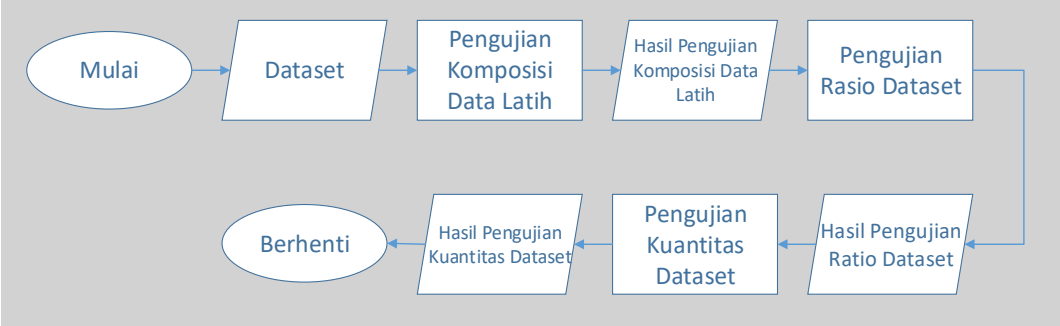
BAB 4
HASIL PENELITIAN

4.1 Hasil Pengujian Dataset

Pengujian penelitian ini menggunakan data ulasan film dengan kuantitas paling minimum adalah 1.000 dokumen untuk data latih dan 100 dokumen untuk data uji. Berdasarkan dari bahasan subbab 3.1.2, penelitian ini merupakan kinerja dari teknik pembobotan TF-IDF dan TF-OR dengan algoritma SVM di mana dilakukan tiga jenis pengujian yaitu pengujian komposisi, rasio, dan kuantitas. Ketiga pengujian tersebut sangat berkaitan erat karena adanya keterkaitan hasil suatu pengujian yang selanjutnya akan digunakan sebagai landasan dalam pengujian lainnya.



Gambar 4-1Keterkaitan Antar Hasil Pengujian

Berdasarkan gambar di atas, masing-masing variabelnya dapat mempengaruhi pengujian yang lainnya, kecuali pada pengujian komposisi data latih karena ini menjadi fondasi awal dalam pengujian yang akan dilakukan. Jumlah data yang digunakan dalam setiap penelitian akan sangat bervariatif. Berikut adalah hasil dari ketiga pengujian tersebut.

4.1.1 Hasil Kinerja Berdasarkan Komposisi Data Latih

Berikut adalah hasil dalam satuan persentase dari pengujian algortiama SVM TF-IDF dan SVM TF-OR berdasarkan penjelasan pada subbab 3.1.2.1.

Tabel 4-1 Hasil Pengujian Komposisi Data Latih

Komposisi	Percobaan																			
	1		2		3		4		5		6		7		8		9		10	
	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF
10+ 90	50	50	55	50	54	50	50	50	51	50	50	50	50	50	50	50	50	51	51	50
15+ 85	52	52	56	50	58	50	56	53	50	50	56	50	53	50	50	50	51	50	55	50
20+ 80	62	58	59	59	62	58	58	51	61	53	64	57	66	59	55	55	50	53	54	54
25+ 75	60	62	59	62	56	69	57	63	60	66	64	63	57	64	54	68	60	66	59	58
30+ 70	73	76	56	72	72	83	71	71	73	77	64	70	64	74	66	71	58	65	70	70
35+ 65	79	82	84	86	84	77	68	82	68	77	73	77	69	75	79	82	70	72	66	79
40+ 60	70	90	84	86	70	82	87	85	82	89	77	83	70	87	73	88	83	80	84	83
45+ 55	80	87	82	82	84	83	85	89	87	86	88	86	89	80	81	79	85	78	86	76
50+ 50	85	82	89	86	80	83	86	83	86	85	84	76	88	86	86	86	88	87	85	81
55+ 45	81	82	84	82	80	83	84	89	84	89	72	85	87	83	85	86	83	82	78	86
60+ 40	83	78	74	87	84	80	80	78	79	77	76	77	81	78	70	86	83	80	78	77
65+ 35	62	80	72	76	79	75	78	74	73	76	80	83	74	71	78	80	77	76	82	81
70+ 30	70	76	79	72	79	76	68	72	69	69	69	76	69	66	73	67	75	72	70	71
75+ 25	69	66	64	68	64	68	55	57	63	68	75	65	69	66	75	67	71	67	66	68

