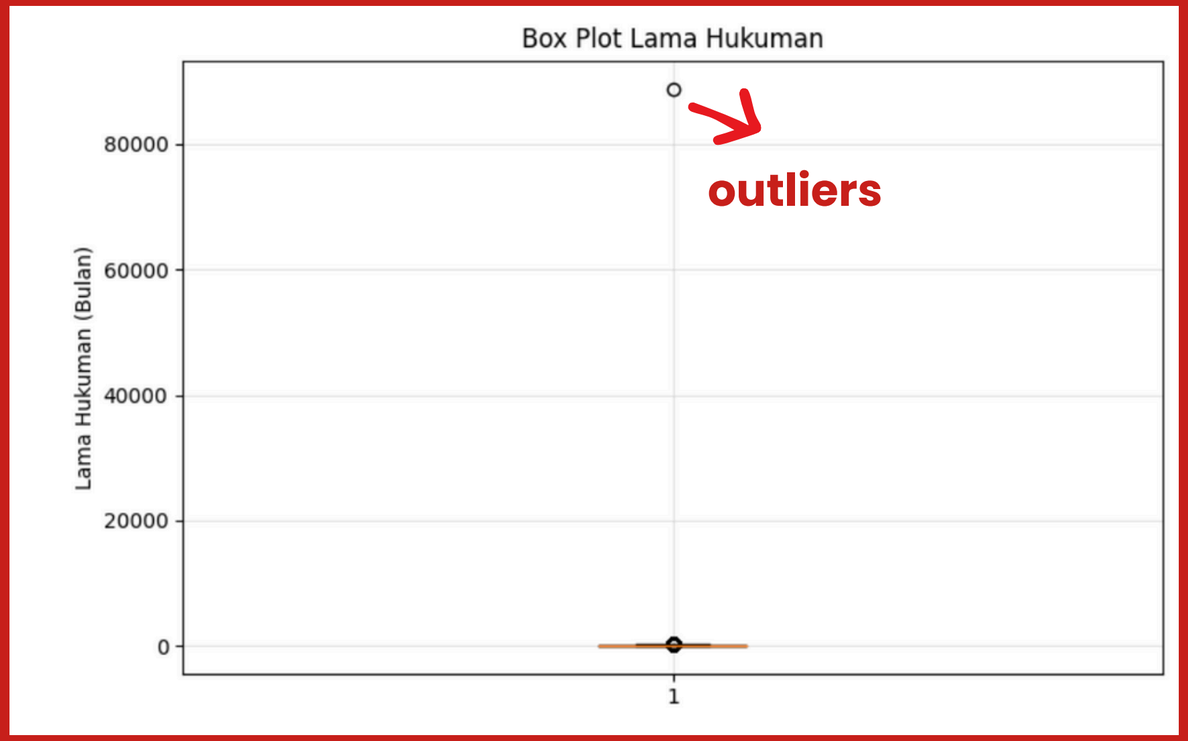
Menerjemahkan teks yang penuh ambiguitas

