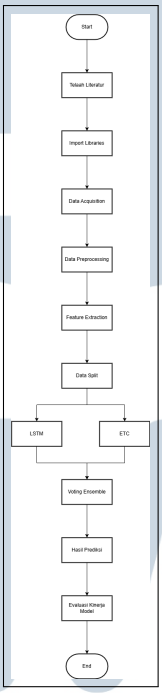
BAB 3
METODOLOGI PENELITIAN

3.1 Alur Penelitian

Alur penelitian merupakan metodologi yang ditampilkan dalam *flowchart* berikut. Langkah alur penelitian secara sistematis dimulai dari studi literatur, *data acquisition* dari *GitHub* penelitian sebelumnya, *data preprocessing*, *feature extraction* dengan *TF-IDF*, *Doc2Vec*, dan *Word2Vec*, *data split*, algoritma *LSTM* dan *ETC*, metode *ensemble learning* dengan cara *soft voting*, dan evaluasi model algoritma.



Gambar 3.1. Flowchart Alur Penelitian

3.2 Metode Penelitian

3.2.1 Telaah Literatur

Telaah literatur dilakukan dengan cara mengumpulkan informasi mengenai penelitian terdahulu yang memiliki hubungan dengan penelitian yang dilakukan. Literatur yang digunakan dalam penelitian ini mengenai lirik lagu, *preprocessing*, algoritma *Extra Trees Classifier*, algoritma *LSTM*, dan *voting classifier*.

3.2.2 Import Libraries

Import Libraries merupakan langkah awal dalam proses pemodelan. *Libraries* merupakan kumpulan *files* yang mengandung *functions* yang bisa digunakan dalam program. *Libraries* dibutuhkan untuk membuat model serta mengevaluasi performa model yang digunakan. Library yang digunakan dalam penelitian adalah *library NLTK (Natural Language Toolkit)*. NLTK berfungsi untuk membbantu memproses *NLP (Natural Language Processing)* dalam bahasa *python*.

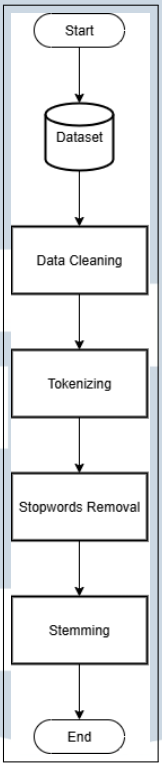
3.2.3 Data Acquisition

Data acquisition merupakan langkah mengunduh *dataset* publik di platform *GitHub*. *Dataset* memiliki kolom seperti nama artis, nama lagu, lirik lagu, dan lagu eksplisit. *Dataset* yang digunakan diambil dari penelitian sebelumnya mengenai deteksi lagu eksplisit dimana terdapat di *GitHub*. Data ini disimpan ke dalam *file* dengan format *Comma Separated Values (CSV)*. *Dataset* memiliki 5566 data dengan 30 atribut, dengan 3 atribut utama yaitu nama lagu (*song*), lirik lagu (*lyrics*), dan klasifikasi lagu eksplisit (*explicit*). Data ini digunakan dalam penelitian untuk menentukan lagu sebagai lagu mengandung konten eksplisit atau tidak.

