

# DOSC

## DATA SCIENCE COMPETITION



# PENDAHULUAN



**Kelemahan sistem referensi perpustakaan menghambat akses pengetahuan.**



**Hambatan** dalam **sistem referensi** menyebabkan kesulitan dalam **mengakses** dan **memilih referensi yang sesuai** untuk karya ilmiah.



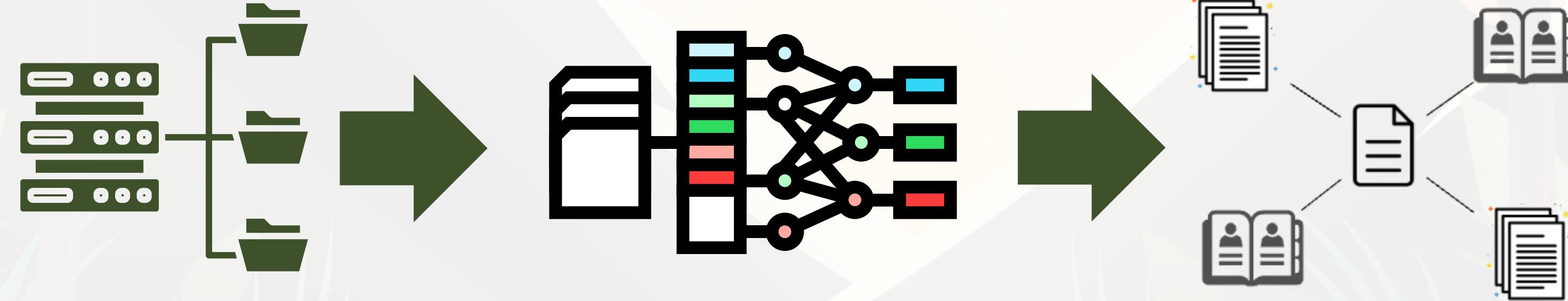
**Kesulitan** dalam **mengakses** dan **memilih** referensi yang sesuai untuk **karya ilmiah** akan **menghambat** proses **penelitian** dan **mengurangi** **kualitas** hasil **penelitian**.

## Lalu, apa yang dapat dilakukan?



# OBJEKTIF

**Mengoptimalkan sistem referensi untuk mempercepat akses informasi dan mendukung kualitas penelitian.**

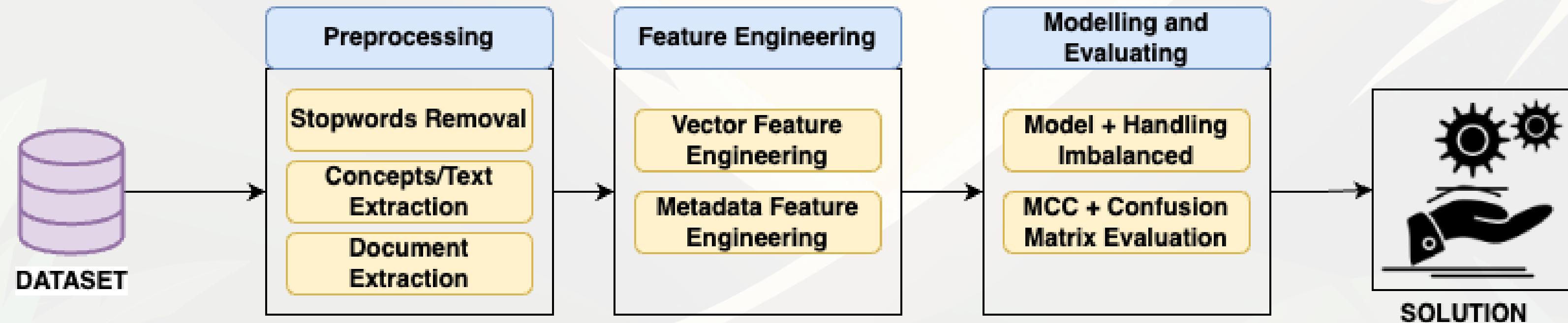


Membuat **perencanaan** untuk meningkatkan efisiensi akses informasi dan **mendukung kualitas penelitian**.