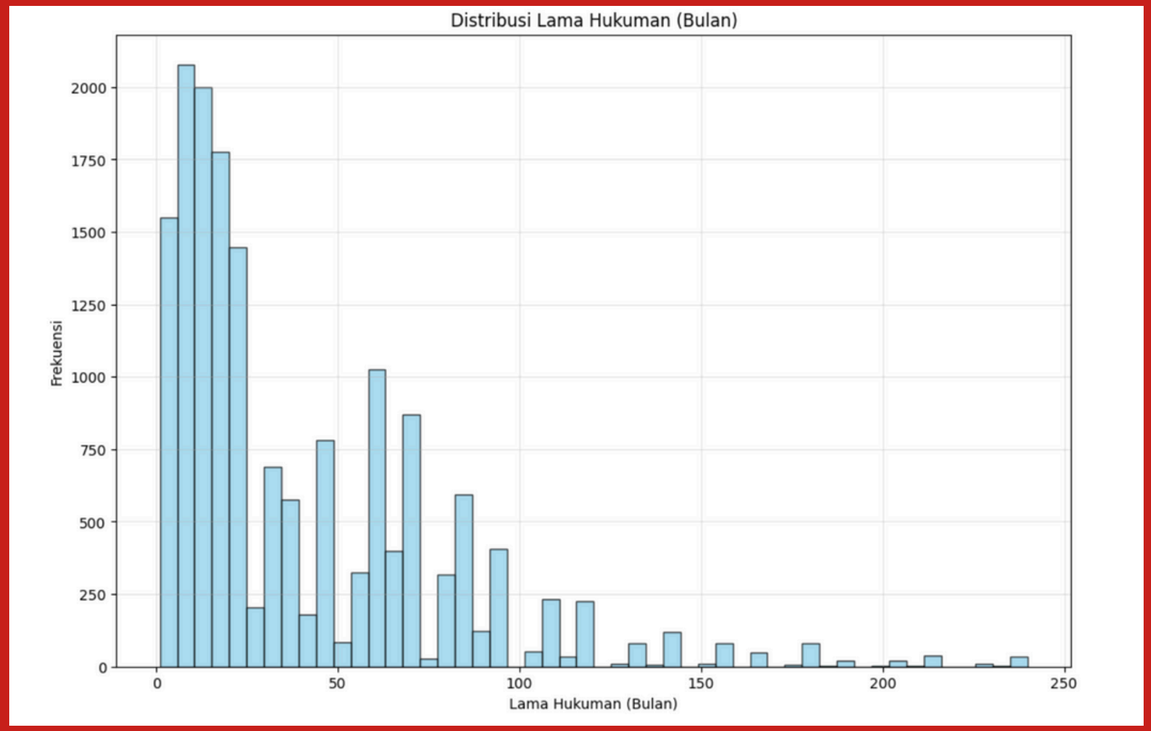
Representasi angka yang objektif

ANALISIS DATA EKSPLORATIF

1. Banyak kesalahan pengetikan
2. Inkonsistensi spasi & kapitalisasi massal.
3. Ribuan tanda baca & simbol tak perlu

ANALISIS DATA EKSPLORATIF





PEMBERSIHAN DATA

Pengumpulan & Penggabungan Data

Ribuan File Putusan

↓

Proses Pengelompokan

↓

Dataset Gabungan (Masih 'Kotor')

Fungsi Pembersihan Teks

Aa → aa

Mengubah ke huruf kecil (lowercase)

@!# →

Menghapus semua tanda baca & simbol

pasal 362 ayat 1 kuhp → **pasal 362 ayat 1 kuhp**

Normalisasi spasi berlebih

HASIL PEMBERSIHAN DATA

...hikama...

...tahan tahan than...

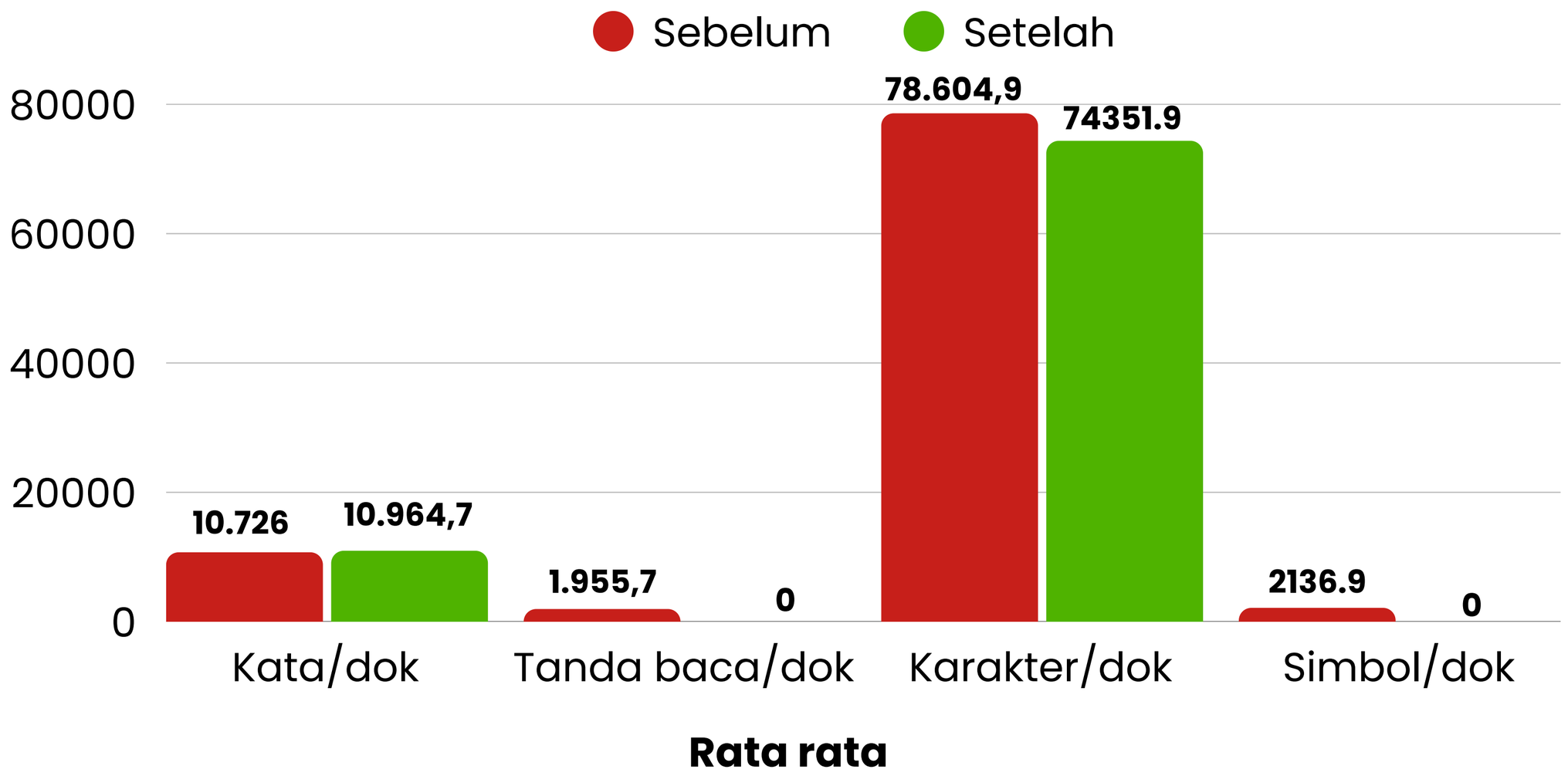
...P U T U S A N...

