BAB 4 PERANCANGAN

4.1 Pengumpulan Data

Pengumpulan data berisi tentang data umum yang menjelaskan sumber data dan data yang akan digunakan dalam penelitian.

4.1.1 Data Umum

Kompas.com merupakan salah satu portal berita online yang menyediakan banyak fitur seperti pengkategorian, pencarian, rekomendasi 5 artikel terpopuler, rekomendasi untuk pembaca berupa artikel dalam kategori yang sama, dan berita terkait. Dalam Detik.com terdapat kategori . Didalam kategori tersebut kemudian dikategorikan lagi menjadi kategori kategori kecil. Contohnya untuk *detikFood* ada kategori resep, tempat makan, kabar kuliner, infografis, anak, dan sehat.

Pengelompokan artikel terkait pada Detik.com dapat dipisahkan dengan garis lurus. Dimana setiap artikel dengan artikel terkaitnya pasti berada dalam kategori kecil yang sama.

4.1.2 Data Artikel *LifeStyle*

Adapun beberapa artikel dalam kategori *detikFood* yang akan digunakan dalam penelitian ini akan dijelaskan dalam tabel 4.1. Data artikel berjumlah 120 dimana 40 merupakan kategori *Eat Good*, 40 *Feel Good*, dan 40 *Look Good*.

Tabel 4. 1 Data Artikel *LifeStyle*

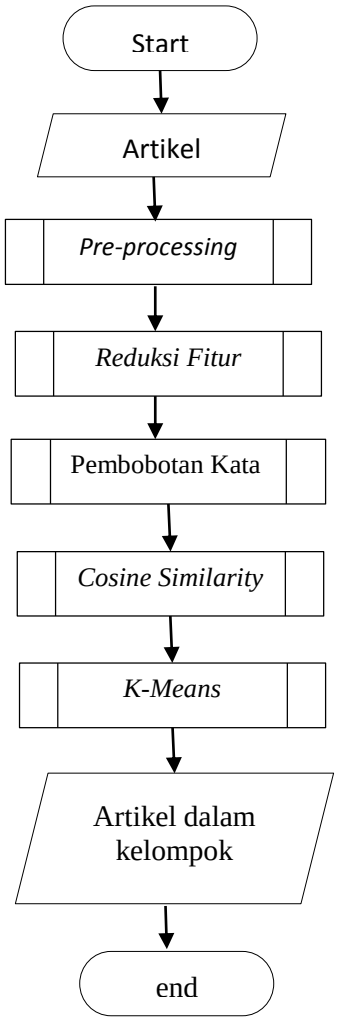
Judul	Kategori
5 Cara Mencegah Makan Berlebihan	*Eat Good*
6 Cara Redakan Nafsu Makan dan Ngemil Berlebih	*Eat Good*
Cara Mudah Kendalikan Nafsu Makan	*Eat Good*
5 Hal yang Harus Dihindari Sebelum Berlari	*Feel Good*
Tips Menjaga Stamina Saat Lari Jarak Jauh	*Feel Good*
…	…

Rajin Olahraga Bikin Nafsu Makan Meningkat	*Look Good*

4.2 Pengolahan Data

Pada pengolahan data terdapat tiga tahap yaitu perancangan pengelompokan dengan reduksi fitur *Information Gain Thresholding* dan metode *k-means*, diagram alir, dan manualisasi. Perancangan pengelompokan dengan metode *k-means* membahas mengenai gambaran atau rancangan metode k-means untuk pengelompokan, diagram alir membahas mengenai langkah langkah dan metode pengelompokan beserta *reduksi fiturnya*, dan manualisasi membahas mengenai perhitungan manual yang ada dalam proses pengelompokan artikel.

4.2.1 Perancangan pengelompokan dengan Metode K-Means



Gambar 4. 1 Perancangan Pengelompokan dengan *Information Gain Thresholding* dan *K-Means*

Dalam perancangan diatas digambarkan proses pengolahan artikel sebelum dikelompokan hingga setiap artikel masuk dalam kelompok tertentu berdasarkan kedekatannya. Proses tersebut meliputi :

1. *Pre-Processing*

   Dalam proses ini, artikel akan dirubah menjadi bentuk kata yang memiliki makna. Kata inilah yang kemudian dijadikan sebagai *fitur* untuk setiap artikel.

2. *Reduksi Fitur*

   Dalam tahapan ini, *fitur* yang didapatkan dari pre-processing akan dikurangi dengan metode *Information Gain Thresholding*. Tujuan dari reduksi *fitur* ini sendiri adalah untuk mengurangi fitur yang ada sehingga meringankan perhitungan.

3. Pembobotan Kata

Pada proses ini, *fitur* yang ada akan diukur bobotnya. Tujuan dari pembobotan ini adalah untuk mendapatkan nilai dari setiap *fitur* yang nantinya akan berguna bagi proses penentuan kedekatan artikel.