**Mengoptimalkan sistem** referensi melalui penerapan **model machine learning** untuk prediksi **hubungan** antar **paper**

**Menerapkan solusi** sistem rekomendasi referensi guna meningkatkan **efektivitas** dan **akurasi** pencarian literatur.

# ALUR PENELITIAN



# DATA PREPROCESSING



PT AYASKARA NISITA SYNERGY  
Market Research and Management Consultants

starcore.co

## Document

PAPER ID	X
<p>Department of Medicine and Institute for Human Genetics, University of California, San Francisco, and California Institute for Quantitative Biosciences, San Francisco, CA t from such learning approaches and use examples from the literature to introduce basic concepts in machine learning. It is important to note.</p>	

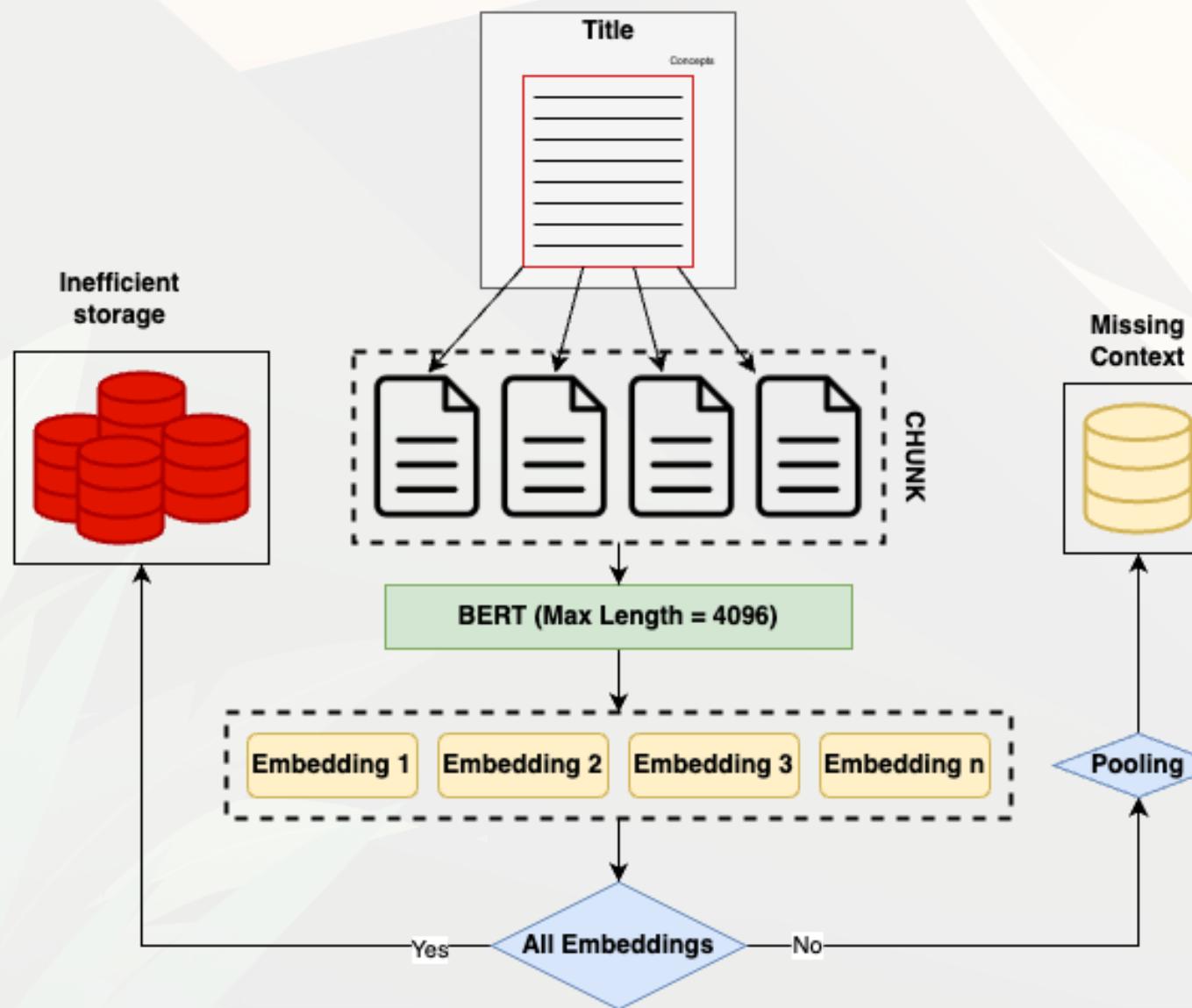
Dokumen berupa teks akademik yang secara mendalam mengulas suatu topik tertentu dalam bentuk pembahasan yang panjang dan terstruktur.