...kepaniteraanmahka
mahagunggoidte...

...umurtanggal....

Dataset masih kotor!

PEMBERSIHAN DATA



STRATEGI EKSTRAKSI FITUR HIBRIDA

- Analisis Kata & Frasa

1. **TF-IDF + N-Gram Kata (1-2)**
2. **TF-IDF + N-Gram Kata (1-3)**
3. **TF-IDF + N-Gram Kata (2-3)**

STRATEGI EKSTRAKSI FITUR HIBRIDA

- Analisis Kata & Frasa

“saksi ahli pidana”

[‘saksi’] [‘ahli’] [‘pidana’]

[‘saksi ahli’] [‘ahli pidana’]

STRATEGI EKSTRAKSI FITUR HIBRIDA

- Analisis Karakter

1. **TF-IDF + N-Gram Karakter (2-4)**

2. **TF-IDF + N-Gram Karakter (3-5)**

STRATEGI EKSTRAKSI FITUR HIBRIDA

- Analisis Karakter

“mahkamah”

['mah', 'ahk', 'hka', 'kam', 'ama']

dan puluhan pola lainnya dari seluruh frasa.

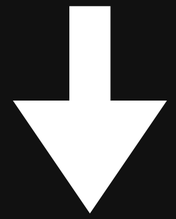
STRATEGI EKSTRAKSI FITUR HIBRIDA

- Analisis Frekuensi

1. **CountVectorizer + N-Gram Karakter (1,2)**

- Analisis Frekuensi

“saksi ahli pidana”

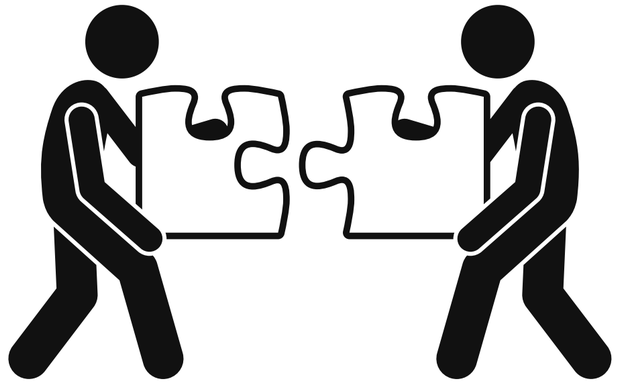


Teks Narasi	saksi	ahli	pidana	saksi ahli	ahli pidana
...menurut keterangan saksi ahli terdakwa terbukti melakukan tindak pidana sesuai pasal 362...	2	1	2	1	1
...penuntut umum telah menghadirkan seorang saksi ahli yang di bawah sumpah keterangannya memberatkan terdakwa...	1	1	1	1	1
...saksi ahli pidana menjelaskan bahwa alat bukti yang digunakan terdakwa bukanlah benda yang lazimnya dibawa sehari-hari...	1	2	1	1	1
...saksi mengatakan bahwa saudara budi mencoba membawa kabur sebuah kendaraan roda empat berwarna hitam...	1	1	1	0	0