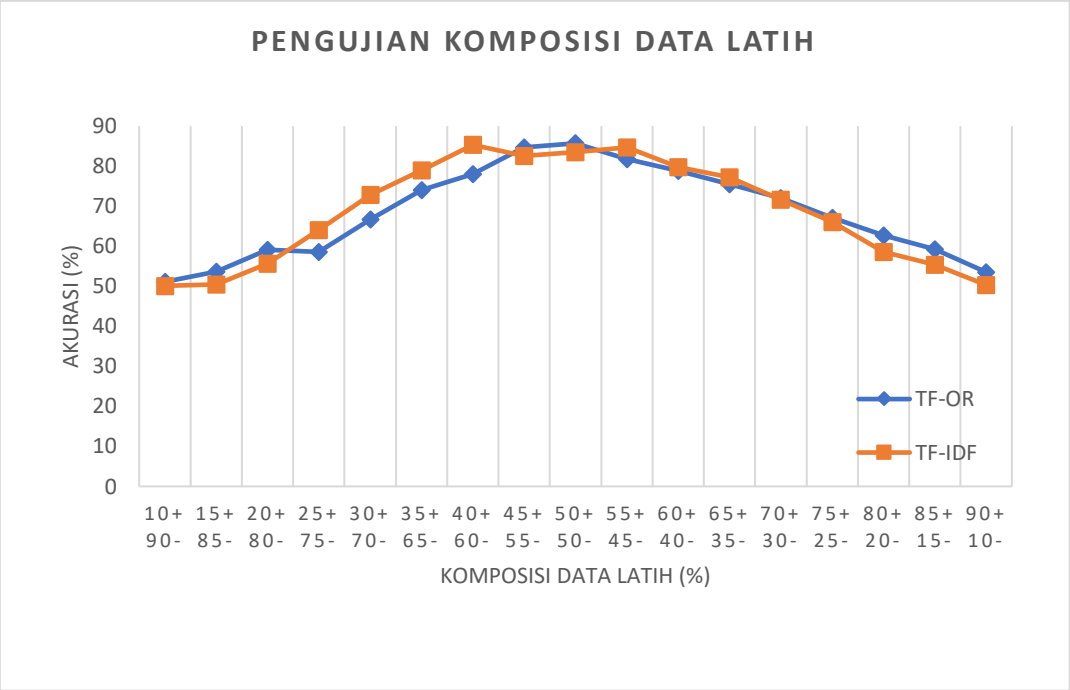
Komposisi	Percobaan																			
	1		2		3		4		5		6		7		8		9		10	
	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF	TF-OR	TF-IDF
80+ 20	61	56	67	56	67	64	53	50	60	64	62	60	56	62	70	62	65	54	66	58
85+ 15	64	56	60	56	60	56	57	53	60	57	51	53	59	60	60	54	66	55	56	54
90+ 10	57	50	54	50	54	50	52	50	57	50	55	52	52	51	52	50	52	50	50	50

Berikut adalah hasil statistik dalam satuan persentase berdasarkan Tabel 4-1 yang terdiri dari 10 kali percobaan, dari 17 komposisi data latih yang berbeda, dengan masing-masing menggunakan dua kombinasi algoritma berbeda yaitu SVM TF-OR dan SVM TF-IDF.

Tabel 4-2 Rangkuman Pengujian Komposisi Data

Komposisi	Rata-Rata (%)		Deviasi		Tertinggi
	TF-OR	TF-IDF	TF-OR	TF-IDF	
10+ 90-	51,1	50,1	1,76	0,30	TF-OR
15+ 85-	53,7	50,5	2,72	1,02	TF-OR
20+ 80-	59,1	55,7	4,68	2,72	TF-OR
25+ 75-	58,6	64,1	2,62	3,08	TF-IDF
30+ 70-	66,7	72,9	5,85	4,66	TF-IDF
35+ 65-	74	78,9	6,54	3,91	TF-IDF
40+ 60-	78	85,3	6,42	3,10	TF-IDF
45+ 55-	84,7	82,6	2,83	4,10	TF-OR
50+ 50-	85,7	83,5	2,41	3,14	TF-OR
55+ 45-	81,8	84,7	4,09	2,61	TF-IDF
60+ 40-	78,8	79,8	4,21	3,52	TF-IDF
65+ 35-	75,5	77,2	5,41	3,49	TF-IDF
70+ 30-	72,1	71,7	3,99	3,44	TF-OR
75+ 25-	67,1	66	5,75	3,16	TF-OR
80+ 20-	62,7	58,6	5,06	4,39	TF-OR
85+ 15-	59,3	55,4	3,93	2,01	TF-OR
90+ 10-	53,5	50,3	2,20	0,64	TF-OR