artist_all	artist_base	rank	song	year	artist_feats	song_clean	artist_clea	lyrics	acousticne	danceabili	energy	explicit	instrument	key	liveness	loudness	mode	popularity	release_da	speechine	tempo	time_signa	va
percy faith	percy faith	1	theme fron	1960		theme fror	percy faith	theres a su	0.631	0.466	0.389	0	0.843	5	0.295	-12.825	1	49	1953	0.0253	92.631	4	
jim reeves	jim reeves	2	he'll have t	1960		hell have t	jim reeves	put your sw	0.909	0.554	0.186	0	0.00144	1	0.11	-15.846	1	53	########	0.0379	81.181	3	
the everly b	the everly b	3	cathy's clo	1960		cathys clo	the everly b	dont want	0.412	0.498	0.582	0	0	7	0.372	-8.961	1	50	########	0.0339	119.809	4	
johnny pre	johnny pre	4	running be	1960		running be	johnny pre	on the ban	0.854	0.772	0.297	0	7.59E-06	5	0.125	-14.679	0	40	########	0.053	119.987	4	
mark dinni	mark dinni	5	teen angel	1960		teen angel	mark dinni	teen angel	0.936	0.57	0.0636	0	0	0	0.122	-18.548	1	18	2005	0.0459	101.517	4	
brenda lee	brenda lee	6	i'm sorry	1960		im sorry	brenda lee	im sorry so	0.928	0.529	0.115	0	0.00333	10	0.121	-16.284	1	48	########	0.0302	101.501	3	
elvis presle	elvis presle	7	it's now or	1960		its now or	elvis presle	its now or	0.642	0.643	0.491	0	0.00972	4	0.286	-9.312	1	58	2002	0.0344	126.399	4	
jimmy jone	jimmy jone	8	handy man	1960		handy man	jimmy jone	hey girls ga	0.408	0.458	0.659	0	0	10	0.247	-8.97	1	28	########	0.0497	144.758	4	
elvis presle	elvis presle	9	stuck on yc	1960		stuck on yc	elvis presle	you can sh	0.758	0.647	0.513	0	8.64E-06	7	0.108	-12.372	1	59	########	0.0421	131.641	4	
chubby ch	chubby ch	10	the twist	1960		the twist	chubby ch	come on b	0.199	0.526	0.633	0	1.24E-06	4	0.0709	-7.119	1	52	########	0.0322	156.405	4	
connie frai	connie frai	11	everybody'	1960		everybody'	connie frai	the tears i	0.826	0.589	0.643	0	3.62E-05	1	0.237	-8.118	1	37	########	0.0304	84.308	4	
bobby ryde	bobby ryde	12	wild one	1960		wild one	bobby ryde	oh wild one	0.709	0.599	0.775	0	0	7	0.0804	-4.566	1	24	########	0.0466	148.507	4	
the brother	the brother	13	greenfields	1960		greenfields	the brother	once there	0.852	0.485	0.182	0	0	11	0.112	-18.083	0	41	1960	0.0434	110.608	1	
jack scott	jack scott	14	what in the	1960		what in the	jack scott	what in the	0.802	0.506	0.176	0	3.93E-06	4	0.0978	-13.962	1	31	########	0.0303	76.273	4	
marty robb	marty robb	15	el paso	1960		el paso	marty robb	out in the w	0.835	0.654	0.452	0	2.89E-05	2	0.16	-9.709	1	58	1959	0.03	106.662	3	
the hollywo	the hollywo	16	alley oop	1960		alley oop	the hollywo	theres a m	0.798	0.65	0.352	0	4.10E-06	2	0.114	-10.304	1	10	########	0.052	64.596	4	
connie frai	connie frai	17	my heart h	1960		my heart h	connie frai	i told this h	0.805	0.338	0.551	0	0	6	0.316	-5.995	1	19	########	0.0465	108.045	4	
brenda lee	brenda lee	18	sweet noth	1960		sweet noth	brenda lee	alright my	0.785	0.787	0.359	0	6.03E-05	5	0.198	-12.824	1	37	########	0.0663	125.216	4	
brian hylar	brian hylar	19	itsy bitsy te	1960		itsy bitsy te	brian hylar	she was af	0.682	0.785	0.4	0	0	2	0.0326	-13.01	1	46	########	0.0749	122.174	4	
roy orbison	roy orbison	20	only the lor	1960		only the lor	roy orbison	dum dum c	0.377	0.57	0.529	0	0.00509	5	0.203	-10.769	1	60	1961	0.028	123.273	4	
dion and th	dion and th	21	where or w	1960		where or w	dion and th	it seems w	0.797	0.454	0.271	0	0	8	0.126	-11.658	1	38	########	0.0283	110.89	3	
connie ste	connie ste	22	sixteen rea	1960		sixteen rea	connie ste	sixteen rea	0.865	0.339	0.406	0	1.10E-05	10	0.11	-8.955	1	30	########	0.0319	109.783	3	
paul anka	paul anka	23	puppy love	1960		puppy love	paul anka	and they cc	0.662	0.29	0.408	0	0	7	0.272	-9.607	1	46	########	0.0318	102.229	3	

Gambar 3.2. Import Dataset

3.2.4 Data Preprocessing



Gambar 3.3. Preprocessing Database

Preprocessing merupakan langkah untuk mengubah data menjadi data bersih sehingga data dapat dianalisis. Berikut proses *preprocessing* yang dilakukan.