Super Vektor → Satu vektor gabungan berisi ribuan fitur

- **Penggabungan**

3 Tahap Analisis Kata & Frasa + 2 Tahap Analisis Karakter + CountVectorizer



Super Vektor → Satu vektor gabungan berisi ribuan fitur

Teks Narasi	saksi	ahli	pidana	saksi ahli	ahli pidana
...menurut keterangan saksi ahli terdakwa terbukti melakukan tindak pidana sesuai pasal 362...	0.12	0.35	0.51	0.41	0.03
...penuntut umum telah menghadirkan seorang saksi ahli yang di bawah sumpah keterangannya memberatkan terdakwa...	0.023	0.58	0.15	0.68	0.12
...saksi ahli pidana menjelaskan bahwa alat bukti yang digunakan terdakwa bukanlah benda yang lazimnya dibawa sehari-hari...	0.07	0.29	0.42	0.34	0.76
...saksi mengatakan bahwa saudara budi mencoba membawa kabur sebuah kendaraan roda empat berwarna hitam...	0.13	0.11	0.05	0.32	0.07

Mengapa Pendekatan ***Hybrid*** Ini **Unggul?**

Menangkap *keyword* penting dari teks yang kompleks

Mampu menghadapi dataset yang kotor

Sumber daya komputasi minimum

Tidak hanya mengubah **teks** menjadi **angka**

tetapi

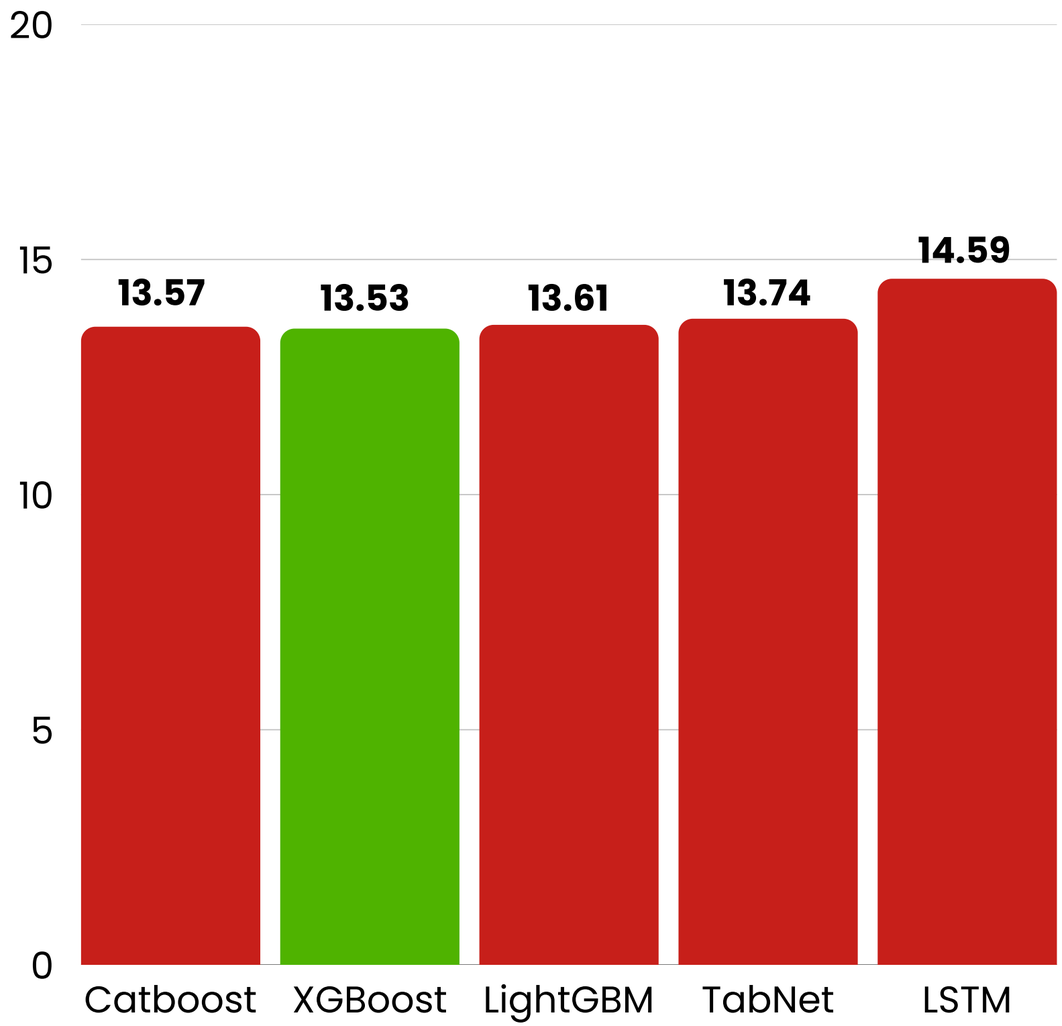
Narasi mentah yang ambigu  Fitur super kaya yang 'bercerita'.