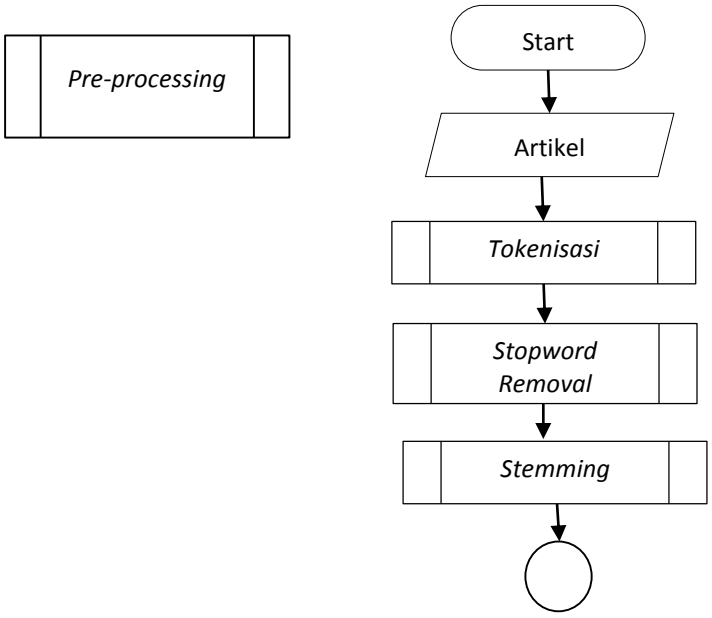
4. *Cosine Similarity*

   Pada tahapan ini artikel satu sama lain akan diukur jarakanya. Jarak terdekat akan dijadikan sebagai titik awal dari artikel tersebut.

5. *K-Means*

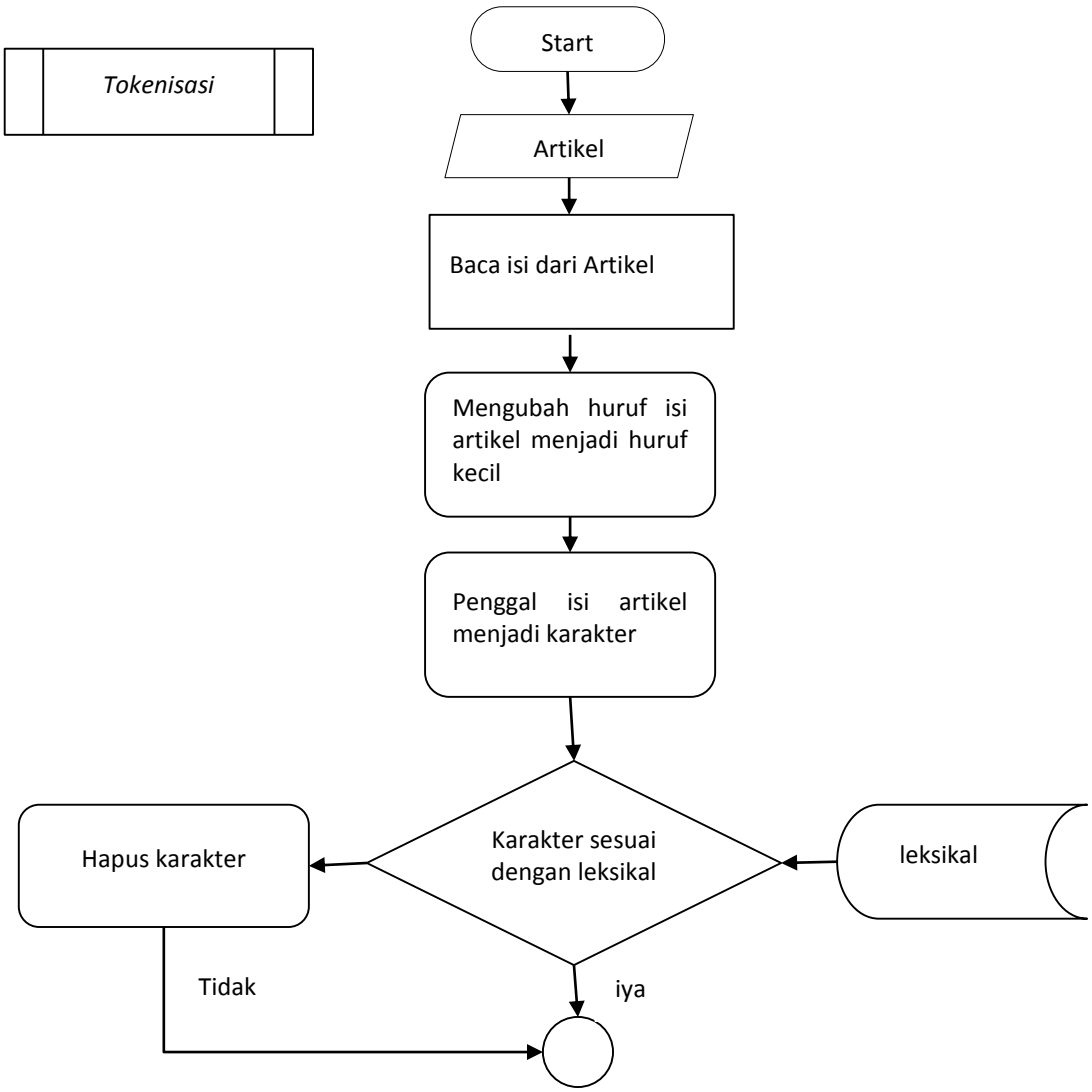
   Pada tahapan ini artikel akan dikelompokan ke dalam kelompok tertentu dimana dalam setiap kelompok, artikel memiliki kedekatan yang besar.