1. *Data Cleaning*
   Data cleaning merupakan proses membersihkan data dimana elemen yang tidak dibutuhkan atau relevan dengan data yang ingin diproses dihilangkan.

2. *Tokenizing*
   Tokenizing merupakan proses memisahkan teks menjadi kata tunggal atau frasa untuk lebih mudah melakukan proses terhadap data.

3. *Normalization*
   Normalization merupakan proses mengubah kata teks yang tidak baku menjadi kata baku.

4. *Stopword Removal*
   Stopword Removal merupakan proses menghapus kata yang tidak memiliki makna penting seperti ”yang”, ”di”, atau ”dan”.

5. *Stemming*

   Stemming merupakan proses mengubah kata imbuhan menjadi kata baku tanpa imbuhan.

3.2.5 Feature Extraction

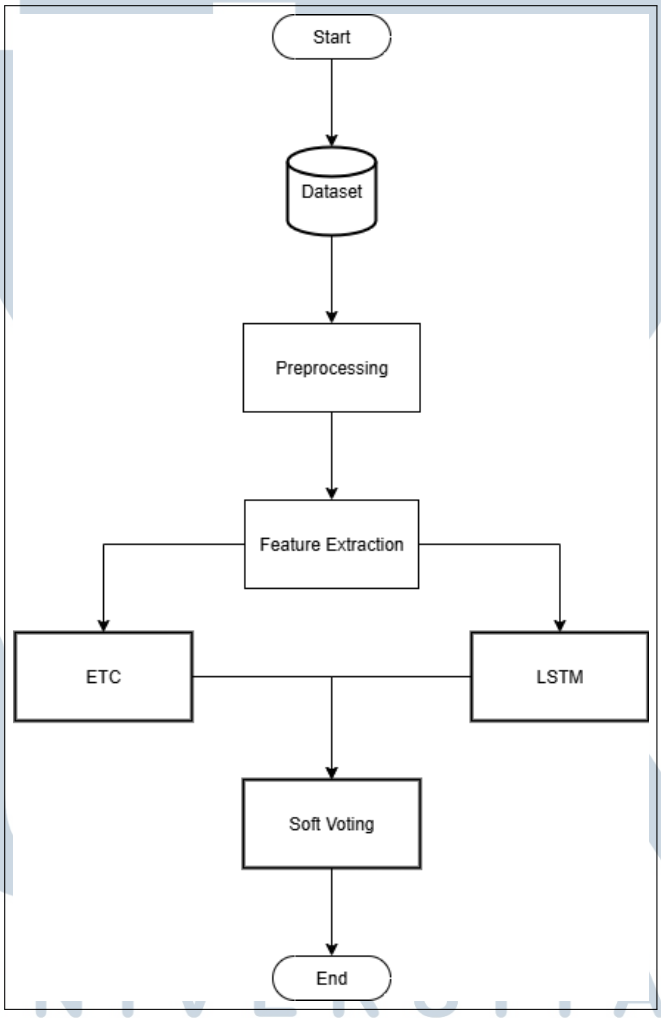
Feature extraction adalah proses yang digunakan untuk mengubah kata dalam teks menjadi nilai *integer* atau *float*. Penelitian ini menggunakan *TF-IDF Vectorizer* dan *Doc2Vec* untuk *machine learning* dan *Word2Vec* untuk *deep learning*. Dalam *feature extraction TF-IDF*, *TF* digunakan dalam mengukur kata muncul dalam dokumen, semakin sering kata muncul maka nilai *TF* akan semakin tinggi. *IDF* merupakan pengukur kata yang dipengaruhi oleh seberapa umum atau langka kata dalam dokumen. Kombinasi *TF* dan *IDF* akan menghasilkan vektor numerik yang menjadi representasi kata dalam dokumen. *Doc2Vec* merupakan *feature extraction* yang bekerja dengan menggunakan *Distributed Memory (DM)* dan *Distributed Bag of Words (DBOW)*. *DM* berfungsi dalam mempertimbangkan konteks dari kata dan ID dokumen sebagai input prediksi kata selanjutnya. *DBOW* berfungsi untuk memprediksi kata dari dokumen dengan menggunakan vektor dokumen sebagai konteks. *Word2Vec* merupakan *feature extraction* yang bekerja dengan menggunakan *Continuous Bag of Words (CBOW)* dan *Skip-gram*. *CBOW* berfungsi dalam memprediksi kata target dengan kata konteks di sekitar. *Skip-gram* berfungsi dalam memprediksi kata konteks dengan kata target.