## Metadata

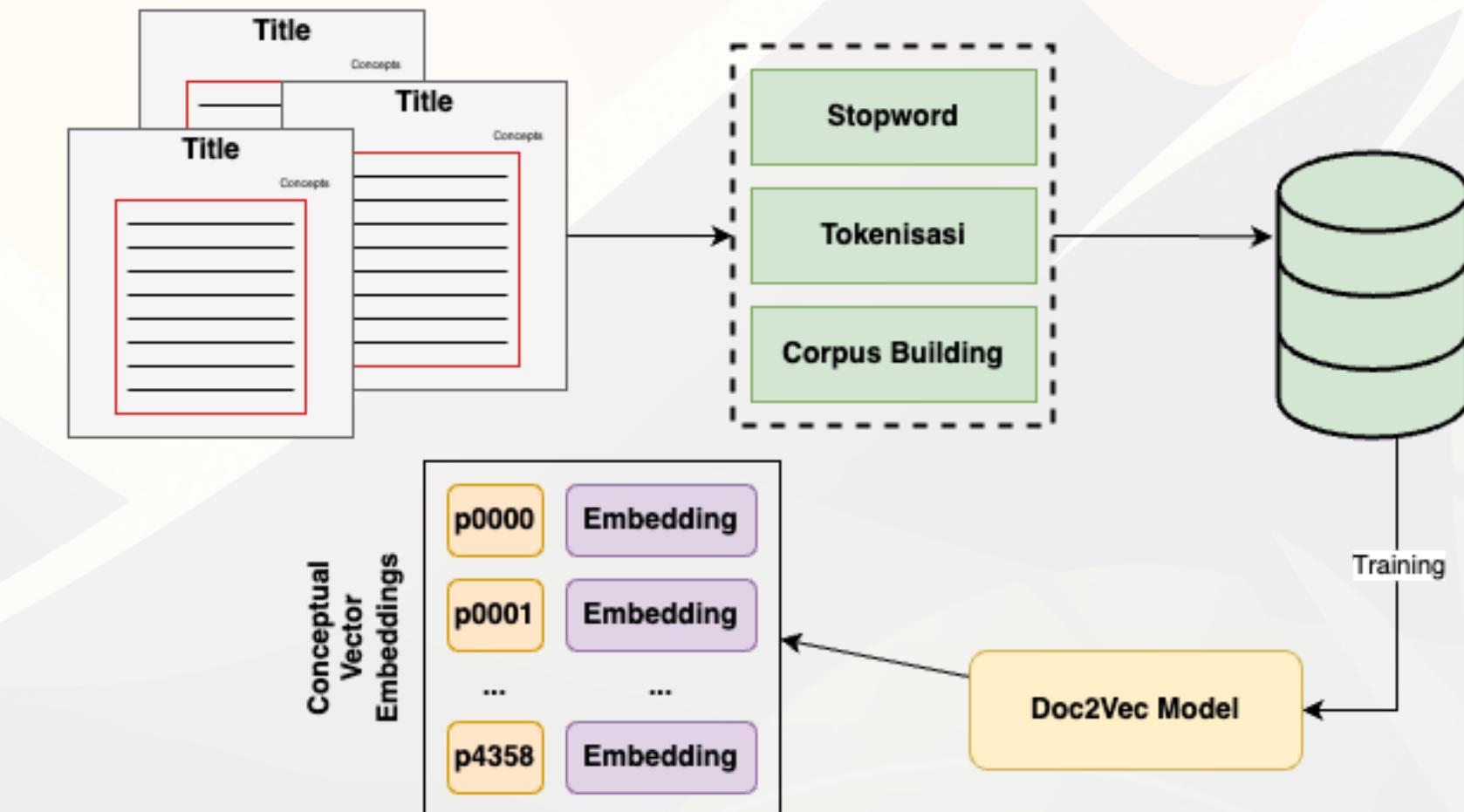
Paper	Title	Concepts	...	Publication Year
p0000	Machine Learning in Medicine	AI;Machine Learning	..	2015
p0001	A literature ... problems	Benchmark; CompSci	..	1998
p0002	Gaussian .... Motion	AI;CompSci	..	2007
...	...	...	...	...
p4358	Applied ... Physic	Physic;Medical	..	2002