4.2.1.1 Pre-processing

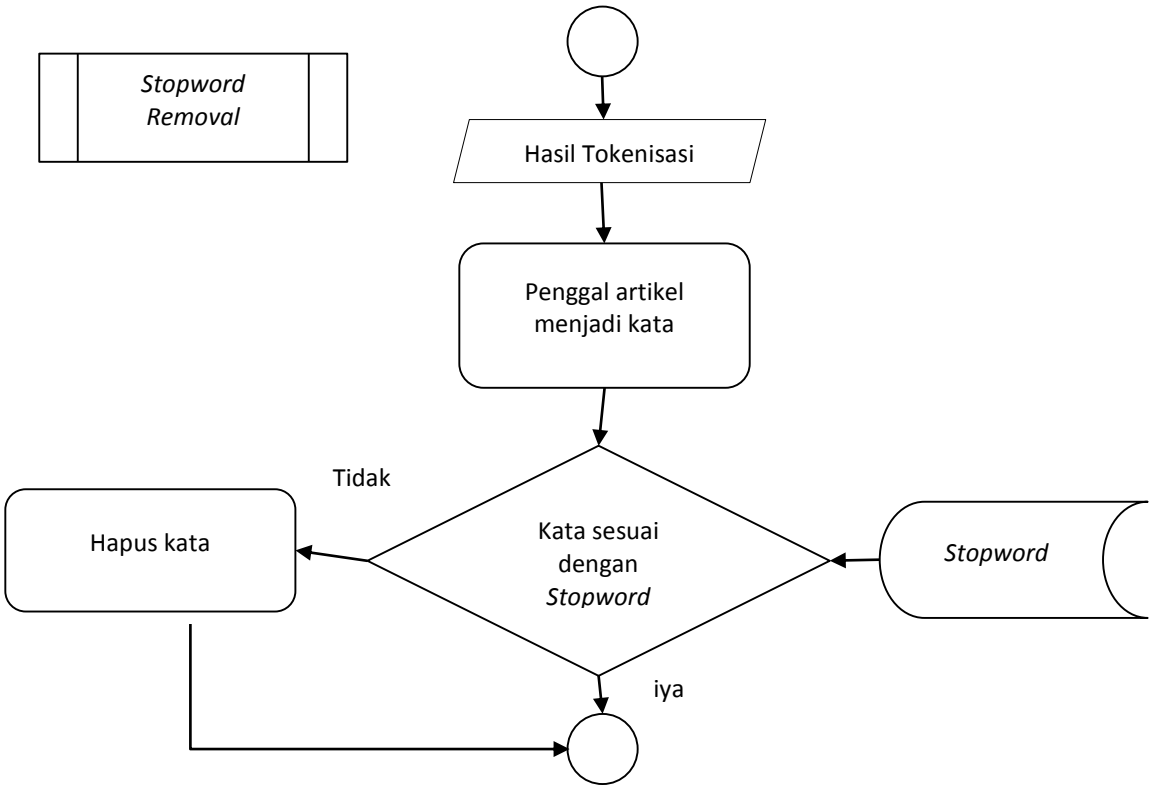


Gambar 4. 2 Diagram Alir *Pre-processing*

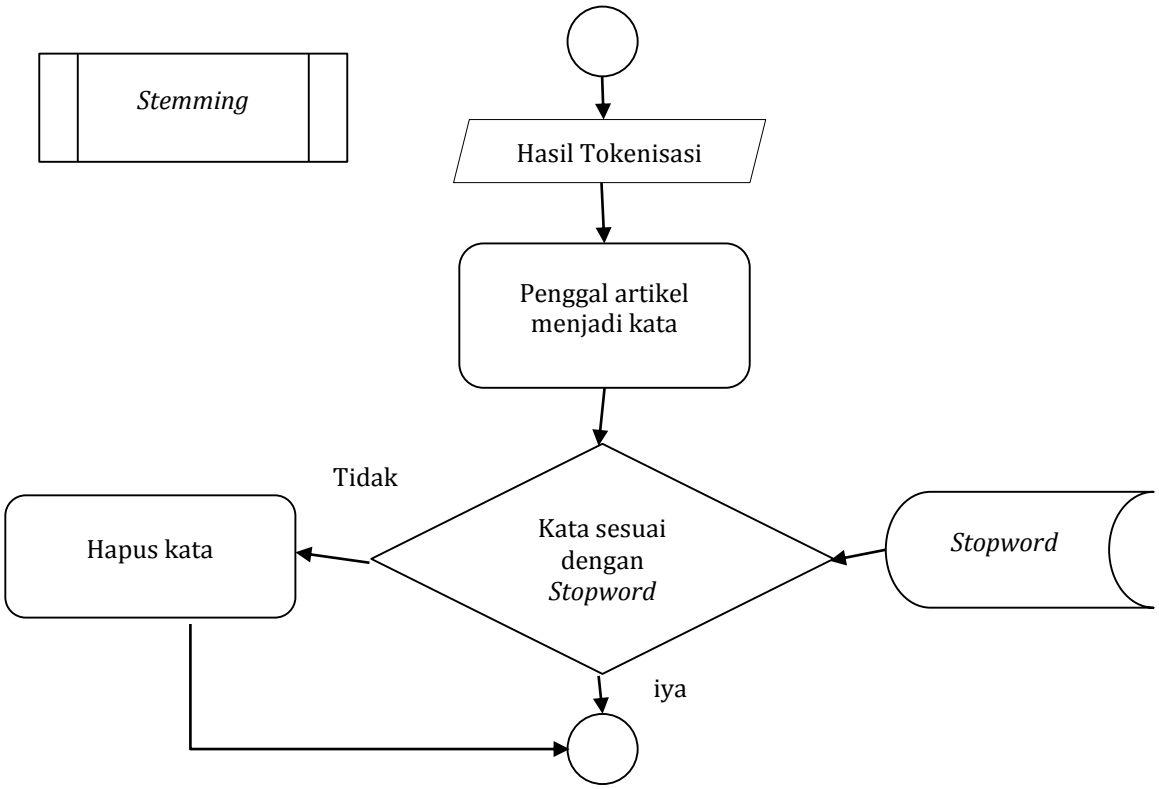
Dalam tahapan pre-processing, diberikan masukan berupa artikel sebanyak N dimana artikel bisa jadi memiliki kelompok artikel terkait yang berbeda. Kemudian dari n artikel akan diubah menjadi kata dasar (stem) yang kemudian dijadikan sebagai fitur untuk setiap artikel. Sebelum kata menjadi stem, terdapat 3 proses yaitu tokenisasi, stopword removal, dan stemming. Adapun proses stemming dijelaskan pada flowchart dibawah :