Tabel diatas menunjukkan suatu pola di mana pada jarak komposisi tertentu, suatu kombinasi algoritma mendominasi. Terhitung kombinasi SVM TF-OR memiliki nilai tertinggi dari 10 komposisi data latih dan di 7 komposisi data latih sisanya, SVM TF-IDF memiliki keunggulan.



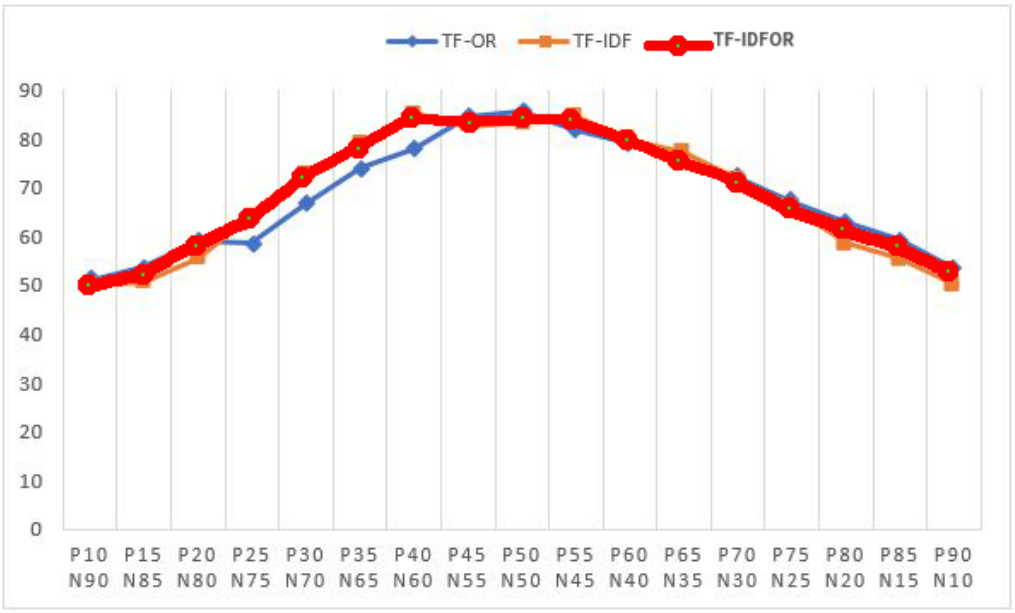
Gambar 4-2 Grafik Hasil Pengujian Komposisi Data Latih

Dari gambar di atas terlihat keberagaman tingkat akurasi yang diperoleh dari setiap komposisi data latih menggunakan algoritma SVM TF-IDF dan SVM TF-OR. Terlihat dari hasil pengujian ini tingkat akurasi tertinggi untuk algoritma SVM TF-IDF adalah 85,3% dengan komposisi data latih 60% berlabel positif dan 40% berlabel negatif, sedangkan untuk algoritma SVM TF-OR adalah 85,7% dengan komposisi data latih 50% berlabel positif dan 50% berlabel negatif. Dapat dirumuskan TF-IDFOR sebagai berikut :

$$TF * \begin{cases} Jika\ 25\% < prevalence \geq\ 10\%, maka\ OR\ (a) \\ Jika\ 45\% < prevalence \geq\ 25\%, maka\ IDF\ (b) \\ Jika\ 55\% < prevalence \geq\ 45\%, maka\ OR\ (c) \\ Jika\ 70\% < prevalence \geq\ 55\%, maka\ IDF\ (d) \\ Jika\ 90\% \leq prevalence \geq\ 70\%, maka\ OR\ (e) \end{cases} \quad \textbf{(4. 1)}$$

Berdasarkan rumus di atas maka TF-IDFOR bekerja dengan cara TF bertemu dengan rumus berikutnya berdasarkan *prevalence* (jumlah kemunculan data latih berlabel positif). Dari lima kondisi *prevalence* dapat dirangkum menjadi tiga kondisi yaitu data latih seimbang ditunjukan oleh kondisi c, data latih tidak seimbang oleh kondisi b dan d, dan kondisi sangat tidak seimbang oleh kondisi a dan e. Maka dengan menggunakan sumber data dari Tabel 4-2, kurva yang akan

diperoleh ketika menggunakan TF-IDFOR akan terlihat sebagai berikut:



Gambar 4-3 Grafik TF-IDFOR

Sumbu TF-IDFOR berwarna merah di atas menunjukkan adanya peningkatan dengan memanfaatkan semua titik keunggulan dari TF-IDF dan TF-OR.