Metadata berupa tabel dengan 17 kolom, terdiri dari kolom bertipe string, seperti kategori dan data tidak terstruktur seperti judul atau konsep dari sebuah jurnal, serta kolom bertipe numerik.

# DOCUMENT EMBEDDING



**BERT Extraction**



**Doc2Vec Extraction**

# DOCUMENT EMBEDDING COMPARISON

Aspek	Doc2Vec	BERT/Transformer-Based
Kebutuhan Sumber Daya	Ringan, tidak perlu GPU <span style="color: green;">✓</span>	Butuh GPU & RAM besar untuk pelatihan/ekstraksi
Pemahaman Konteks	Cukup baik, konteks berdasarkan urutan kata	Sangat baik, konteks dua arah (bidirectional) <span style="color: green;">✓</span>
Dokumen Panjang	Sangat cocok, tidak dibatasi panjang input <span style="color: green;">✓</span>	Terbatas (umumnya 512 token), kecuali model khusus seperti Longformer
Kecepatan Pelatihan	Cepat, lebih ringan dan langsung bisa memproses dokumen panjang <span style="color: green;">✓</span>	Relatif lebih lambat karena model besar dan perlu membagi dokumen panjang menjadi chunk terlebih dahulu

Doc2Vec dipilih karena efisien secara komputasi, murah, mampu menangani dokumen panjang, dan memberikan hasil yang cukup baik tanpa memerlukan sumber daya besar seperti BERT.