Gambar 4. 3 Diagram Alir Tokenisasi

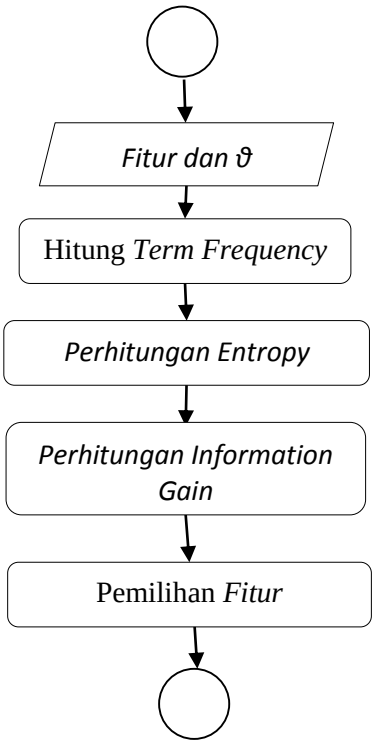


Gambar 4. 4 Diagram Alir Stopword Removal



Gambar 4. 5 Diagram Alir Stemming

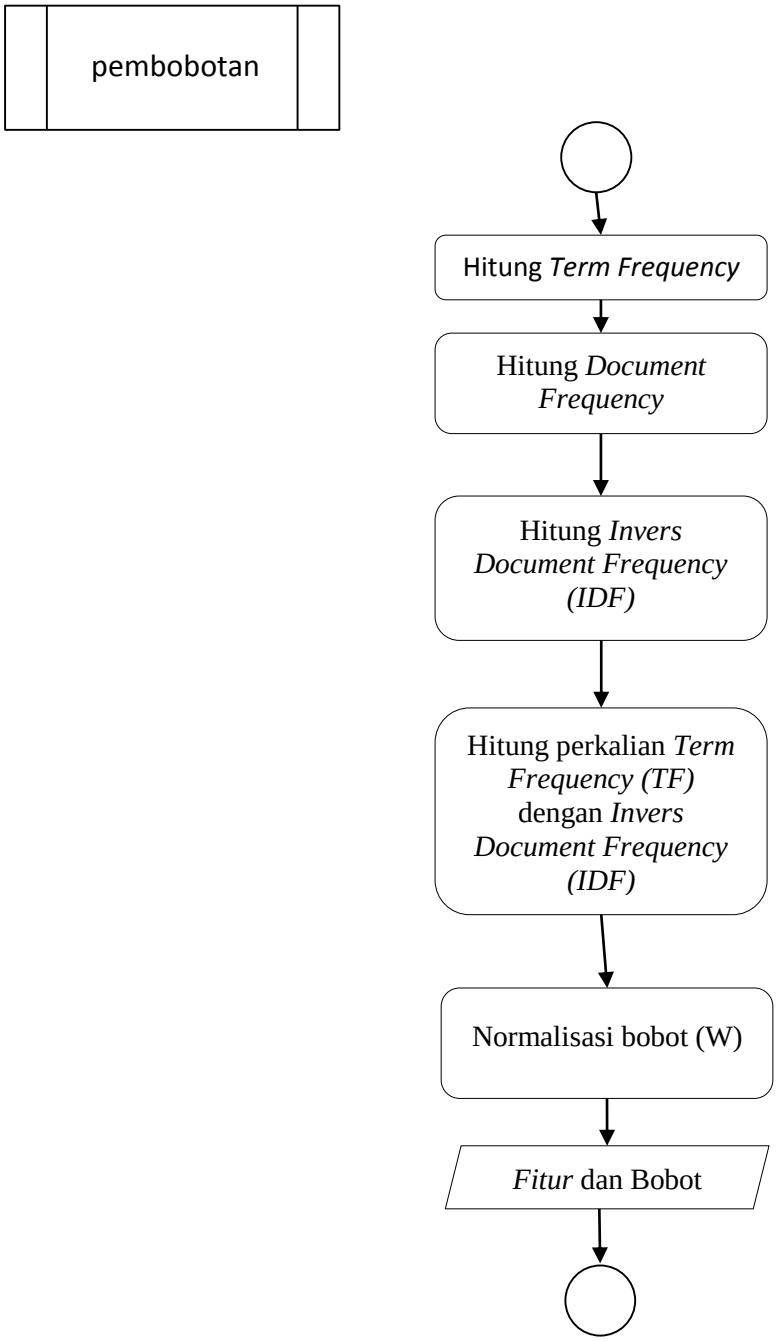
4.2.1.2 Reduksi Fitur



Gambar 4. 6 Diagram Alir *Reduksi Fitur*

Dari tahap pre-processing, telah didapatkan fitur untuk setiap artikel. Pada tahapan ini, fitur tersebut akan dijadikan sebagai masukan. Selain itu pada tahapan ini akan digunakan batas ambang (θ) yang digunakan sebagai nilai batas bawah suatu fitur dapat digunakan. Hasil keluaran dari proses ini adalah fitur baru yang lebih sedikit jumlahnya dibandingkan fitur semula.