3.2.6 Algoritma Klasifikasi

Penelitian ini menggunakan algoritma *Extra Trees Classifier* dan *LSTM* untuk mencari klasifikasi terbaik dalam analisis kata seksual dalam lirik lagu. Tahap ini dilakukan untuk mencari nilai hasil dari akurasi algoritma terbaik untuk mengukur kinerja model dari *confusion matrix* dalam melakukan klasifikasi kata seksual dalam lirik lagu. Tahapan yang dilakukan adalah *data acquisition*, *data preprocessing*, *feature extraction* dengan menggunakan *TF-IDF* dan *Doc2Vec* untuk algoritma *ETC* dan *Word2Vec* untuk algoritma *LSTM*, *data split*, *modeling ETC*, dan evaluasi. Di awal, dilakukan *preprocessing* dengan *data cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming* untuk *dataset*. Lalu data akan dibuat menjadi nilai dengan *feature extraction* dengan *TF-IDF*, *Word2Vec*, dan *Doc2Vec*. Tahap berikutnya adalah melakukan *modeling* dengan menggunakan

ETC dan *LSTM*, dimana digunakan *hyperparameter* dari penelitian sebelumnya dan menggunakan *cross validation* untuk mencari rata-rata akurasi model. Kemudian dilakukan evaluasi *confusion matrix* dari setiap model. Performa model dilihat dari nilai tabel *confusion matrix*.

3.2.7 Metode Ensemble Learning



Gambar 3.4. Metode Ensemble Learning

Ensemble Learning merupakan proses pembelajaran dimana menggabungkan beberapa model dasar untuk mendapatkan model baru dengan hasil yang lebih baik dibandingkan model dasar. Algoritma yang digunakan dalam *ensemble learning* adalah algoritma *LSTM* dan *ETC*. Metode *ensemble learning* yang digunakan dalam penelitian sendiri menggunakan *soft voting* dimana bekerja

dengan cara mengambil hasil dari gabungan model algoritma dan mengambil rata-rata dari probabilitas semua model yang digunakan untuk menghasilkan hasil akhir. Dalam penelitian, *soft voting* yang digunakan merupakan *ETC + LSTM* dengan *soft voting* pertama menggunakan *feature extraction TF-IDF* dalam *ETC* dan *soft voting* kedua menggunakan *feature extraction Doc2Vec* dalam *ETC*.

A K-Fold Cross Validation

K-fold cross validation merupakan metode validasi yang membagi dataset menjadi beberapa bagian *(fold)* dengan ukuran yang sama. Dalam penelitian ini digunakan *10-fold cross validation*, di mana dataset dibagi menjadi 10 bagian. Setiap iterasi menggunakan satu *fold* sebagai *testing set* dan sembilan *fold* lainnya sebagai *training set*. Proses ini dilakukan sebanyak 10 kali untuk memastikan generalisasi model terhadap data yang berbeda pada setiap iterasi.

3.2.8 Testing Model

Setelah model algoritma dibangun menggunakan pendekatan *ensemble voting*, langkah selanjutnya adalah melakukan pengujian terhadap model tersebut. Pengujian dilakukan dengan menggunakan data berupa lirik lagu, dan hasil dari prediksi akan menunjukkan apakah lirik tersebut tergolong *Explicit* atau *Non Explicit*, sesuai dengan klasifikasi model.

3.2.9 Hasil Prediksi

Hasil prediksi merupakan visualisasi evaluasi dari pendekatan *ensemble learning* yang terdiri dari dua model *soft voting* dengan teknik *feature extraction* yang berbeda. Visualisasi ini bertujuan untuk membandingkan kinerja model berdasarkan representasi fitur yang digunakan serta menilai akurasi klasifikasi dalam mendeteksi konten eksplisit pada lirik lagu.