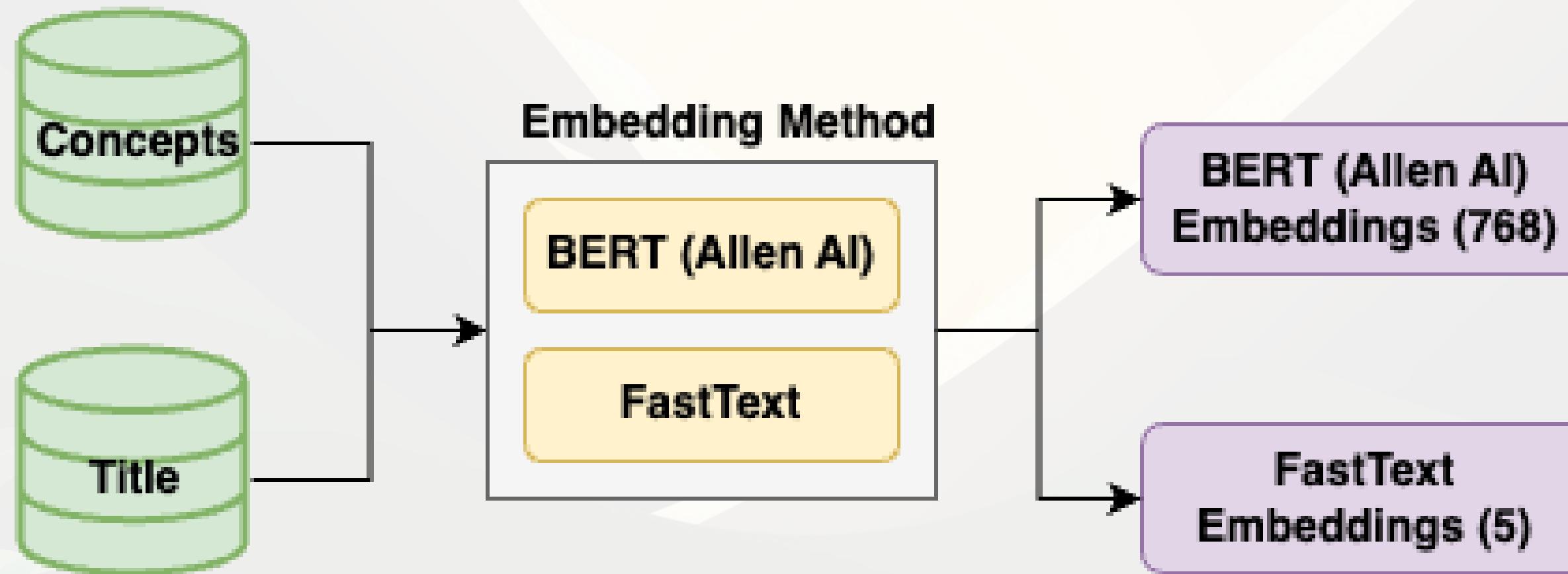
# METADATA EMBEDDING

Paper	Title	Concepts	...	Publication Year
p0000	Machine Learning in Medicine	AI;Machine Learning	..	2015
p0001	A literature ... problems	Benchmark; CompSci	..	1998
p0002	Gaussian .... Motion	AI;CompSci	..	2007
...	...	...	...	...
p4358	Applied ... Physic	Physic;Medical	..	2002

Title Embedding

Concepts Embedding

# METADATA EMBEDDING



# FEATURE ENGINEERING

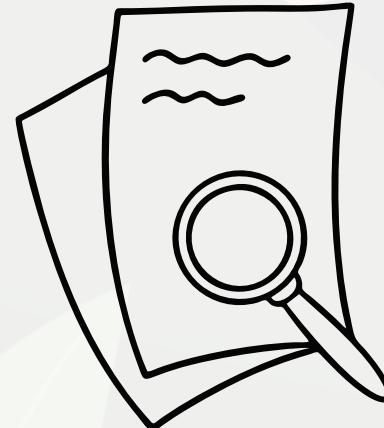


PT AYASKARA NISITA SYNERGY  
Market Research and Management Consultants

starcore.co

# VECTOR FEATURE ENGINEERING

## DISTANCE CALCULATION (2 VECTOR )



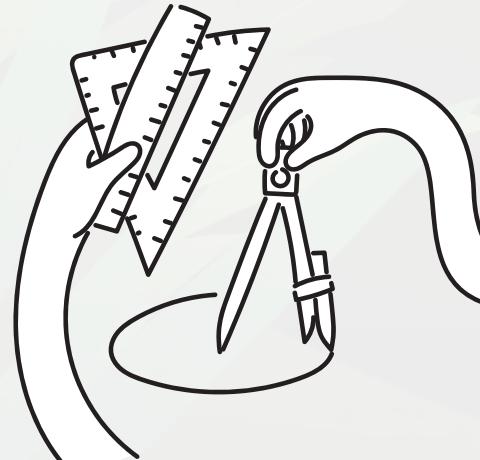
Cosine Similarity  
Euclidian Distance  
Manhattan Distance  
Pearson Correlation

## SCALAR CALCULATION (2 VECTOR)

Mean Combined  
Standar Deviation Diff  
Squared Diff Sumation  
Absolute Diff Sumation