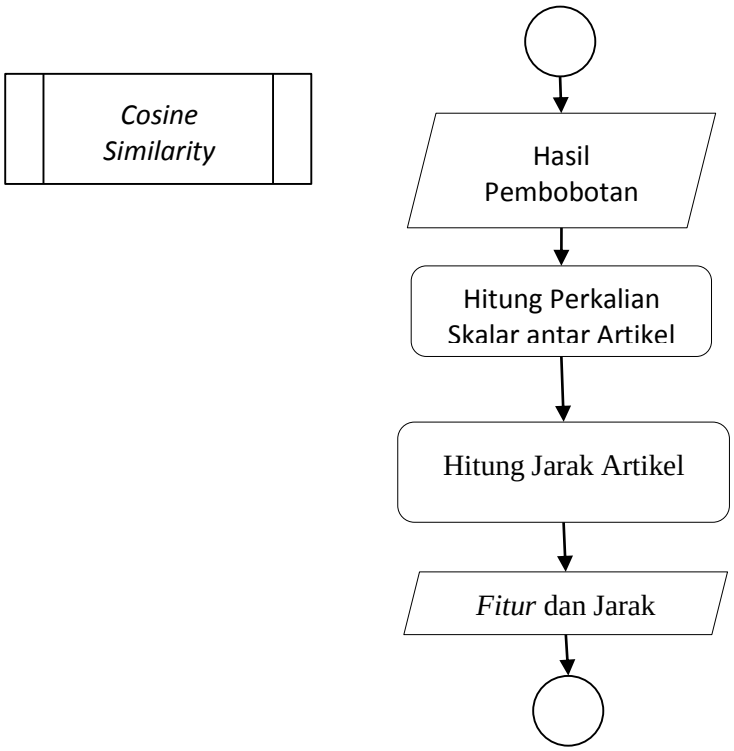
4.2.1.3 Pembobotan



Gambar 4. 7 Diagram Alir Pembobotan Kata

Dalam tahapan ini fitur baru yang yang didapatkan dari proses *reduksi fitur* dihitung bobotnya. Perhitungan bobot dilakukan dengan menghitung kemunculan kata di dalam satu artikel, kemudian menghitung kemunculan kata dalam seluruh artikel, dan dari hasil tersebut didapatkan bobot yang kemudian harus dinormalisasi agar nilainya tidak terlalu tinggi.