4.1.2 Hasil Kinerja Berdasarkan Rasio Dataset

Berdasarkan hasil pengujian komposisi data latih yang telah dilakukan, didapat bahwa titik komposisi tertinggi untuk SVM TF-DF adalah 60% positif dan 40% negatif (60+/40-), sedangkan untuk SVM TF-OR adalah 50% positif dan 50% negatif (50+/50-). Dalam pengujian kali ini, angka 1 dalam rasio ini berarti 100 dokumen telah digunakan. Maka dalam pengujian berikut akan dilakukan perubahan rasio data latih pada data uji terhadap masing-masing komposisi data tersebut sesuai dengan yang telah dijelaskan pada subbab 3.1.2.2.

Tabel 4-3 Hasil Pengujian Rasio Data Latih 50+/50-

Percobaan		Rasio				
		1:100	1:10	1:1	10:1	100:1
1	TF-OR (%)	66	70	61	87	91
	TF-IDF (%)	72	70	77	85	91
2	TF-OR (%)	72	68	65	86	90
	TF-IDF (%)	73	72	66	86	91
3	TF-OR (%)	67	60	65	86	91

Percobaan		Rasio				
		1:100	1:10	1:1	10:1	100:1
	TF-IDF (%)	66	67	68	87	91
4	TF-OR (%)	67	66	63	87	91
	TF-IDF (%)	72	71	69	82	88
5	TF-OR (%)	68	63	59	84	88
	TF-IDF (%)	71	70	68	89	87
6	TF-OR (%)	63	68	63	85	93
	TF-IDF (%)	71	68	73	86	91
7	TF-OR (%)	65	54	69	85	94
	TF-IDF (%)	68	66	64	79	90
8	TF-OR (%)	71	69	63	85	90
	TF-IDF (%)	73	70	63	85	93
9	TF-OR (%)	66	71	69	86	92
	TF-IDF (%)	71	73	68	88	90
10	TF-OR (%)	64	68	67	89	94
	TF-IDF (%)	71	72	69	87	92
RATA-RATA	TF-OR (%)	66,9	65,7	64,4	86	91,4
	TF-IDF (%)	70,8	69,9	68,5	85,4	90,4
DEVIASI	TF-OR	2,85	5,27	3,27	1,41	1,90
	TF-IDF	2,20	2,28	4,09	2,95	1,78

Berdasarkan hasil pengujian komposisi data latih 50+/50- pada Tabel 4-3 di atas, didapatkan bahwa rasio minimal data latih terhadap data uji yang dibutuhkan untuk mendapatkan tingkat akurasi di atas 80% adalah 10:1.

Tabel 4-4 Hasil Pengujian Rasio Data Latih 60+/40-

Rasio		Rasio				
		1:100	1:10	1:1	10:1	100:1
1	TF-OR (%)	55	59	61	76	90
	TF-IDF (%)	55	57	58	87	91
2	TF-OR (%)	57	55	50	86	88
	TF-IDF (%)	53	53	56	71	93
3	TF-OR (%)	63	58	57	71	93
	TF-IDF (%)	56	57	54	80	93
4	TF-OR (%)	55	58	56	85	89
	TF-IDF (%)	56	53	51	89	90
5	TF-OR (%)	55	54	59	87	90

	TF-IDF (%)	53	55	73	92	89
6	TF-OR (%)	62	60	61	77	94
	TF-IDF (%)	56	60	58	80	91
7	TF-OR (%)	56	53	57	85	91
	TF-IDF (%)	53	54	54	82	90
8	TF-OR (%)	64	61	64	76	89
	TF-IDF (%)	56	58	55	88	91
9	TF-OR (%)	54	58	60	77	92
	TF-IDF (%)	53	57	63	83	92
10	TF-OR (%)	55	55	48	77	89
	TF-IDF (%)	57	55	57	83	91
RATA-RATA	TF-OR (%)	57,6	57,1	57,3	79,7	90,5
	TF-IDF (%)	54,8	55,9	57,9	83,5	91,1
DEVIASI	TF-OR (%)	3,84	2,69	4,99	5,52	1,96
	TF-IDF (%)	1,62	2,28	6,19	5,95	1,29

Berdasarkan hasil pengujian komposisi data latih 60+/40- pada Tabel 4-4 di atas, rasio minimal data latih terhadap data uji yang dibutuhkan untuk mendapatkan tingkat akurasi di atas 80% adalah 10:1.