KOMPARASI MODEL TERBAIK

Metriks Evaluasi:

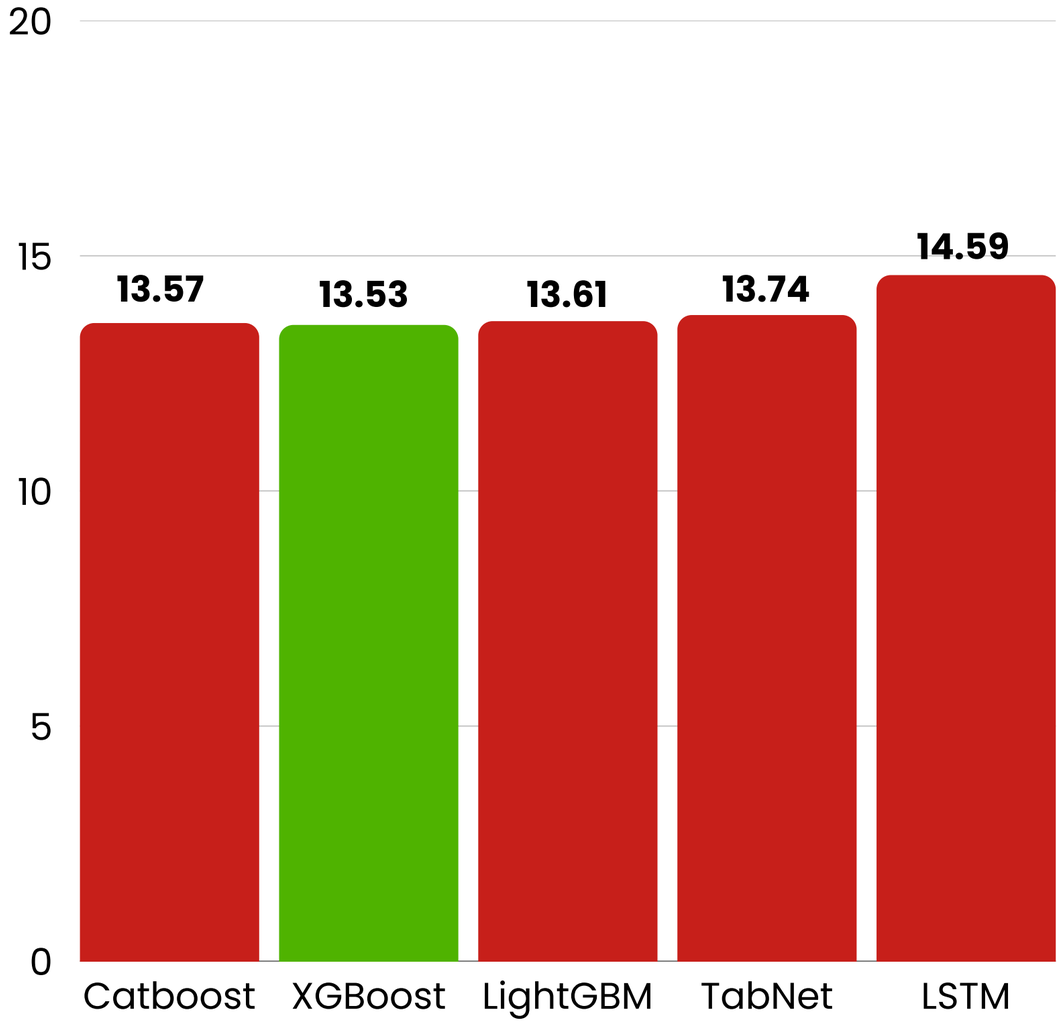
Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