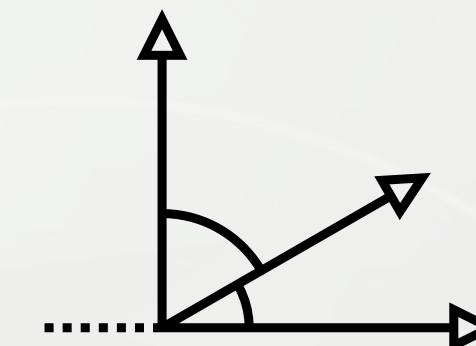
## PROJECTION CALCULATION (2 VECTOR )



Vector Projection

## ANGULAR CALCULATION (2 VECTOR )

Angular Cosine Similarity



## NORMALIZED PUBLICATION YEAR (2 PAPER)

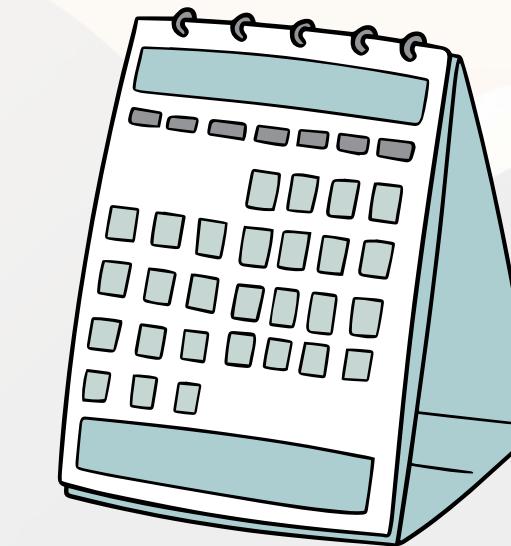


Now Year Diff  
Diff Age  
Diff Ratio  
Ratio Age

# METADATA FEATURE ENGINEERING

## DATE TIME EXTRACTION (2 PAPER)

Diff Date  
Diff Month  