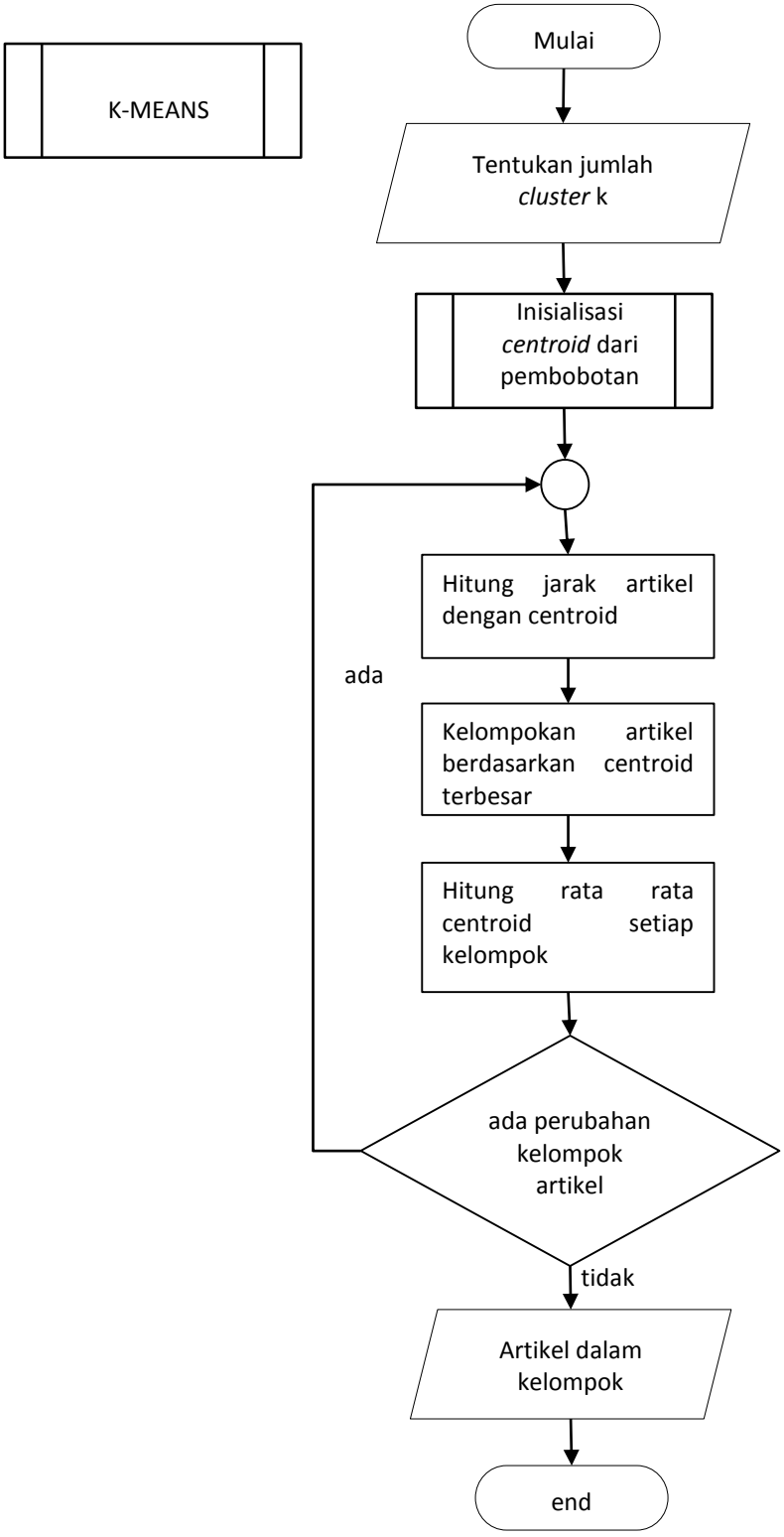
4.2.1.4 Cosine Similarity



Gambar 4. 8 Diagram Alir *Cosine Similarity*

Dalam tahapan ini fitur yang telah memiliki bobot dapat dijadikan sebagai alat ukur untuk menentukan kedekatan artikel satu dengan artikel lainnya. Pertama hitung perkalian antar dokumen (fitur artikel) yang dijadikan sebagai query dengan dokumen lainnya. Kemudian jumlahkan hasil perkalian dokumen untuk mendapatkan jarak.

4.2.1.5 K-Means



Gambar 4. 9 Diagram Alir K-Means

4.2.2 Manualisasi

Manualisasi berfungsi untuk memperhitungan artikel agar artikel dapat dikelompokan dalam kelompok yang sesuai dengan kedekatannya. Dalam tahapan manualisasi digunakan 5 data yaitu 3 artikel *eat good* dan 2 artikel *life good*. 5 data akan melalui 5 tahap yaitu *pre-processing*, *reduksi fitur*, penentuan kedekatan artikel dengan *cosine similarity*, dan pengelompokan dengan metode *k-means*. Pada tabel 4.1 dijelaskan mengenai fitur kata dari 5 data yang digunakan dalam perhitungan manual

Dokumen 1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
Masalahnya,	Studi	Sayangnya,	Berlari	Lari
kita	dari	Tiap	adalah	jarak
seringkali	USDA	Orang	olahraga	jauh
...	...	...	...	...
ini.	tersebut.	coba.	lari.	kecapekan.

Tabel 4. 2 Data untuk Perhitungan Manual

4.2.2.1 Pre-Processing

Dalam tahapan pre-processing ada 3 tahap yaitu *tokenisasi, stopword removal*, dan *stemming*.

1.2.3.1.1 Tokenisasi

Pada tahapan ini proses telah dijelaskan pada subab 2.3.1.1 dan melalui diagram alir 1.2.1.1.1. Hasil dari tokenisasi selanjutnya akan digunakan padatahap *Stopword Removal*.

Dokumen 1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
masalahnya	studi	sayangnya	berlari	lari
kita	dari	tiap	adalah	jarak
seringkali	usda	orang	olahraga	jauh
…	…	…	…	…
ini	tersebut	coba	lari	kecapekan

Tabel 4. 3 Hasil *Tokenisasi*

4.2.2.2 Stopword Removal

Pada tahapan ini proses telah dijelaskan pada subab 2.3.1.2 dan melalui diagram alir 1.2.1.1.2. Hasil dari *Stopword Removal* selanjutnya akan digunakan padatahap *Stemming*.

Dokumen 1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
sadar	studi	orang	berlari	lari
mengasup	usda	dinilai	olahraga	jarak
kalori	human	jenis	efektif	butuh

…	…	…	…	…
tips	dorongan	coba	lari	kecapekan

Tabel 4. 4 Hasil *Stopword Removal*

4.2.2.3 Stemming

Pada tahapan ini proses telah dijelaskan pada subab 2.3.1.3 dan melalui diagram Alir 1.2.1.1.3. Hasil dari *stemming* berupa *stem* akan dijadikan sebagi fitur sementara dari setiapartikel.

Dokumen 1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
sadar	studi	orang	lari	lari
asup	usda	nilai	olahraga	jarak
kalori	human	jenis	efektif	butuh
…	…	…	…	…
tips	dorong	coba	lari	capek

Tabel 4. 5 Hasil *Stemming*

4.2.3 Reduksi Fitur

Dalam tahapan *reduksi fitur* terdapat 3 proses yaitu perhitungan kemunculan kata dalam artikel, menghitung nilai *entropy* dengan persamaan 2.6 pada bab II, menghitung information gain dengan persamaan 2.7 pada bab II, dan terakhir memilih fitur berdasarkan batas ambang.