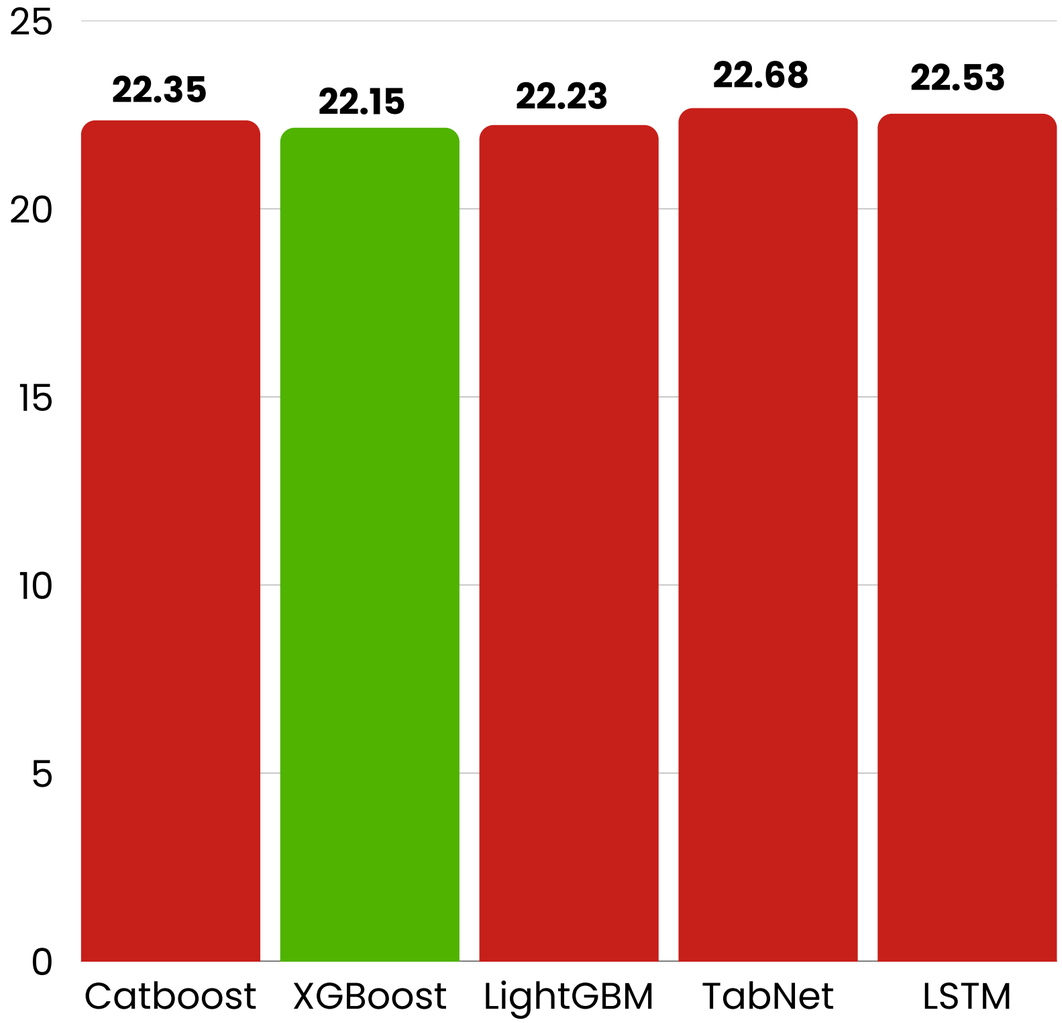


Komparasi Pendekatan *Feature Extraction*

Pendekatan Usulan



IndoBERT



KESIMPULAN

- **Frasa-frasa kunci** (seperti pasal, jenis tindak pidana, keterangan saksi) **sangat berpengaruh terhadap vonis hukuman**, sehingga kami merancang strategi untuk mengisolasi dan memperkuat sinyal-sinyal penting ini.

- **Optimalisasi pengaturan N-gram:**
  - **N-gram Kata**: Mengubah frasa krusial seperti 'pasal 363 ayat 1' menjadi satu fitur tunggal yang kuat.
  - **N-gram Karakter**: Memberikan 'kekebalan' terhadap typo, singkatan, dan variasi ejaan, memastikan tidak ada informasi yang hilang

- Pentingnya metode pra-pemrosesan yang tepat agar hasil model memuaskan

- *State-of-the-Art* seperti IndoBERT tidak selalu menjadi jawaban terbaik

Terima kasih.*