Diff Week  
Diff Year



## AUTHOR AND JOURNAL NAME EXTRACTION



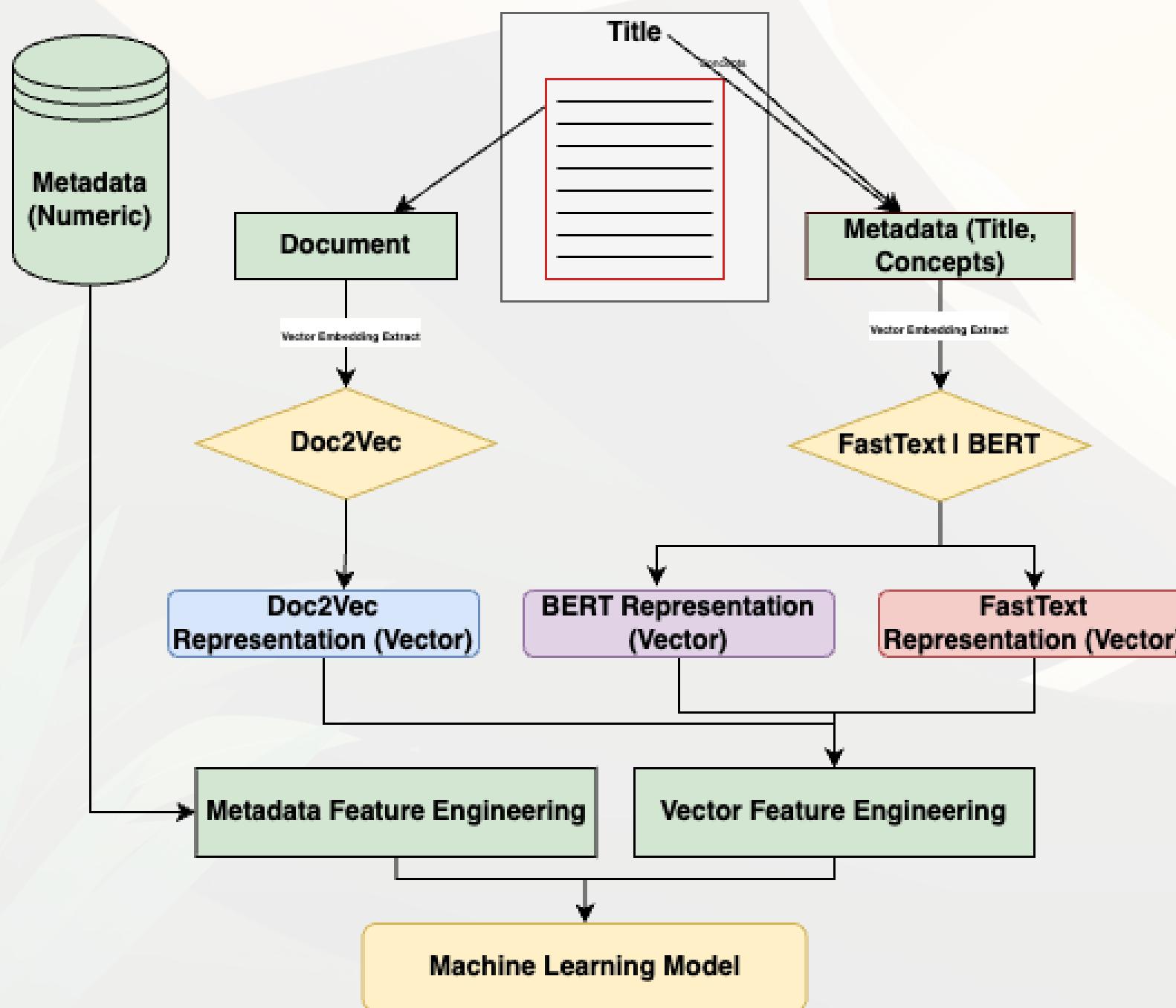
Primary Author  
Journal Name

## CITED COUNT EXTRACTION

CITATION DIFF  
CITATION RATIO



# KEY POINTS



Document Representation

Metadata Representation

# MODELLING



PT AYASKARA NISITA SYNERGY  
Market Research and Management Consultants

[starcore.co](http://starcore.co)