4.1.3 Hasil Kinerja Berdasarkan Kuantitas Dataset

Berdasarkan hasil pengujian pada Tabel 4-3 dan Tabel 4-4 menunjukkan bahwa rasio perbandingan data latih dan data uji yang ideal sebagai langkah awal membangun aplikasi sentimen analis adalah 10:1. Pada pengujian kali ini dari rasio tersebut akan dinaikan dari perbandingan data latih dan data uji 10:1 hingga menjadi 100:10 atau 10.000 dokumen data latih dan 1.000 data uji.

Tabel 4-5 Hasil Pengujian Kuantitas Data Latih Rasio 10:1 Komposisi 50+/50-

Percobaan		Kuantitas		
		1x	5x	10x
1	TF-OR (%)	66	83	83
	TF-IDF (%)	66	81	83
2	TF-OR (%)	65	82	84
	TF-IDF (%)	71	82	83
3	TF-OR (%)	64	82	85
	TF-IDF (%)	68	81	83
4	TF-OR (%)	62	81	84
	TF-IDF (%)	62	80	83

Percobaan		Kuantitas		
		1x	5x	10x
5	TF-OR (%)	64	75	84
	TF-IDF (%)	74	80	83
6	TF-OR (%)	71	82	80
	TF-IDF (%)	72	81	83
7	TF-OR (%)	72	80	85
	TF-IDF (%)	75	81	83
8	TF-OR (%)	60	82	82
	TF-IDF (%)	66	81	83
9	TF-OR (%)	68	82	83
	TF-IDF (%)	71	82	83
10	TF-OR (%)	66	82	84
	TF-IDF (%)	71	82	83
RATA-RATA	TF-OR (%)	65,8	81,1	83,4
	TF-IDF (%)	69,6	81,1	83
DEVIASI	TF-OR (%)	3,74	2,28	1,51
	TF-IDF (%)	4,03	0,74	0,32

Berdasarkan hasil pengujian dengan komposisi 50+/50- dengan ratio 10:1 pada Tabel 4-5 di atas, menunjukkan bahwa kuantitas dataset yang optimal untuk menggunakan TF-OR sebagai teknik pembobotan SVM adalah 10.000:100.

Tabel 4-6 Hasil Pengujian Kuantitas Data Latih Rasio 10:1 Komposisi 60+/40-

Percobaan		Kuantitas		
		1x	5x	10x
1	TF-OR (%)	56	69	79
	TF-IDF (%)	55	75	80
2	TF-OR (%)	61	79	79
	TF-IDF (%)	56	76	80
3	TF-OR (%)	54	53	79
	TF-IDF (%)	55	76	81
4	TF-OR (%)	58	77	77
	TF-IDF (%)	57	77	80
5	TF-OR (%)	54	63	81
	TF-IDF (%)	61	77	80
6	TF-OR (%)	63	71	79
	TF-IDF (%)	58	76	81

Percobaan		Kuantitas		
		1x	5x	10x
7	TF-OR (%)	62	70	72
	TF-IDF (%)	54	76	80
8	TF-OR (%)	62	70	76
	TF-IDF (%)	61	76	81
9	TF-OR (%)	54	72	81
	TF-IDF (%)	56	77	81
10	TF-OR (%)	59	75	76
	TF-IDF (%)	56	76	81
RATA-RATA	TF-OR (%)	58,3	69,9	77,9
	TF-IDF (%)	56,9	76,2	80,5
DEVIASI	TF-OR (%)	3,62	7,45	2,73
	TF-IDF (%)	2,42	0,63	1,49

Berdasarkan hasil pengujian dengan komposisi 60+/40- dengan ratio 10:1 pada Tabel 4-6 menunjukkan bahwa kuantitas dataset mulai optimal untuk menggunakan TF-IDF sebagai teknik pembobotan SVM adalah 5000:500.