4.2.3.1 Perhitungan Kata

Pada tabel 4.5 terdapat data berupa fitur yang digunakan berdasarkan hasil *pre-processing* dan jumlah kemunculan dalam setiap artikel

tf(i,j)	d1	d2	d3	d4	d5	df(j)
sadar	2	0	0	0	0	1
asup	1	0	0	0	0	1
kalori	1	0	1	1	0	3
…	…	…	…	…	…	…
capek	0	0	0	0	1	1

Tabel 4. 6 Perhitungan Frekuensi Kata

4.2.3.2 Hitung *Entropy*

Untuk menghitung entropy pertama fitur kata yang sudah ada akan dikelompokan dalam 2 atribut yaitu Ada dan Tidak. Jumlah Ada dihitung berdasarkan jumlah kemunculan fitur dalam artikel pada kategori yang sama. Sebaliknya, jumlah Tidak didapatkan dari jumlah ketidakmunculan fitur dalam artikel pada kategori yang sama. C1 dan C2 merupakan kategori dari artikel.

ADA	1	2		TIDAK	1	2	

Cn	C1	C2	Total	Cn	C1	C2	Total
sadar	1	0	1	sadar	2	2	4
asup	1	0	1	asup	2	2	4
kalori	2	1	3	kalori	1	1	2
…	…	…	…	…	…	…	…
capek	0	1	1	capek	3	1	4

Tabel 4. 7 Fitur Ada dan Tidak

Setelah pengelompokan atribut, entropy baru bisa dihitung dengan persamaan 2.6 pada bab II. Hasilnya ditunjukan pada tabel 4.6

Entropy Ada		Entropy Tidak	
Cn	Ada	Cn	Tidak
sadar	0	sadar	1
asup	0	asup	1
kalori	0.9183	kalori	1
…	…	…	…
capek	0	capek	0.8113

Tabel 4. 8 Hasil Perhitungan *Entropy*

4.2.3.3 Perhitungan Information Gain

Pada tabel 4.8 ditunjukan hasil perhitungan dari information gain, dimana untuk mendapatkan hasil tersebut digunakan persamaan 2.7 pada bab II.

Cn	
sadar	0.171
asup	0.171
kalori	0.02
…	…
capek	0.322

Tabel 4. 9 Hasil Perhitungan *Information Gain*

4.2.3.4 Pemilihan Fitur

Tabel 4.9 merupakan fitur baru yang akan digunakan untuk proses pengelompokan artikel. Hasil tersebut dipilih berdasarkan fitur yang memiliki nilai lebih sama dengan 0,4. Sedangkan fitur yang kurang dari nilai tersebut dianggap tidak penting sehingga dapat dihilangkan.

Cn	
makan	1
coba	1
orang	1
diet	1
lari	1

Tabel 4. 10 Fitur Baru

4.2.4 Pembobotan Kata

Pada tahap ini terdapat 4 proses yaitu menghitung kemunculan kata pada artikel, inverse dari jumlah artikel yang memuat fitur sama, pembobotan TF-IDF, dan normalisasi dari pembobotan TF-IDF

4.2.4.1 Perhitungan *Term Frequency dan Inverse Document Frequency*

Pada tahap ini digunakan persamaan 2.1 untuk menghitung kemunculan kata dalam satu artikel dan persmaan 2.2 untuk menghitung jumlah artikel yang memuat kata yang sama. Hasil dijelaskan pada tabel 4.10.

	tf(i,j)	makan	coba	orang	diet	lari
1	d1	1	1	0	0	0
2	d2	1	0	1	1	0
3	d3	1.47712125	1	1	1	0
4	d4	0	0	0	0	1.30103
5	d5	0	0	0	0	1.30103
	idf(j)	0.22184875	0.39794001	0.39794001	0.39794001	0.39794001

4.2.4.2 Pembobotan TF-IDF

Pada tahap ini digunakan persamaan 2.3 pada bab II. Hasil didapatkan dari perkalian antara nilai TF dan IDF yang didapatkan pada proses sebelumnya.Hasil dijelaskan pada tabel 4.11.

	tf(i,j)	makan	coba	orang	diet	lari
1	d1	0.221849	0.39794	0	0	0
2	d2	0.221849	0	0.39794	0.39794	0
3	d3	0.327698	0.39794	0.39794	0.39794	0
4	d4	0	0	0	0	0.517732
5	d5	0	0	0	0	0.517732

Tabel 4. 11 Hasil Pembobotan TF-IDF

4.2.4.3 Normalisasi Pembobotan TF-IDF

Pada tahap ini hasil pembobotan TF-IDF dinormalisasi dengan persamaan 2.4 pada bab II. Hasil normalisasi dijelaskan pada tabel 4.12.

	tf(i,j)	makan	coba	orang	diet	lari
1	d1	0.486935	0.873438	0	0	0
2	d2	0.36674	0	0.657838	0.657838	0
3	d3	0.42938	0.521419	0.521419	0.521419	0
4	d4	0	0	0	0	1
5	d5	0	0	0	0	1

Tabel 4. 12 Hasil Normalisasi Pembobotan TF-IDF

4.2.5 Cosine Similarity

Pada tahap ini digunakan persamaan 2.5 pada bab II. Dimana pada tahap ini perhitungan cosine similarity dilakukan dengan 2 proses yaitu menghitung perkalian dari tiap dokumen. Proses ini dijelaskan pada tabel 4.13.