# MODEL EXPERIMENT FLOW

Model

MCC

CatBoost + TF IDF Document + Metadata Mentah

0.32

CatBoost Tuned + TF IDF + TF IDF Document + Metadata Mentah

0.33

LGBM Tuned + TF IDF + Cosine Document + Metadata Mentah

0.41

RandomForest Tuned + TF IDF + Cosine Document + Metadata Mentah

0.47

CatBoost + Doc2Vec Document + Metadata Mentah

0.5

# MODEL EXPERIMENT FLOW

Model

MCC

CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + Metadata Mentah

0.518

CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + FE Metadata

0.52

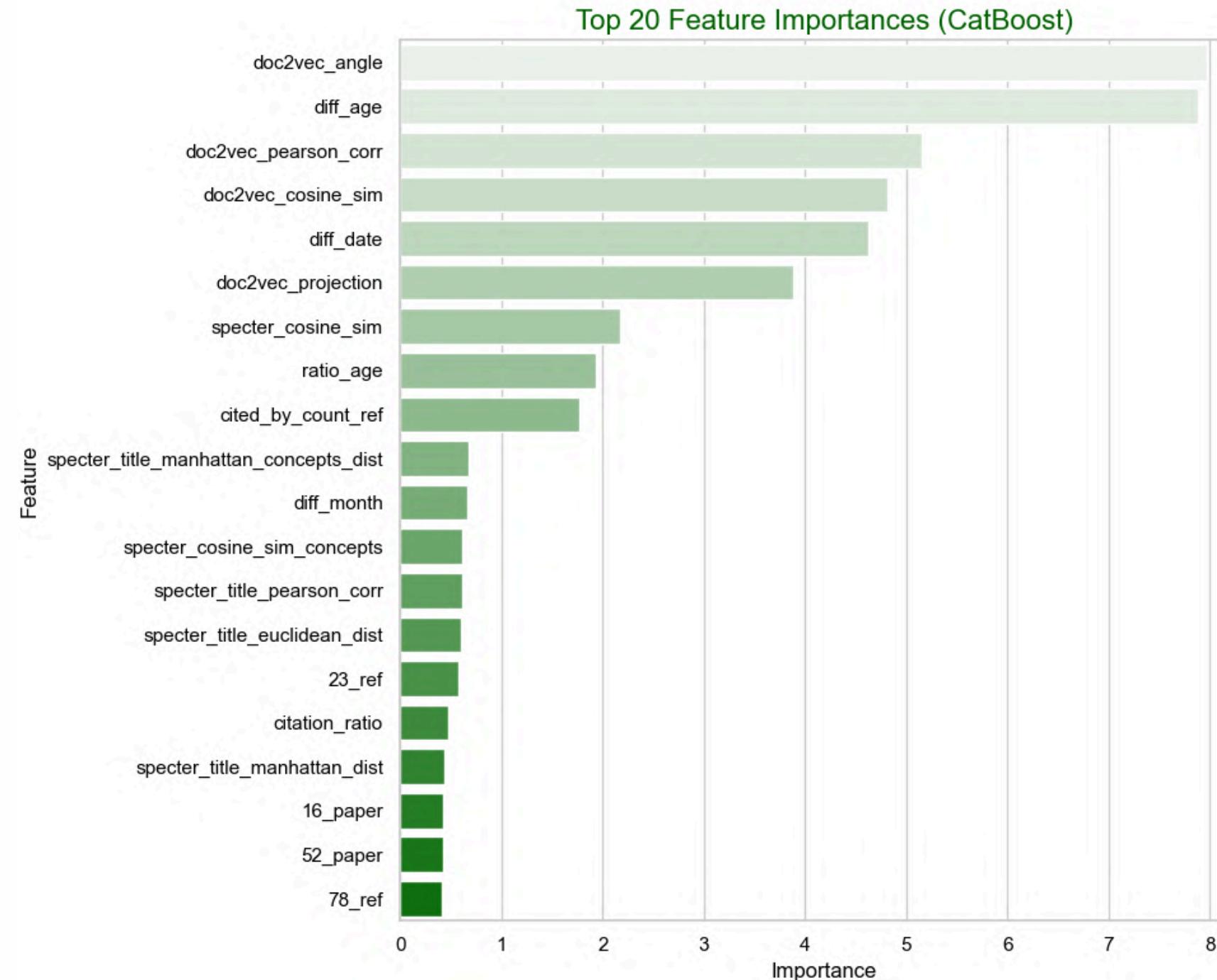
CatBoost + Doc2Vec Document + Fast Text (Concepts + Title) + BERT (Concepts + Title)  
+ FE Metadata + FE Vector Embeddings

0.54

CatBoost Tuned + Doc2Vec Document + Fast Text (Concepts + Title) + BERT (Concepts +  
Title) + FE Metadata + FE Vector Embeddings

0.568

# BEST MODEL EVALUATE



## Confusion Matrix

69045	44
527	192

Model masih mengalami kesalahan dalam mengklasifikasikan kelas minoritas, namun karena menggunakan MCC (Matthews Correlation Coefficient), evaluasi tetap adil meskipun data tidak seimbang.

# CONCLUSION

1. Pemanfaatan NLP (representasi **vektor embedding**) merupakan **kunci** dalam merepresentasikan sebuah dokumen. Selain itu tanggal **penerbitan** juga menjadi **kunci** untuk **mendeteksi hubungan** antar dokumen
2. Hubungan antar dokumen dapat **direpresentasikan** secara matematis dengan **menghitung dua vektor embedding** antar dokumen.
3. Model Machine Learning Sederhana dapat membangun sistem referensi standar yang cukup baik dengan skor MCC sekitar **0,568** mendekati **0,57**.
4. Dengan **biaya yang cukup murah**, cara ini dapat menghasilkan **hasil** yang **cukup baik**



# THANK YOU