	tf(i,j)	makan	coba	orang	diet	lari
1	d1d1	0.23710617	0.76289383	0	0	0
2	d1d2	0.17857875	0	0	0	0
3	d1d3	0.20908053	0.45542691	0	0	0
4	d1d4	0	0	0	0	0
5	d1d5	0	0	0	0	0
6	d2d2	0.13449827	0	0.43275086	0.43275086	0
7	d2d3	0.15747098	0	0.34300907	0.34300907	0
8	d2d4	0	0	0	0	0
9	d2d5	0	0	0	0	0
10	d3d3	0.18436748	0.27187751	0.27187751	0.27187751	0
11	d3d4	0	0	0	0	0
12	d3d5	0	0	0	0	0
13	d4d4	0	0	0	0	1
14	d4d5	0	0	0	0	1
15	d5d5	0	0	0	0	1

Tabel 4. 13 Hasil Perkalian Antar Dokumen

Selanjutnya proses kedua yaitu mencari jarak antar dokumen dijelaskan pada tabel 4.14.

	d1	d2	d3	d4	d5	c1	c2
d1	1	0.17857875	0.66450744	0	0	0.17857875	0
d2	0.17857875	1	0.84348912	0	0	1	0
d3	0.66450744	0.84348912	1	0	0	0.84348912	0
d4	0	0	0	1	1	0	1
d5	0	0	0	1	1	0	1

Tabel 4. 14 Jarak tiap Dokumen

4.2.6 Perhitungan manual Pengelompokan

Untuk pengelompokan artikel, jarak yang didapatkan dari proses selanjutnya dipilih secara acak untuk dijadikan centroid. Centroid yang digunakan dalam perhitungan manual dituliskan pada tabel 4.15

	tf(i,j)	makan	coba	orang	diet	lari
1	d1	0.48693549	0.87343794	0	0	0
2	d2	0.36674006	0	0.65783802	0.65783802	0
3	d3	0.42938035	0.52141874	0.52141874	0.52141874	0
4	d4	0	0	0	0	1
5	d5	0	0	0	0	1
6	c1	0.36674006	0	0.65783802	0.65783802	0
7	c2	0	0	0	0	1

Tabel 4. 15 Centroid Pertama

Namun untuk centroid selanjutnya, centroid dicari dengan menggunakan persamaan 2.8 pada bab II.

	tf(i,j)	makan	coba	orang	diet	lari
1	d1	0.48693549	0.87343794	0	0	0
2	d2	0.36674006	0	0.65783802	0.65783802	0
3	d3	0.42938035	0.52141874	0.52141874	0.52141874	0
4	d4	0	0	0	0	1
5	d5	0	0	0	0	1
6	c1	0.4276853	0.46495223	0.39308559	0.39308559	0
7	c2	0	0	0	0	1

Tabel 4. 16 Centroid Baru

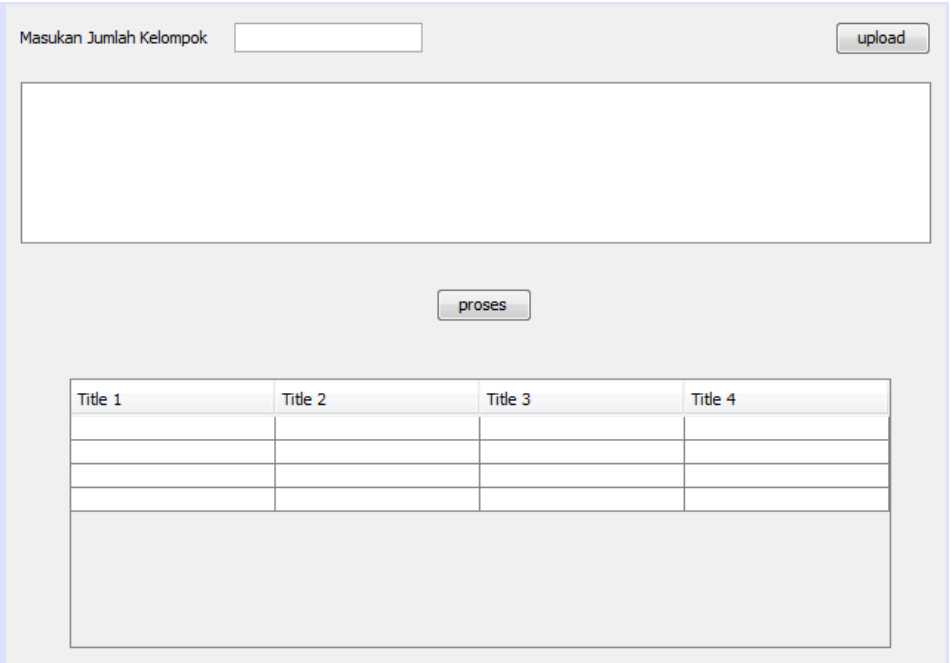
Selama centroid terus berubah dan letak artikel terus berubah maka akan terus dilakukan perulangan hingga mencapai maksimal iterasi, atau jika tidak ada perubahan maka proses berhenti dan perulangan terakhir dijadikan kelompok terbaru.

C1	d1	d2	d3	d4	d5
C2	d1	d2	d3	d4	d5
CLUSTER	1	1	1	2	2

Tabel 4. 17 Hasil Pengelompokan

4.3 Perancangan Antarmuka

Rancangan antarmuka dibuat untuk memberikan gambaran mengenai tampilan dari sistem pengelompokan artikel. Tampilan dibuat sederhana agar sistem dapat digunakan dengan mudah. Rencana antarmuka terdiri dari fungsi unggah data, tombol pemrosesan, dan hasil keluaran. Adapun rencana implementasi antarmuka sistem dijelaskan pada gambar 4.10.



Gambar 4. 10 Rancangan Antarmuka